An Adaptation of Stochastic Curtailment
to Truncate Wald's SPRT
in Computerized Adaptive Testing

CSE Report 606

Matthew Finkelman
CRESST/Stanford University

September 2003

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Matthew Finkelman, Project Director, CRESST/Stanford University

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Institute of Education Sciences, or the U.S. Department of Education.

AN ADAPTATION OF STOCHASTIC CURTAILMENT TO TRUNCATE WALD'S SPRT IN COMPUTERIZED ADAPTIVE TESTING [1]

Matthew Finkelman
CRESST/Stanford University

Abstract

Computerized adaptive testing (CAT) has been shown to increase efficiency in educational measurement. One common application of CAT is to classify students as either proficient or not proficient in ability. A truncated form of Wald's sequential probability ratio test (SPRT), in which examination is halted after a prespecified number of questions, has been proposed to provide a diagnosis of proficiency. This article studies the further truncation provided by stochastic curtailment, where an exam is stopped early if completion of the remaining questions would be unlikely to alter the classification of the examinee. In a simulation study presented, the increased truncation is shown to offer substantial improvement in test length with only a slight decrease in accuracy.

[1] The author would like to thank David Rogosa, Edward Haertel, David J. Weiss, David Siegmund, and Wim J. van der Linden for their help and good advice.

A fundamental part of President George W. Bush's "No Child Left Behind" accountability plan is to determine whether each child has achieved an acceptable degree of proficiency in academic subjects. With the onset of this plan, classification of students as either "proficient" or "not proficient" has an increasingly important role in policy relevance. As the availability of computers becomes more widespread, computerized adaptive testing (CAT) may serve as an attractive alternative to paper-and-pencil tests in this context, for CAT offers the sizable advantage of shorter exam lengths while maintaining comparable accuracy.

A major challenge in CAT proficiency testing is to strike a balance between confidence of a correct decision and economy of the items administered. Therefore, the test termination rule constitutes an integral part of any CAT procedure. The sequential probability ratio test (SPRT) has been widely studied as one such rule. See Wald (1947) for the sequential analysis roots of this test; see Eggen (1999), Lin & Spray (2000), and Spray & Reckase (1996) for applications in educational measurement.

Because an upper bound must be placed on the number of items asked in a CAT testing session, administrators may employ a truncated version of the SPRT in real-life examinations. A typical method of ensuring a timely end to the exam is to set an upper limit to the number of questions presented, after which the test terminates automatically (Spray & Reckase, 1996). In this article, a more aggressive approach to test termination is discussed. This method retains the original upper bound, but it inserts an additional stopping rule, ceasing the test if future questions are unlikely to change the classification decision. It is adapted from the idea of stochastically curtailed tests (Lan, Simon, & Halperin, 1982).

The article begins with a description of the use of item response theory in educational testing. Then an explanation of the truncated SPRT is followed by the proposal of a new procedure, which is referred to as the "stochastically curtailed sequential probability ratio test". Finally, a simulation study is presented that compares the different methods.

Item Response Theory and Maximum Likelihood Estimation

In item response theory (IRT), the ith student's ability is denoted as a latent variable θ_i. Although θ is assumed to vary from student to student, the subscript i is dropped at this point for simplicity. The probability that a student of ability θ gives the right answer to question j is modeled by a known class of functions. Let

$$u_j = \begin{cases} 1 & \text{if a given student answers question j correctly} \\ 0 & \text{if the student answers question j incorrectly} \end{cases} \tag{1}$$

Then under the two-parameter logistic (2-PL) model (Lord, 1980),

$$p_j(\theta) \equiv p(u_j = 1|\theta) = \frac{1}{1 + e^{-1.7a_j(\theta - b_j)}}. \tag{2}$$

The 2-PL model will be used in this paper, and it will be assumed that a_j and b_j are known exactly for all j. In practice, these parameters can only be imperfectly estimated; however, a study has suggested that the classifications made by the SPRT are fairly robust to substitution of estimates in place of the true parameters a_j and b_j(Spray & Reckase, 1987).

For the 2-PL model, a sufficient statistic for estimating θ exists and is given by $\sum_{j=1}^{k} a_j u_j$, where k refers to the number of items presented to the examinee (Lord, 1980). The maximum likelihood estimate $\hat{\theta}$ of θ is then the solution to the equation

$$\sum_{j=1}^{k} a_j p_j(\theta) = \sum_{j=1}^{k} a_j u_j \tag{3}$$

(Lord, 1980). Solving this equation may not lead to a finite estimate of θ, for instance if all questions have been answered correctly (or all incorrectly). If the maximum likelihood estimate (MLE) is finite, an asymptotic $100(1-\xi)\%$ confidence interval for θ is given by

$$\hat{\theta} \pm z_{1-\xi/2} \frac{1}{\sqrt{\sum_{j=1}^{k} I(\hat{\theta}, u_j)}}, \tag{4}$$

where (Lord, 1980)

$$I(\hat{\theta}, u_j) = \frac{1.7^2 a_j^2}{e^{1.7a_j(\hat{\theta}-b_j)}(1+e^{-1.7a_j(\hat{\theta}-b_j)})^2} \tag{5}$$

is the information function of question j evaluated at $\hat{\theta}$, and $z_{1-\xi/2}$ is the $1-\xi/2$ quantile of the normal distribution. As will be seen later, the stochastically curtailed SPRT uses one-sided confidence intervals, which are of the form

$$\theta > \hat{\theta} - z_{1-\xi} \frac{1}{\sqrt{\sum_{j=1}^{k} I(\hat{\theta}, u_j)}} \tag{6}$$

and

$$\theta < \hat{\theta} + z_{1-\xi} \frac{1}{\sqrt{\sum_{j=1}^{k} I(\hat{\theta}, u_j)}}. \tag{7}$$

The right-hand side of Inequality 6 will be referred to as the endpoint of the upper one-sided asymptotic confidence interval for θ, and the right-hand side of Inequality 7 will be referred to as the endpoint of the lower one-sided asymptotic confidence interval for θ.

The Sequential Probability Ratio Test and CAT

Following Eggen (1999), the proficiency testing scenario begins by setting a cut point θ_0, which separates the students who are proficient ($\theta > \theta_0$) from those who are not ($\theta < \theta_0$). The SPRT then assigns an indifference region $(\theta_0 - \delta, \theta_0 + \delta)$, where δ is typically small. The indifference region may be thought of as the values of ability where it is most difficult to determine whether a student is a master or a nonmaster. Its introduction is necessitated by the fact that the SPRT, as it is used in this setting, requires two discrete values of θ at which to compare the likelihoods.

The hypothesis of inadequate proficiency is given by

$$H_0 : \theta \leq \theta_0 - \delta = \theta' \tag{8}$$

and the hypothesis of acceptable proficiency is given by

$$H_1 : \theta \geq \theta_0 + \delta = \theta''. \tag{9}$$

The desired Type I and Type II error rates are then set to α and β, respectively, by the administrator of the test.

In order to conduct an examination via CAT, an algorithm for selecting items must be specified. For the SPRT, Spray and Reckase (1996) cited work by Spray and Reckase (1994) to suggest maximizing the information function at the cut point θ_0. That is, among items that have not yet been administered, the one with the highest value of

$$I(\theta_0, u_j) = \frac{1.7^2 a_j^2}{e^{1.7a_j(\theta_0 - b_j)}(1 + e^{-1.7a_j(\theta_0 - b_j)})^2} \tag{10}$$

is presented next. Note that use of the criterion in Equation 10 results in a test where all students receive the same questions in the same order.

To make a decision to accept or reject the hypothesis H_0, the SPRT utilizes the likelihood function. Suppose that k items have been presented to an examinee, who has given a vector of responses $\mathbf{U_k} = (u_1, u_2, ..., u_k)$. The likelihood of θ in this case represents the probability that a student of ability θ would give the exact response pattern $\mathbf{U_k}$ to the set of k selected items. The likelihood of θ for one response u_j is given by

$$L(\theta; u_j) = p_j(\theta)^{u_j}(1 - p_j(\theta))^{1-u_j}. \tag{11}$$

Again following Eggen (1999), under the assumption of local independence (conditional independence of responses, given θ), the likelihood of the entire response pattern is equal to

$$L(\theta; \mathbf{U_k}) = \prod_{j=1}^{k} L(\theta; u_j). \tag{12}$$

The log likelihood ratio of the values θ'' and θ' is then defined as

$$\log LR(\mathbf{U_k}) = \log \frac{L(\theta''; \mathbf{U_k})}{L(\theta'; \mathbf{U_k})} = \sum_{j=1}^{k} \log \frac{L(\theta''; u_j)}{L(\theta'; u_j)}. \tag{13}$$

The SPRT makes the following decisions based on the test statistic $\log LR(\mathbf{U_k})$ (Wald, 1947):

Stop testing and accept H_0 if

$$\log LR(\mathbf{U_k}) \leq \log B(\alpha, \beta); \tag{14}$$

stop testing and reject H_0 if

$$\log LR(\mathbf{U_k}) \geq \log A(\alpha, \beta); \tag{15}$$

continue testing if

$$\log B(\alpha, \beta) < \log LR(\mathbf{U_k}) < \log A(\alpha, \beta), \tag{16}$$

where $A(\alpha, \beta)$ and $B(\alpha, \beta)$ are constants depending on the given error rates α and β. In the context of computerized adaptive testing, $A(\alpha, \beta)$ is usually set to $\frac{1-\beta}{\alpha}$ and $B(\alpha, \beta)$ is set to $\frac{\beta}{1-\alpha}$. In this article, this will be assumed to be the case for the SPRT and its extensions. Strictly speaking, these values of $A(\alpha, \beta)$ and $B(\alpha, \beta)$ do not guarantee that both of the desired error rates will be achieved exactly. However, they do ensure that if α' and β' are the respective true Type I and Type II errors, then

$$\alpha' + \beta' \leq \alpha + \beta \tag{17}$$

for the test of hypotheses given in Equations 8 and 9 (Wald, 1947). Inequality 17 implies that at least one of the desired error inequalities must hold for this test.

Equations 14 and 15 state that testing ceases if likelihood ratio becomes smaller than $B(\alpha, \beta)$ or larger than $A(\alpha, \beta)$. From a Bayesian viewpoint, the test has the interpretation of stopping when the posterior probability of proficiency leaves a region determined by $A(\alpha, \beta)$, $B(\alpha, \beta)$, and the prior probability. Let P_p be the prior probability that a student is proficient. Then the conditions in Equations 14 and 15 are equivalent to the following: Stop testing and accept H_0 if

$$\frac{P_p LR(\mathbf{U_k})}{P_p LR(\mathbf{U_k}) + (1 - P_p)} \leq \frac{B(\alpha, \beta) P_p}{B(\alpha, \beta) P_p + (1 - P_p)}; \tag{18}$$

Stop testing and reject H_0 if

$$\frac{P_p LR(\mathbf{U_k})}{P_p LR(\mathbf{U_k}) + (1 - P_p)} \geq \frac{A(\alpha, \beta) P_p}{A(\alpha, \beta) P_p + (1 - P_p)}; \tag{19}$$

continue testing if

$$\frac{B(\alpha, \beta) P_p}{B(\alpha, \beta) P_p + (1 - P_p)} < \frac{P_p LR(\mathbf{U_k})}{P_p LR(\mathbf{U_k}) + (1 - P_p)} < \frac{A(\alpha, \beta) P_p}{A(\alpha, \beta) P_p + (1 - P_p)}. \tag{20}$$

The left-hand side of Inequalities 18 and 19 represents the posterior probability of proficiency. The test thus accepts H_0 if this posterior probability gets unduly small and rejects H_0 if the posterior gets unduly large.

In real-life applications of the SPRT, testing cannot continue indefinitely until either Inequality 14 or 15 is enforced. The truncated SPRT provides a solution to this problem by placing an upper bound N on the number of questions that can be asked in a given session. If the exam has not ended after N questions have been asked, the test is halted and a classification is made on the basis of the test statistic $\log LR(\mathbf{U_N})$. A common-sense rule is then to declare the student proficient if $\log LR(\mathbf{U_N}) > 0$ and to declare the student not proficient if $\log LR(\mathbf{U_N}) < 0$. The truncated SPRT (hereafter TSPRT) stopping rules can thus be expressed as follows:

If $k < N$:

Stop testing and accept H_0 if

$$\log LR(\mathbf{U_k}) \leq \log B(\alpha, \beta); \tag{21}$$

stop testing and reject H_0 if

$$\log LR(\mathbf{U_k}) \geq \log A(\alpha, \beta); \tag{22}$$

continue testing if

$$\log B(\alpha, \beta) < \log LR(\mathbf{U_k}) < \log A(\alpha, \beta). \tag{23}$$

If $k = N$:

Stop testing.

Accept H_0 if

$$\log LR(\mathbf{U_k}) < 0; \tag{24}$$

reject H_0 if

$$\log LR(\mathbf{U_k}) > 0. \tag{25}$$

The Stochastically Curtailed Sequential Probability Ratio Test

In this section, a newly proposed procedure is introduced as a more aggressive alternative to the TSPRT. Its appeal lies in the fact that when testing for proficiency, the desire to maximize accuracy must be balanced by the desire to minimize the number of items presented. Many common loss functions increase with test length and with the respective numbers of masters and nonmasters who are misclassified (for example, Lewis & Sheehan, 1990). Therefore, if further testing cannot possibly change the classification decision associated with a provided rule, it is optimal for the exam to cease immediately. Note that this statement is true regardless of whether the current classification is correct or incorrect.

The method of stochastic curtailment (Lan, et. al, 1982) exploits this fact, but it also extends the observation to the case where a change in decision is possible but unlikely. Stochastic curtailment ceases testing and rejects H_0 if given k observations, the probability under H_0 that a decision D will accept H_0, $P_k(D = H_0|H_0)$, is no more than a prescribed value $1 - \gamma$. It stops testing and accepts H_0 if this probability under H_1, $P_k(D = H_0|H_1)$, is at least γ'. For the purposes of this article, γ' will be taken to be equal to γ.

