Graduate
Management
Admission
Council®

Constrained Item Selection Using a Stochastically Curtailed SPRT

Jasper T. Wouda and Theo J.H.M. Eggen

Cito and Twente University

Presented at the CAT Research and Applications Around the World Poster Session, June 2, 2009

Abstract

Computerized classification testing (CCT) provides the means to increase efficiency in educational testing. The stochastically curtailed sequential probability ratio test (SCSPRT) (Finkelman 2003, 2004, 2008) has been tested as an efficient alternative to the most frequently used decision algorithm in CCT for mastery testing, the sequential probability ratio test (SPRT). However, in order to be applied in operational tests, non-statistical constraints must also be considered. In this study, the efficiency gain of different constraint handling methods were compared, together with different item selection methods. The applied constraints were content balancing and exposure control. The methods for exposure control compared for the SPRT and SCSPRT were the Sympson-Hetter method, the progressive method, and alpha-stratified testing. The methods for content balancing compared were the Kingsbury-Zara (Kingsbury & Zara, 1989; Kingsbury & Zara, 1991) approach and the weighted deviation method (WDM) of Stocking and Swanson (1993). Results show that the Kingsbury-Zara method, combined with the Sympson-Hetter approach, showed the largest gain in efficiency for the SCSPRT.

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. and Eggen, T. J. H. M. (2009). Constrained item selection using a stochastically curtailed SPRT. 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

Constrained Item Selection Using a Stochastically Curtailed SPRT

The goal of many applications of educational measurement is to distinguish between masters and non-masters of a subject or skill. In general, we refer to this type of testing as *mastery testing*. When the administration of such a test takes place by means of a computer, this is called a computerized classification test (CCT). CCT is, nevertheless, also the correct term for computerized classification routines with more than two classification levels. In CCT it is often important to use no more items than is necessary to classify an examinee. The exact latent ability level (θ_i) is not important as it is in traditional CAT; instead, it is important that a person is correctly classified in one of a few predefined classification levels. Generally CCT is more efficient than computerized adaptive testing (CAT) with respect to the number of items used. For both computerized methods, the item banks that are usually used have been calibrated using item response theory (IRT).

The sequential probability ratio test (SPRT) is a method relatively widely used in CCT to decide between pass/fail or mastery/non-mastery. The SPRT has no theoretical maximum on the number of items (Finkelman, 2008). