Graduate
Management
Admission
Council®

Computerized Classification Testing in More Than Two Categories by Using Stochastic Curtailment

Jasper T. Wouda

and

Theo J. H. M. Eggen

Cito and Twente University

Presented at the CAT for Classification Paper Session, June 2, 2009

Abstract

Computerized classification testing (CCT) can be used to increase efficiency in educational measurement. The truncated sequential probability ratio test (TSPRT) has been widely studied as a decision algorithm in CCT for two or more categories. Finkelman (2003) added an algorithm to the TSPRT in the form of stochastic curtailment, to classify an examinee in an even earlier stage of testing. This stochastically curtailed SPRT (SCSPRT) halts testing when a change of classification is possible, but unlikely. Finkelman (2003) adapted the algorithm for two categories and with fixed item ordering. The current study replicates his results, replicates it in realistic settings, and subsequently generalizes the SCSPRT to three categories while using adaptive item selection. The results show increased efficiency both when using one and two cut points. Different item selection methods are discussed.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2009 by the Authors

Citation

Wouda, J. T., & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Jasper Wouda, Plantage Middenlaan 42-D, 1018 DH Amsterdam, Netherlands.
Email: jtwouda@gmail.com or jasper@psychometrics.nl

Computerized Classification Testing in More Than Two Categories by Using Stochastic Curtailment

Applications of computerized classification testing (CCT) show how adaptive testing can be used to increase efficiency in educational measurement. CCT differs from conventional computerized adaptive testing (CAT) in that in CCT, a person's exact (latent) ability level, θ_i, is not especially important, as long as the person is correctly classified after reaching a certain threshold. Classifying a person can be important in determining whether someone has reached a certain level of proficiency in, for instance, mathematics. For both CCT and CAT, most item banks used have been calibrated using item response theory (IRT).

The SPRT is a widely used method for classification problems in CCT (see Thompson, 2007, for an overview of the different methods used). The SCSPRT (Finkelman, 2003, 2004, 2008) is an expansion of the SPRT, but as it is used now it is only applicable for mastery testing (CCT with one cut point). This is because it needs information of items N (the maximum number of items) minus k (the current number of items administered) to come to a classification decision. To be able to obtain information of these "future" items, using the SCSPRT as it is used now a test is needed that uses fixed item ordering. Fixed item ordering is thus far only useful in classification problems that use one cut point.

The Current Study

In the current study the SCSPRT will be generalized to three levels or two cut points. Consequently, the item selection algorithm has to be adaptive and the result is that items cannot be fixed in the test (Eggen & Straetmans, 2000). Put differently, in order to be effective, the SCSPRT has to be adapted to make it work with three classification levels. First the SPRT will be explained, then the SCSPRT will be discussed. Subsequently, the SCSPRT will be modified for three proficiency levels. Then the simulations conducted by Finkelman (2003) will be replicated and the algorithm will be tested with realistic settings. This replication was carried out not only to test our implementation of the SCSPRT, but also because it is not an exact replication; in this study. a slightly different version of the item information function was used, together with a different likelihood estimator.

Finally the adapted SCSPRT will be tested with two cut points, using simulated data with real parameters, realistic θ distributions, and sensible cut points. The goal of this study was to test the SCSPRT in somewhat more realistic settings, generalize it to more than two levels, and to explore which item selection method has the largest efficiency gain, while keeping usability in mind.

The IRT Model

Let u_{ij} be 1 if a given respondent i has answered item j correctly, and 0 if the respondent has answered the item incorrectly. In IRT, the i^{th} examinee's ability is regularly denoted as a latent variable θ_i. Although θ is assumed to vary from person to person, the subscript i is dropped at this point for simplicity. In this study, the IRT model used was the two-parameter logistic (2PL) model, which is given by (Birnbaum, 1968) as

$$p_j(\theta) = P(U_{j=1} \mid \theta) = \frac{\exp\left[a_j(\theta - b_j)\right]}{1 + \exp\left[a_j(\theta - b_j)\right]}, \qquad (1)$$

in which b_j is the difficulty parameter and a_j the discrimination parameter. The estimator of θ used here to estimate the maximum likelihood of θ is the weighed maximum likelihood (WML) estimator developed by Warm (1989). In the 2PL model the WML estimator is given by

$$\hat{\theta}_{(k)} = \max_\theta \left(\sqrt{\left(\sum_{j=1}^{k} I_j(\theta) \right) \prod_{j=1}^{k} p_j(\theta)\left[1 - p_j(\theta)\right]} \right), \qquad (2)$$

which is the generally statistically superior variant of the unweighted maximum likelihood estimator (Eggen, 1999), in which k is the number of items and $I_j(\theta)$ is the Fisher item information of item j. In the 2PL model this function is given by