In the SPRT setting, it may be computationally intensive to calculate the probability that H_0 will eventually be accepted. Moreover, H_0 and H_1 are composite hypotheses, so this probability will vary depending on the value of θ within H_0 or H_1. Although attention could focus on the simple hypotheses $H_0 : \theta = \theta_0 - \delta$ and $H_1 : \theta = \theta_0 + \delta$, the current approach is to plug in an estimate of θ, denoted $\tilde{\theta}$, to estimate the respective probabilities of classification made by the TSPRT. This variation on stochastic curtailment, applied to the TSPRT, will be called the "stochastically curtailed sequential probability ratio test" (hereafter SCSPRT).

To elaborate, suppose the TSPRT is to be used to classify a given student, and the maximum number of items presented to the student is set at

N. If the student has answered k questions ($k < N$) and a decision has not yet been reached, the TSPRT only has a maximum of $N - k$ remaining items at its disposal. If additionally the item-selection criterion is maximum information at the cut point θ_0 (Equation 10), then it is known to the administrator which items have a chance to be asked in the future. Let the current classification of the SCSPRT be acceptable proficiency if $\hat{\theta} > \theta_0$, and inadequate proficiency if $\hat{\theta} < \theta_0$. Then provided $\tilde{\theta}$ is precise enough to estimate θ accurately, the probability that the TSPRT's final classification will be the same as the current one may also be reasonably estimated. If this probability is greater than or equal to γ, the SCSPRT halts testing immediately.

The reader may wonder at this point why the current classification is based upon $\hat{\theta}$ rather than comparing $\log LR(\mathbf{U_k})$ to 0. In practice, these two criteria will give nearly identical results. In this article, use of the $\hat{\theta}$ condition was made because $\hat{\theta}$ considers a range of values of θ, rather than only evaluating the likelihood function at two values, as does $\log LR(\mathbf{U_k})$.

A Mathematical Look at the SCSPRT Algorithm

With the introduction complete, this section gives a mathematical description of the SCSPRT procedure. As will be seen, the SCSPRT halts the examination session whenever the TSPRT would do so, but it also stops testing under several other scenarios. Suppose maximum information at the cut point (Equation 10) has been chosen as the item selection criterion. Recall that Equation 3, which uses the sufficient statistic $\sum_{j=1}^{k} a_j u_j$, provides the solution to the MLE. Therefore, plugging in $\theta = \theta_0$ and $k = N$ to the left-hand side of Equation 3 yields a threshold value for proficiency if all N questions are to be asked. That is, $T = \sum_{j=1}^{N} a_j p_j(\theta_0)$ may be viewed as a cutoff value for a student who answers all N questions. The student may be classified as proficient if $\sum_{j=1}^{N} a_j u_j > T$, and not proficient otherwise.

Let $S_k = \sum_{j=1}^{k} a_j u_j$ denote the sufficient statistic for θ when k questions have been asked. In order to exceed the threshold T, the student must achieve the remaining $R_k = T - S_k$ units of the sufficient statistic. Let $\tilde{\theta}$ be an estimate of θ. Then an estimate of the expected value of S_N given the k responses already observed is

$$\hat{E}[S_N|\mathbf{U_k}] = S_k + \sum_{j=k+1}^{N} a_j p_j(\tilde{\theta}) \tag{26}$$

and an estimate of its standard error is

$$\sqrt{V\hat{a}r[S_N|\mathbf{U_k}]} = \sqrt{\sum_{j=k+1}^{N} a_j^2 p_j(\tilde{\theta})(1 - p_j(\tilde{\theta}))}. \tag{27}$$

In Equation 27 the first k questions do not contribute to $\sqrt{V\hat{a}r[S_N|\mathbf{U_k}]}$ since this value is conditioned on the k responses that have already been observed, which are thus made deterministic. The probability that a complete test of N questions would result in a decision D of inadequate proficiency is estimated by the normal approximation to the distribution of $S_N|\mathbf{U_k}$. Define the approximation of this probability as

$$\begin{aligned}\hat{P}_k(D = H_0) &\equiv \Phi\left(\frac{T - \hat{E}[S_N|\mathbf{U_k}]}{\sqrt{V\hat{a}r[S_N|\mathbf{U_k}]}}\right) && (28)\\ &= \Phi\left(\frac{R_k - \sum_{j=k+1}^{N} a_j p_j(\tilde{\theta})}{\sqrt{V\hat{a}r[S_N|\mathbf{U_k}]}}\right). && (29)\end{aligned}$$

If the TSPRT results in an acceptance of H_0, or if both $\hat{\theta} < \theta_0$ and $\hat{P}_k(D = H_0) \geq \gamma$ for a given γ, the SCSPRT accepts H_0. If the TSPRT results in a rejection of H_0, or if both $\hat{\theta} > \theta_0$ and $\hat{P}_k(D = H_0) \leq 1 - \gamma$, the SCSPRT rejects H_0. Additionally, if $T > S_k + \sum_{j=k+1}^{N} a_j$, the SCSPRT accepts H_0; if $T < S_k$, the SCSPRT rejects H_0. The latter conditions correspond to the cases where S_N cannot exceed T and where S_N must exceed T, respectively.

The estimate $\hat{P}_k(D = H_0)$ considers only the case where all N questions are to be asked, ignoring the possibility of an early end to the exam. Additionally, it uses S_N as its criterion rather than $\log LR(\mathbf{U_k})$ as used by the TSPRT. However, these facts should have little consequence for the error rate of the SCSPRT. The only sample paths neglected by this estimate are the improbable scenarios where the final classification decision made from $\log LR_{k'}(\mathbf{U})$, $k < k' \leq N$, by the TSPRT is different from the decision based on S_N that would be made under completion of the entire N questions.

A critical issue in this procedure is what value to use for $\tilde{\theta}$, which is pivotal in the calculation of $\hat{P}_k(D = H_0)$. A natural choice would be to take $\tilde{\theta} = \hat{\theta}$; however, due to sampling variation, this selection may lead to substantial overestimation or underestimation of the probability that H_0 will be accepted in the future. Since overestimation is a larger concern when deciding to stop the test in favor of H_0, and underestimation is a larger concern when stopping the test in favor of H_1, the following conservative approach may be taken. Take as $\tilde{\theta}$ the endpoint of an appropriate asymptotic one-sided confidence interval:

$$\tilde{\theta} = \hat{\theta} + z_{1-\xi} \frac{1}{\sqrt{\sum_{j=1}^{k} I(\hat{\theta}, u_j)}} \tag{30}$$

when $\hat{\theta} < \theta_0$ and take

$$\tilde{\theta} = \hat{\theta} - z_{1-\xi} \frac{1}{\sqrt{\sum_{j=1}^{k} I(\hat{\theta}, u_j)}} \tag{31}$$

when $\hat{\theta} > \theta_0$. The value obtained in Equation 30 is the endpoint of the lower one-sided asymptotic confidence interval for θ. It is therefore a conservatively high estimate for θ to be used when $\hat{\theta} < \theta_0$, so that acceptance of H_0 is being considered. Conversely, the value obtained in Equation 31 is the endpoint of the upper one-sided asymptotic confidence interval for θ. It is thus a conservatively low estimate for θ, which is appropriate when $\hat{\theta} > \theta_0$. In the improbable event that $\hat{\theta} = \theta_0$, so that the MLE is equal to the cut point, it is apparent that the true classification is unsure enough that testing should probably continue unless $k = N$. If $\hat{\theta} = \theta_0$, the regular TSPRT conditions are taken as the stopping rule (Equations 21 through 25). In this case, the experimenter may be indifferent between the two possible classifications.

Each of the estimates $\tilde{\theta}$ described above involves the MLE $\hat{\theta}$. The precision of the crucial estimate $\hat{P}_k(D = H_0)$ thus depends highly on the precision of $\hat{\theta}$. A well-known result is that $\hat{\theta}$ is asymptotically normally distributed with mean θ and standard error $1/\sqrt{\sum_{j=1}^{k} I(\theta, u_j)}$ (Hambleton, Swaminathan, & Rogers, 1991). Before the estimate of $\hat{\theta}$ has become precise enough, stopping decisions should not be made on its basis. Therefore, it is recommended that the stopping rules proposed in this section be imposed only when a precision condition has been met. Examples of such conditions would be requiring an estimate of the standard error of $\hat{\theta}$ to be less than a given value SE^*, or else requiring the number of questions administered to be greater than or equal to a set value $k^* < N$. In this report, the latter approach is taken; the difference between the two criteria may be the subject of future research.

Use of the normal approximation in Equations 28 and 29 is justified when many questions remain to be asked. For small values of $N - k$, an improvement may result from using the Edgeworth expansion, which includes additional terms in the approximation. See Barndorff-Nielsen and Cox (1989) for details. The current article considers only the normal approximation, but the Edgeworth expansion should be considered for use in the future.