$$I_j(\theta) = a_j^2 p_j(\theta)\left[1 - p_j(\theta)\right]. \qquad (3)$$

The Sequential Probability Ratio Test

The Sequential Probability Ratio Test (SPRT) is a widely used method in CCT for determining to which of a limited number of categories the θ level of an examinee belongs. The first applications of Wald's (1947) SPRT in CCT (Lewis & Sheenan, 1990; Reckase, 1983; Spray & Reckase 1994, 1996) involved only one cut point on the θ scale with two categories (e.g. mastery and non-mastery). Eggen and Straetmans (2000) and Spray (1993) showed that the SPRT can also be utilized in a CCT for classification into one of three categories. Users of the SPRT procedure (e.g., Jiao, Wang, & Lau, 2004; Eggen & Straetmans, 2000) typically set a maximum to the number of items to be administered under real testing conditions. This is often implemented as an extra stopping rule at $k = N$ items, where k is the number of currently administered items and N is the defined maximum number of items to be administered. This feature makes this procedure a truncated form of the conventional SPRT (TSPRT).

The Truncated Sequential Probability Ratio Test (TSPRT)

The TSPRT is a sequential testing procedure, in which the likelihoods of a statistical hypothesis and an alternative are compared. Cut points have to be set first on the θ scale to separate the classification levels. Subsequently, indifference regions δ are assigned around these cut points (see Eggen & Straetmans, 2000). These indifference regions can be seen as the areas in which it is most difficult for the TSPRT to be able to classify an examinee. Therefore the outer bounds of these regions are used for comparing the ratio test. The cut points plus or minus the indifference regions mark the two values for which the TSPRT compares the likelihoods. The basic TSPRT rationale is the evaluation of the ratio of two likelihoods. These values are then used to test two statistical hypotheses against one another. In the case of two cut points, θ_1 and θ_2, these hypotheses are (Eggen & Straetmans, 2000)

$$H_{01}: \theta \leq \theta_1 - \delta_{11} \quad \text{(Level 1)} \quad H_{11}: \theta \leq \theta_1 - \delta_{12} \quad \text{(higher than 1)}$$

$$H_{02}: \theta \leq \theta_2 - \delta_{12} \quad \text{(lower than 3)} \quad H_{12}: \theta \leq \theta_2 - \delta_{22} \quad \text{(Level 3)}$$

. Table 1 displays the ratio tests in the form of stopping rules of the TSPRT, as used by Eggen and Straetmans (2000) and Jiao, Wang and Lau (2004). These are in fact generalizations of the stopping rules as used by Finkelman (2003, 2008) in the one cut point case. In Table 1, α

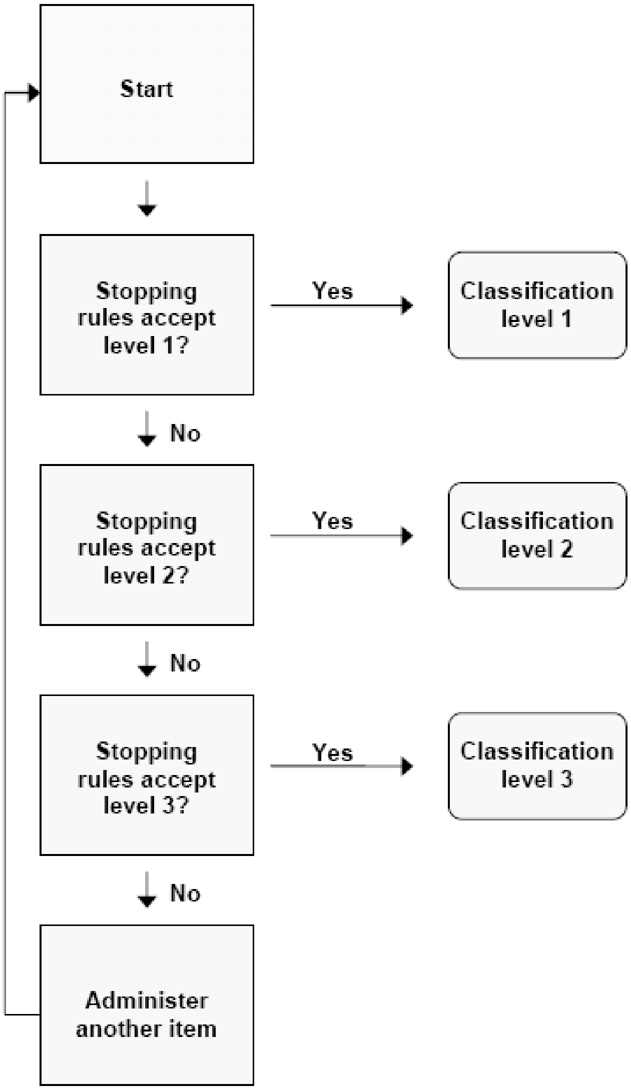
and *β* are constants representing the allowed decision error rates of the two statistical tests. The first stopping rule stops testing and accepts Level 1 if the weighed sum score is smaller than or equal to the log of the constant (which depends on the error rates) minus the sum of the log likelihood ratios of the values at which the hypotheses are tested, divided by the sum of the *δ* values around that cut point. At $k = N$, testing is ceased and a classification decision is forced. The decision rules at $k = N$ are the ratio tests as shown in Table 1, which are evaluated against the weighed sum. These decision rules are the same as the stopping rules, but without the ln(*β/*(*1 – α*)) or ln((*1 – β*)*/α*) parts.