The SCSPRT stopping rules can thus be summarized as follows:

If $k < k^*$:

Stop testing and accept H_0 if

$$\log LR(\mathbf{U_k}) \leq \log B(\alpha, \beta); \tag{32}$$

stop testing and reject H_0 if

$$\log LR(\mathbf{U_k}) \geq \log A(\alpha, \beta); \tag{33}$$

continue testing if

$$\log B(\alpha, \beta) < \log LR(\mathbf{U_k}) < \log A(\alpha, \beta). \tag{34}$$

If $k^* \leq k < N$:

Stop testing and accept H_0 if

$$\log LR(\mathbf{U_k}) \leq \log B(\alpha, \beta) \tag{35}$$

or if

$$T > S_k + \sum_{j=k+1}^{N} a_j \tag{36}$$

or if

$$\hat{\theta} < \theta_0 \qquad \text{and} \qquad \hat{P}_k(D = H_0) \geq \gamma; \tag{37}$$

stop testing and reject H_0 if

$$\log LR(\mathbf{U_k}) \geq \log A(\alpha, \beta) \tag{38}$$

or if

$$T < S_k \tag{39}$$

or if

$$\hat{\theta} > \theta_0 \qquad \text{and} \qquad \hat{P}_k(D = H_0) \leq 1 - \gamma; \tag{40}$$

continue testing otherwise.

If $k = N$:

Stop testing.

Accept H_0 if

$$\hat{\theta} < \theta_0; \tag{41}$$

reject H_0 if

$$\hat{\theta} > \theta_0. \tag{42}$$

For the case when $k = N$, the decision is made based on $\hat{\theta}$ rather than on $\log LR(\mathbf{U_N})$ in order to achieve consistency with the early truncation procedure of the SCSPRT, which bases decisions on $\hat{\theta}$. Use of $\hat{\theta}$ here makes virtually identical decisions with a rule specifying mastery if $\log LR(\mathbf{U_N}) > 0$. As stated in the section describing the SPRT, the SCSPRT will take $A(\alpha,\beta) = \frac{1-\beta}{\alpha}$ and $B(\alpha,\beta) = \frac{\beta}{1-\alpha}$ for the purposes of this article. Note that since the SCSPRT's stopping rules contain those of the TSPRT, the TSPRT can never result in a shorter test length than the SCSPRT for any examinee.

It should be noted that although the previous discussion assumes use of maximum information at θ_0 as the item selection criterion, another criterion may be chosen instead. As the SCSPRT is a test termination rule, it can be used with any item selection technique, including those with content and/or exposure constraints. However, use of more complicated selection rules necessitates a new method for calculating $\hat{P}_k(D = H_0)$, since the algorithm above assumes knowledge of future questions. This poses a difficult challenge if the SCSPRT is to be extended to general selection techniques. Nevertheless, the effort may be fruitful if the SCSPRT is proven successful in significantly reducing test lengths.

Potential Dominance of the SCSPRT over the TSPRT

As it is an aggressive stopping rule, the SCSPRT will typically have shorter average test length than the TSPRT, but this will be accompanied by larger false positive and false negative rates. However, if γ is set equal to 1, then the SCSPRT can only stop earlier than the TSPRT if a decision based on comparing $\hat{\theta}$ to θ_0 cannot possibly change under hypothetical completion of the exam. Therefore, insofar as decisions made using $\hat{\theta}$ match decisions made using $\log LR(\mathbf{U_k})$, the SCSPRT dominates the TSPRT when $\gamma = 1$. Moreover, if a simple modification of the SCSPRT is made, where $\log LR(\mathbf{U_k})$ controls decisions rather than $\hat{\theta}$, then exact dominance holds. In this case, the SCSPRT must make the same classification as the TSPRT (and thus must have identical error rates); however, the SCSPRT may record a shorter test length than the TSPRT, and never a greater length. The process of shortening another test in this way is referred to simply as "curtailment" in sequential analysis (for example, Lan, et. al, 1982). The following such cousin of the SCSPRT dominates the TSPRT and will be referred to as the "curtailed SPRT" (CSPRT):

If $k < N$:

Stop testing and accept H_0 if

$$\log LR(\mathbf{U_k}) \leq \log B(\alpha, \beta) \tag{43}$$

or if

$$\log LR(\mathbf{U_k}) + \sum_{j=k+1}^{N} \log \frac{p_j(\theta'')}{p_j(\theta')} < 0; \tag{44}$$

stop testing and reject H_0 if

$$\log LR(\mathbf{U_k}) \geq \log A(\alpha, \beta) \tag{45}$$

or if

$$\log LR(\mathbf{U_k}) + \sum_{j=k+1}^{N} \log \frac{1 - p_j(\theta'')}{1 - p_j(\theta')} > 0; \tag{46}$$

continue testing otherwise.

If $k = N$:

Stop testing.

Accept H_0 if

$$\log LR(\mathbf{U_k}) < 0; \tag{47}$$

reject H_0 if

$$\log LR(\mathbf{U_k}) > 0. \tag{48}$$

Inequality 44 denotes the case where the final log likelihood ratio must be negative, even if all remaining questions are answered correctly; inequality 46 denotes the case where the final log likelihood ratio must be positive, even if all remaining questions are answered incorrectly. From a decision standpoint, it is obvious that in these cases the test should immediately be halted with the appropriate classification made. Thus, this slight alteration of the SCSPRT yields a stopping rule that can never make a different decision than the TSPRT, but it may result in a shorter test length.

Simulation Method and Results

Because of the similarity between the TSPRT, SCSPRT, and CSPRT, simulation provided a matched comparison of them. For every simulee, the TSPRT, the CSPRT, and two versions of the SCSPRT were run simultaneously, using the same response pattern $\mathbf{U_k}$ to make a classification. If one stopping rule ended before the others, the remaining methods were allowed to ask more questions until they reached their respective stopping points. Thus, every simulee represented a set of matched observations, and statistics involving accuracy and test length could be computed for each method.

Both forms of the SCSPRT explained in this article, as well as the CSPRT, were simulated and compared to the TSPRT. The first form used $\tilde{\theta} = \hat{\theta}$, and the second used the conservative confidence interval endpoints for $\tilde{\theta}$ (Equations 30 and 31). To distinguish the two versions, the former will hereafter be referred to as SCSPRT1, and the latter as SCSPRT2. Note that the CSPRT requires no value of $\tilde{\theta}$.

Mimicking van der Linden (1998), 300 items following the 2-PL model were created randomly with $a_j \sim U[0.5, 1.5]$ and $b_j \sim U[-4, 4]$ independently for all j. Following one version of simulations conducted by Lin & Spray (2000), δ was set to .2; the error rates α and β were set to .05; and the cut score θ_0 was set to -.325, close to the value of -.32 chosen in the article being followed. Mimicking Spray & Reckase (1996), the maximum number of questions allowed, N, was set at 50. Regarding the parameters used only by the SCSPRT1 and SCSPRT2, the point at which further measures of truncation could start, k^*, was set to 20. γ was given a value of .95 for SCSPRT1 and SCSPRT2; for the CSPRT, $\gamma = 1$ as it requires. For all one-sided confidence intervals, 90% confidence ($\xi = .1$) was used.

In order to obtain a well-rounded test of the methods, multiple values of θ were used, with 2000 replications at each value. A correct classification decision was defined to be "proficient" for $\theta > -.325$ and "not proficient" for $\theta < -.325$. Simulation began by replicating at -.4 (a point near $\theta_0 = -.325$) and continued incrementally by .1 in the positive and negative directions until the percentage of correct decisions made by both the TSPRT and SCSPRT approached 100. This resulted in use of θ at all values between -.9 and .2, inclusive, incremented by .1.

Table 1 compares the percentages of correct decisions made in simulation by the TSPRT, SCSPRT1, SCSPRT2, and CSPRT. The accuracy rate of the TSPRT was never more than 1.5% greater than the accuracy rate of the SCSPRT1 at the corresponding θ value. 8 of the 12 values had a difference of no more than 1%, and the SCSPRT1 approached perfect classification about as quickly as the TSPRT. Not surprisingly, the SCSPRT1 was less accurate than the SCSPRT2, which succeeded in maintaining an accuracy level close to that of the TSPRT. The accuracy rate of the TSPRT was never more than 0.5% higher than that of the corresponding accuracy rate of the SCSPRT2. In fact, at $\theta = -.3$ and $\theta = -.2$, the SCSPRT2 actually had a slightly higher accuracy for this data due to the randomness involved in simulation. Finally, the CSPRT by definition gave the exact same error rates as the TSPRT. Overall, the difference in classification accuracy was relatively small.

Table 1: Percentages of Correct Decisions Made by the TSPRT, SCSPRT1, SCSPRT2, and CSPRT

θ	TSPRT	SCSPRT1	SCSPRT2	CSPRT
-.9	99.95	99.90	99.95	99.95
-.8	99.60	99.40	99.60	99.60
-.7	98.85	98.15	98.75	98.85
-.6	95.80	94.75	95.65	95.80
-.5	87.80	86.45	87.30	87.80
-.4	69.20	67.85	68.70	69.20
-.3	55.10	54.45	55.20	55.10
-.2	78.30	78.00	78.35	78.30
-.1	91.35	89.85	91.30	91.35
0	97.45	96.45	97.35	97.45
.1	99.75	99.50	99.65	99.75
.2	99.95	99.95	99.95	99.95

As the new approaches are more aggressive than the TSPRT, they made gains in test length. Table 2 gives the average test lengths for the TSPRT, SCSPRT1, SCSPRT2, and CSPRT. The SCSPRT1 recorded average test lengths at least 10 items shorter than those of the TSPRT for the 3 values of θ closest to the cut (-.4, -.3, and -.2). Moreover, at each of the 3 values of θ where the difference in accuracy between the TSPRT and SCSPRT1 was highest (-.5, -.4, and -.1), the SCSPRT1 recorded an average decrease in test length of at least 7.5 items. Additionally, even at the fringes of the θ values simulated (-.9, -.8, .1, .2), where the SCSPRT1 was found to be at least 99.4% accurate, reasonably large decrements of test length were found. For each of these values, average improvement in test length was at least 0.85 items presented.