Table 1. Stopping Rules of the Truncated Sequential Probability Ratio Test

	If $k < k^* < N$
Stop testing and accept Level 1 if	$\sum_{j=1}^{k} a_j u_j \le \dfrac{\ln\left(\dfrac{\beta_1}{1-\alpha_1}\right) - \sum_{j=1}^{k} \ln\left(\dfrac{1-p_j(\theta_1+\delta_{12})}{1-p_j(\theta_1-\delta_{11}}\right)}{\delta_{11}+\delta_{12}}$
Stop testing and accept Level 2 if	$\dfrac{\ln\left(\dfrac{1-\beta_1}{\alpha_1}\right) - \sum_{j=1}^{k} \ln\left(\dfrac{1-p_j(\theta_1+\delta_{12}}{1-p_j(\theta_1+\delta_{11}}\right)}{\delta_{11}+\delta_{12}} \le \sum_{j=1}^{k} a_j u_j \le \dfrac{\ln\left(\dfrac{\beta_2}{1-\alpha_2}\right) - \sum_{j=1}^{k} \ln\left(\dfrac{1-p_j(\theta_2+\delta_{22})}{1-p_j(\theta_2-\delta_{21}}\right)}{\delta_{21}+\delta_{22}}$
Stop testing and accept Level 3 if	$\sum_{j=1}^{k} a_j u_j \ge \dfrac{\ln\left(\dfrac{1-\beta_2}{\alpha_2}\right) - \sum_{j=1}^{k} \ln\left(\dfrac{1-p_j(\theta_2+\delta_{22})}{1-p_j(\theta_2-\delta_{21}}\right)}{\delta_{21}+\delta_{22}}$
else	continue testing

Figure 1 represents a flow chart of how the different classification decisions take place. This flow chart is applicable to the SPRT as well as the SCSPRT.

Figure 1. Flow Chart of the Stopping Rules of the (SC)SPRT



The SCSPRT

Finkelman (2004) states that the TSPRT is inefficient in that there are cases in which it presents another item while this item cannot change the classification decision about the examinee. The SCSPRT intervenes in these cases and halts testing when a change in classification is impossible. This part is called curtailment. However, the SCSPRT also halts testing in cases in which the probability of a change of classification decision is smaller than a predefined value, which is called *stochastic* curtailment. Finkelman (2003, 2008) showed that the TSPRT can be stochastically curtailed in order to shorten test length, while gaining in classification accuracy. Usually, in CAT, it is optimal to choose items that provide maximum information at the current θ estimate. In this way the test is adapted to the difficulty an examinee can handle. In mastery testing, it is optimal to choose items which have maximum information at the cut point (Eggen, 1999). This way the test only stays adaptive in test length, but is not adaptive in the selection of items. The problem of CCT with three levels is that there are two cut points. With one cut point it would have been clear around which cut point items have to be administered, because there is only one. With two cut points there is no best choice. Around which cut point items will be most informative for an examinee differs from one examinee to the other. This difference is explained by differences in ability between examinees. The other problem is that the SCSPRT, as applied by Finkelman (2003), is constructed in such a way that it needs information of all to be administered (N - k) items to make a classification decision. The next section formulates an answer to these problems in order to make the SCSPRT work in the case of three proficiency levels.

Extension. As can be seen in Finkelman (2003, 2008), the SCSPRT is an extension of the SPRT. It adds stochastic curtailment in the form of two extra stopping rules per level. Stochastic curtailment ceases testing and rejects H_{01} if given k observations, the probability that a decision D will accept H_{01}, $P_k(D = H_{01})$, is not higher than a set value $1 - \gamma$. Testing stops and accepts H_{01} if this probability is at least γ. As mentioned before, the TSPRT and the SCSPRT as used in a two-category problem have a set maximum number of items N (e.g. Spray & Reckase, 1996; Eggen, 1999; Finkelman 2003, 2004, 2008). When using one cut point, one can easily set item selection to be maximally informative at the cut point θ_0. An advantage of maintaining this selection criterion is that all items that will be administered are known. If k is the current number of items, then the possible future items, $N - k$, are known. Consequently it can also be estimated whether, given the expected value of the rest of the items, future items would change the current classification at k items.

Future items. However, in this study we would extend the SCSPRT to three proficiency levels. This means that there are two cut-scores, so maximum information at the cut-score cannot be used as a selection criterion (Eggen, 1999). How to provide the SCSPRT with information of future ($N - k$) items? It is not possible to know at which cut point one has to start administering items. So one solution is to calculate the optimal descending ordering of item information after every administered item and to plug that into the SCSPRT as information about “future items.” Other solutions could be to select items with highest information around the cut point nearest to the current θ estimate, or to select items in the middle, exactly between the two cut points.

The SCSPRT for Three Proficiency Levels

The SCSPRT uses fewer items than the SPRT. This means that there is less information to make a classification decision and this could cause higher decision error rates. In order to prevent misclassification due to imprecise estimations of θ, $\hat{\theta}$, we set, following Finkelman (2003), a minimum number of $k^* < N$ items that have to be completed before the extra stopping rules of the SCSPRT can take effect. So until $k^* < k$, only the stopping rules of the TSPRT apply (see Table 1). Let (Finkelman, 2003)

$$T = \sum_{j=1}^{N} a_j p_j(\theta_1), \qquad (4)$$

and let

$$S_k = \sum_{j=1}^{N} a_j u_j. \qquad (5)$$

Then the first additional stopping rule stops testing and accepts H_0 if

$$T > S_k + \sum_{j=k+1}^{N} a_j. \qquad (6)$$

This is the case in which the weighted sum of item scores that have highest information around the first cut point exceed the weighted sum of item scores of the administered items, plus the maximum score on the items to be administered which have highest information around the current θ estimate. If the administered items have a higher value for the weighted sum of item scores $\sum_{j=1}^{k} a_j u_j$ than the value of T, then the rule breaks off testing and accepts H_1. This is the curtailment stopping rule. The second additional stopping rule, denoted as $P_k(D = H_0)$, breaks off testing and accepts H_0 if

$$\Phi\left(\frac{T - S_k - \sum_{j=k+1}^{N} a_j p_j(\tilde{\theta})}{\sqrt{\sum_{j=k+1}^{N} a_j^2 p_j(\tilde{\theta})\left[1 - p_j(\tilde{\theta})\right]}}\right) \geq \gamma. \qquad (7)$$

If the normal approximation (Φ) of the weighted sum of item scores that have highest information around the first cut point, minus the expected value at N items, given k already administered items, divided by the standard error of that expected value, is greater than or equal to a set value of γ, then the SCSPRT stops testing and accepts H_0. However, this must go together with $\hat{\theta} < \theta_1$. Equation 7 is the stochastic curtailment stopping rule. The estimation of theta, $\tilde{\theta}$, as used in Equation 7, is a somewhat more conservative estimator of θ_k than $\hat{\theta}$. Finkelman (2003) sets this theta estimation to be

$$\tilde{\theta} = \hat{\theta} + z_{1-\zeta} \frac{1}{\sqrt{\sum_{j=1}^{k} I(\hat{\theta}, u_j)}} \qquad (8)$$