Table 2: Average Test Lengths of the TSPRT, SCSPRT1, SCSPRT2, and CSPRT

θ	TSPRT	SCSPRT1	SCSPRT2	CSPRT
-.9	13.94	13.05	13.57	13.90
-.8	16.72	14.83	15.91	16.62
-.7	21.57	17.57	19.77	21.22
-.6	26.80	20.40	23.88	26.12
-.5	32.63	23.76	28.52	31.53
-.4	37.50	26.40	32.40	35.99
-.3	38.67	26.88	33.53	37.15
-.2	34.97	24.67	30.33	33.60
-.1	29.61	21.96	26.06	28.60
0	22.07	17.67	20.13	21.67
.1	17.09	14.99	16.10	16.92
.2	13.51	12.66	13.13	13.46

Since using one-sided confidence intervals for $\tilde{\theta}$ is more conservative than plugging in $\tilde{\theta} = \hat{\theta}$, the SCSPRT2's gains in test length were more modest than those of the SCSPRT1. However, for each θ value between -.5 and -.1, the SCSPRT2's average test length was at least 3.5 items less than that of the TSPRT. Additionally, the smallest decrement in average test length (occurring at $\theta = -.9$) was 0.37 for the SCSPRT2, so non-trivial improvement was found relatively far from the cut point. Finally, since the CSPRT had error rates identical to those of the TSPRT, any improvement in test length would recommend its use over that of the TSPRT. As it is more conservative than both the SCSPRT1 and SCSPRT2, the CSPRT was only slightly shorter than the TSPRT. For every θ value simulated, the average test length of the TSPRT was never more than 2 questions greater than the corresponding average test length of the CSPRT. Additionally, at the three θ points furthest from the cut (-.9, -.8, and .2), the mean improvement was never more than .1 items. These statistics suggest that use of the $\hat{P}_k(D = H_0)$ condition is a major force behind the comparatively large improvements in test length that are exhibited by the SCSPRT1 and SCSPRT2.

Conclusion

Wald's SPRT and TSPRT have been staples of adaptive proficiency testing for many years. The results of this study suggest that the SCSPRT may be an attractive alternative to the TSPRT when shorter test length is desired. In the simulations conducted in this article, the substantial gains in test length made by the SCSPRT were accompanied by relatively small losses in classification accuracy.

The SCSPRT procedure, as described in this article, assumes that the administrator of the exam knows which questions have a chance to be presented, and in what order. This is the case when the items are selected by maximum information at the cut point θ_0. In order to be applied more generally, the SCSPRT algorithm must be adjusted for the situation where the questions presented are random.

References

Barndorff-Nielsen, O.E., & Cox, D.R. (1989). *Asymptotic Techniques for Use in Statistics.* London: Chapman & Hall.

Eggen, T.J.H.M. (1999). Item Selection in Adaptive Testing with the Sequential Probability Ratio Test. *Applied Psychological Measurement, 23,* 249-261.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park: SAGE Publications.

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., & Sheehan, K. (1990). Using Bayesian Decision Theory to Design a Computerized Mastery Test. *Applied Psychological Measurement, 14,* 367-386.

Lin, C.-J., & Spray, J. (2000). Effects of Item-Selection Criteria on Classification Testing with the Sequential Probability Ratio Test (Research Report 2000-8). Iowa City, IA: American College Testing.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Spray, J.A., & Reckase, M.D. (1987). The Effect of Item Parameter Estimation Error on Decisions Made Using the Sequential Probability Ratio Test (Research Report 1987-17). Iowa City, IA: American College Testing.

Spray, J.A. & Reckase, M.D. (April, 1994). *The selection of test items for decision making with a computer adaptive test.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Spray, J.A., & Reckase, M.D. (1996). Comparison of SPRT and Sequential Bayes Procedures for Classifying Examinees Into Two Categories Using a Computerized Test. *Journal of Educational and Behavioral Statistics, 21,* 405-414.

van der Linden, W.J. (1998). Bayesian Item Selection Criteria for Adaptive Testing. *Psychometrika, 63,* 201-216.

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