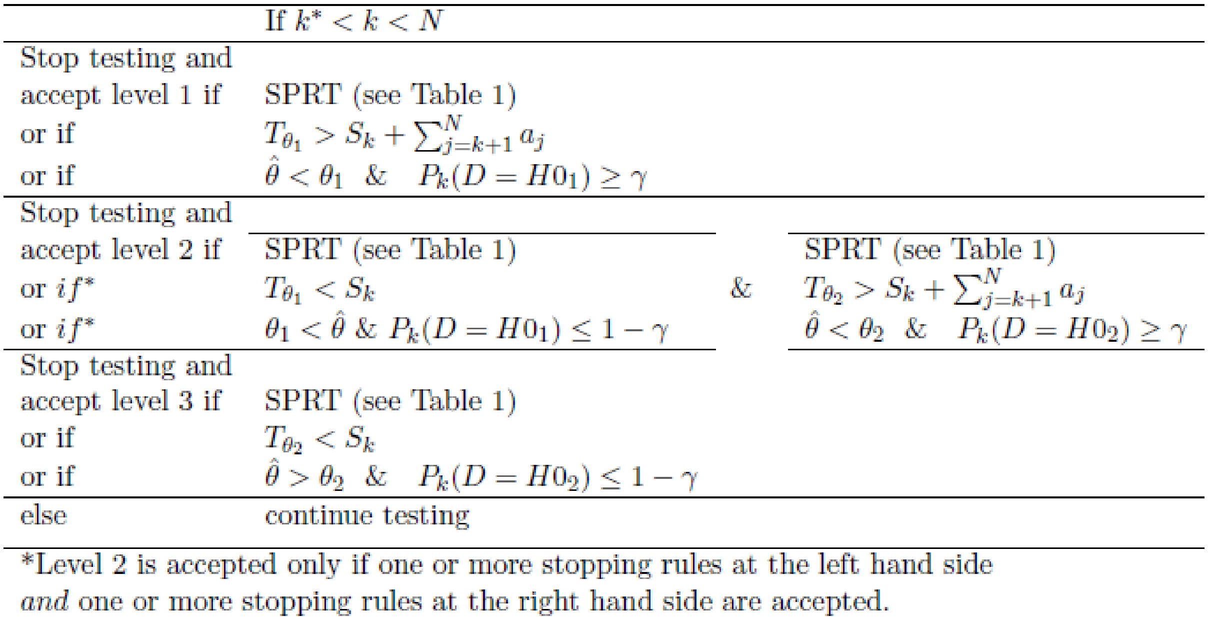
if the value of $\hat{\theta}$ is smaller than the respective cut point value, and

$$\tilde{\theta} = \hat{\theta} - z_{1-\zeta} \frac{1}{\sqrt{\sum_{j=1}^{k} I(\hat{\theta}, u_j)}} \qquad (9)$$

if the value of $\hat{\theta}$ is larger than the respective cut point value. These $\tilde{\theta}$ are the endpoints of the respective lower (Equation 8) and upper (Equation 9) one-sided confidence intervals of a θ estimate.

As can be seen in Equations 6 and 7, both additional stopping rules need information of items that could be administered in the future ($N - k$), in order to be able to assign examinees to a certain level. To get around this potential problem, we calculate the optimal descending ordering of Fisher item information at $\hat{\theta}_k$ after every administered item. Table 2 presents the stopping rules as they are applied when $k^* < k < N$. At $k = N$, the test is ended, and the decision rules from the TSPRT are applied.

Table 2. Stopping Rules of the SCSPRT for Three Levels



	If $k^* < k < N$		
Stop testing and			
accept level 1 if	SPRT (see Table 1)		
or if	$T_{\theta_1} > S_k + \sum_{j=k+1}^{N} a_j$		
or if	$\hat{\theta} < \theta_1$ & $P_k(D = H0_1) \geq \gamma$		
Stop testing and			
accept level 2 if	SPRT (see Table 1)		SPRT (see Table 1)
or if^*	$T_{\theta_1} < S_k$	&	$T_{\theta_2} > S_k + \sum_{j=k+1}^{N} a_j$
or if^*	$\theta_1 < \hat{\theta}$ & $P_k(D = H0_1) \leq 1 - \gamma$		$\hat{\theta} < \theta_2$ & $P_k(D = H0_2) \geq \gamma$
Stop testing and			
accept level 3 if	SPRT (see Table 1)		
or if	$T_{\theta_2} < S_k$		
or if	$\hat{\theta} > \theta_2$ & $P_k(D = H0_2) \leq 1 - \gamma$		
else	continue testing		

*Level 2 is accepted only if one or more stopping rules at the left hand side *and* one or more stopping rules at the right hand side are accepted.

Simulation Studies

Three simulation studies were conducted. Two studies were conducted using only one cut point, and one study using two cut points. In all studies, the same generated data was used per experiment for both sequential rules. The simulations were programmed in the R software environment for statistical computing and graphics (R Development Core Team, 2008). In our simulations we used the more conservative $\tilde{\theta}$ as an estimate for θ.

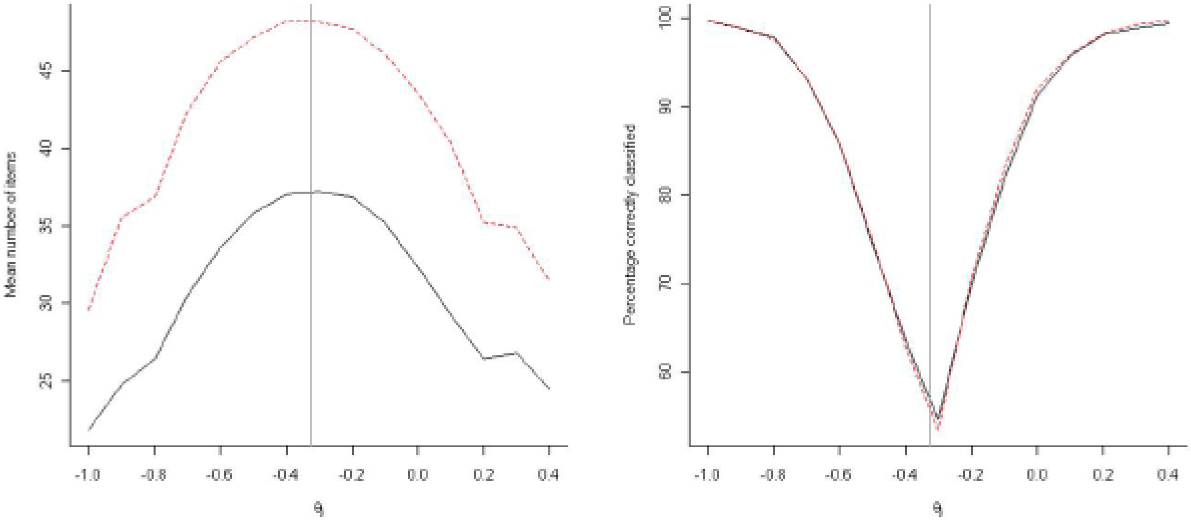
Simulation 1: Replication Study

The first study was conducted to replicate the results of Finkelman (2003). The same distribution and number of item parameters were used, a_j U[0.5, 1.5], b_j U[-4, 4], and the item bank consisted of 300 items. The same cut point, maximum number of items, and values for γ, α, and β were used: $N_{max}=50$, $\theta_0 = -0.325$, $\gamma = 0.95$, δ = 0.2, α = 0.05, β = 0.05. The only differences were the use of the WML estimator (Equation 2) and the use of a slightly different version of the Fisher information function (Equation 3).

Fifteen equidistant θ values were used as ability parameters for the simulees, with 2,000 replications per value. Figure 2 depicts the percentage of correctly classified simulees per θ value and the mean number of items used before a classification decision was taken.

Results. The percentage of correctly classified simulees was the same for the TSPRT as for the SCSPRT. The difference in mean number of items used between the SPRT and SCSPRT was quite large, with a minimum of 6.93 items, and a maximum of 13.12 items near the cut point. This is a larger mean difference than Finkelman (2003) reported.

Figure 2. Study 1: Mean Number of Items Used Before a Classification Decision Was Made and Percentage of Correctly Classified Simulees Per θ Value

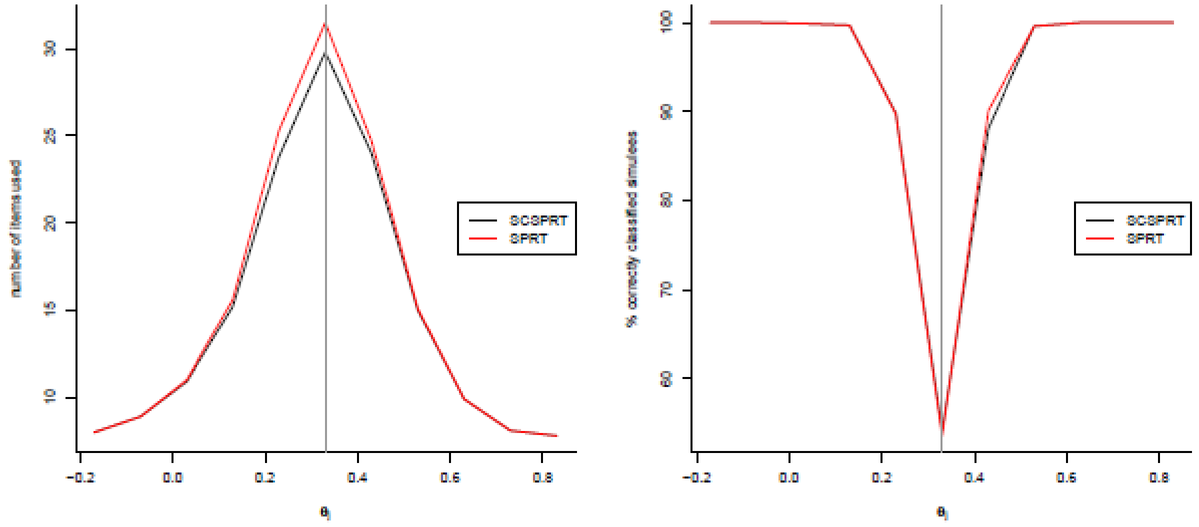


Simulation 2: Study With Real Item Parameters

The second study was conducted using real item parameters from the Eggen and Straetmans (2000) study. These item parameters originated from a mathematics item bank consisting of 250 calibrated items. To satisfy the necessary scaling constraints, the geometric mean of the estimated discrimination indices was fixed at $\left(\prod \hat{a}_j\right)^{1/250}$, and the sum of the estimated difficulty parameters was fixed at $\sum_{j=1}^{N} \hat{b}_j = 0$. The cut points between the three θ levels, $\theta_1 = -0.13$ and $\theta_2 = 0.33$, were set as realistic values (see Eggen & Straetmans, 2000). However, because only one cut point was needed in this study, we used θ_2 as the one cut point. The error rates were set at $\alpha_1 = \beta_1 = 0.05$, and the δ regions at $\delta_{11} = \delta_{22} = 0.2$. The maximum number of items N was set at $N_{max} = 40$. The value of γ was set at $\gamma = 0.975$. The θ_j of every simulee was randomly drawn from a distribution with a mean of 0.294, and a standard deviation of 0.522; this was the θ distribution as estimated in the calibration study. The point at which further measures of truncation could start was set to, following Finkelman (2003), $k^* = 20$. The number of simulees was set to 5,000.

Results. Results showed that with a realistic cut point, θ distribution, realistic parameters, δ regions, and error rates, the percentage of correctly classified simulees was exactly the same for the SPRT and the SCSPRT, 95.05%. The mean difference between the number of items used by the SPRT and SCSPRT was 0.42 items, in favor of the SCSPRT. This is a much smaller difference than what had been shown in Study 1. Figure 3 shows the results for 10 equidistant θ points with 2,000 replications per point. This figure shows that the SCSPRT used slightly fewer items than the SPRT, especially near the cut point, while maintaining virtually the same percentage of correctly classified simulees. An explanation for this relatively small observed difference (Finkelman, 2008) is that the SCSPRT works less efficiently when there are items administered with high a parameters, with b parameters around the cut point. In the studied mathematics item bank there are such items, which have high information around the cut point, so this might be the reason for the smaller observed difference. To test this hypothesis, we reduced the mathematics item bank to the items with low *a* parameters. The items with an *a* parameter above 3.0 were omitted, resulting in an item bank consisting of 52 items with *a* = 2, and 115 items with *a* = 3 (167 items total). The result of this exploration was exactly as Finkelman (2008) predicted. With 2,000 simulees, the difference between the TSPRT and SCSPRT in number of items used, before a classification decision was made, was 2.71 items. This is still a more modest result than with the theoretical item bank, but it is a substantial difference. The percentage of correctly classified simulees was 94.35% for the TSPRT and 94.45% for the SCSPRT.

Figure 3. Study 2: Mean Number of Items Used Before a Classification Decision Was Made and Percentage of Correctly Classified Simulees Per θ Value



Simulation 3: Study With 2 Cut Points

Item bank. The same item bank, model configurations, and data simulation methods were used as in Study 2. The θ parameters were also randomly drawn from the same distribution. The difference in terms of data and item configuration was the use of an additional cut point, $\theta_1 = -0.13$ (see Eggen & Straetmans, 2000). In order to be able to produce figures for the most efficient item selection method, ten equidistant θ values were used as ability parameters for the simulees, with 2,000 replications per value.

Item selection. The most important difference was that with two cut points, it was not clear which item selection method would produce the best results. Therefore, three different item selection methods were explored: selection in the exact middle between the two cut points, selection at the nearest cut point, and selection at the current θ estimate $\hat{\theta}_k$. Information about "future" items was also chosen at these points. For the third method this resulted in different "future" items after every item administration.

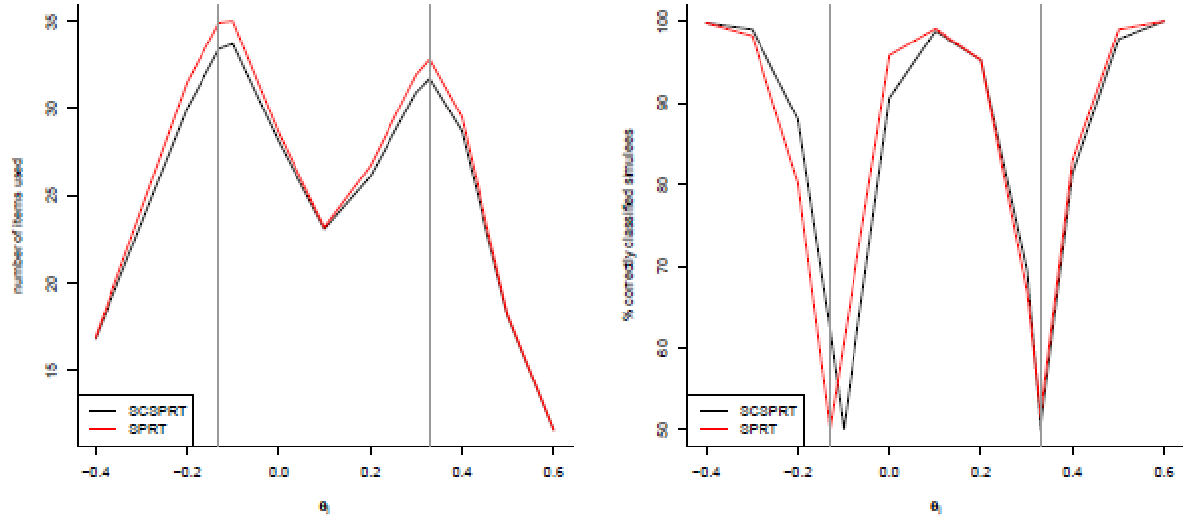
Results. As can be seen in Table 3, the results of selection with maximum information at the nearest cut point and selection at the current θ estimate were similar. Results of item selection in the middle, between the two cut points, were somewhat worse. The preferred method here was selection with maximum information at the current θ estimate, because this method is much more flexible regarding exposure control and content balancing methods. This means that this selection method is of much more practical use, because a larger part of the item bank can be used.

Table 3. Results of Different Item Selection Methods With the SPRT vs SCSPRT

Item selection	SCSPRT	SPRT	SCSPRT	SPRT
method	% correct	% correct	N	N
Max I in the middle	91.1	91.9	21.7	22.9
Max I at nearest cut point	90.7	91.2	19.2	20.1
Max I at current $\tilde{\theta}$	90.9	91.3	19.3	19.7

Figure 4 gives the mean number of items used per equidistant θ value and the mean percentage correct per θ value. This figure shows that the percentage correctly classified simulees was similar for the SCSPRT and SPRT, but that the SCSPRT used fewer items on average, especially near the cut points.

Figure 4. Study 3: Mean Number of Items Used Before a Classification Decision Was Made and Percentage of Correctly Classified Simulees Per θ Value



Discussion

One Cut Point

When using one cut point, the performance of the SCSPRT with a theoretical item bank was much more efficient compared to the SPRT. With a real item bank, under realistic circumstances, the gain in efficiency of the SCSPRT compared to the SPRT was more modest, but still quite substantial. Finkelman (2008) is correct in his conclusion that when highly discriminating items are omitted, the performance of the SCSPRT increases. However, our conclusion is that the results must be highly item bank dependent because the results with realistic settings, even when highly discriminating items are omitted, are more modest compared to our results with the theoretical item bank.

Two Cut Points

A generalization to more than two categories is possible. The performance of the SCSPRT with a real item bank was in general better than the SPRT with respect to the number of items needed for classification. This cost only a minimal loss in the percentage correct classification decisions. For both the SCSPRT and SPRT the selection method which chooses items with maximum information at the midpoint between the cutting points performed worse than the two other explored methods. As is the case with the SPRT, there were no large differences in performance between the remaining item selection methods; maximum information at the nearest cut point and maximum information at the current θ estimate performed similarly.

Future Research

Additional results showed that if the cut points were set further apart, while keeping everything else the same, the efficiency gain became larger for the SCSPRT as opposed to the SPRT. This must be due to the effect Finkelman (2008) described that the gain is larger when there are only low discriminating items available in a certain θ area. However, different types of item banks should be explored to test the SCSPRT in different situations.

References

Birnbaum, A. Some latent trait models. In Lord, F.M. and Novick, M.F. *Statistical theories of mental test scores.* Reading: Addlson-Wesley, 1968.

Davey, T., Godwin, J., & Mittelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment. *Journal of Educational Measurement, 34*, 21-41.

Eggen, T. J. H. M. (1999). Item selection in 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, 60* , 713-734.

Finkelman, M. (2003). *An adaptation of stochastic curtailment to truncate Wald's SPRT in computerized adaptive testing* (Tech. Rep.). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

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

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*, 442-463.

Jiao, H., Wang, S., & Lau, C. A. (2004). *An investigation of two combination procedures of SPRT for three-category classification decisions in computerized classification test.* Paper presented at the annual meeting of the American Educational Research Association, San Antonio, April 2004.

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

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

Spray, J. (1993). *Multiple-category classification using a sequential probability ratio test* (Tech. Rep.). Iowa City: ACT Research Report Series.

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.

R Development Core Team (2008). R: A language and environment for statistical computing [computer software]. Vienna, Austria: R Foundation for statistical computing. (Retrieved from http://www.R-project.org)

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12 (1)*, 1-13. (http://pareonline.net/getvn.asp?v=12&n=1)

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54 (3)*, 427-450.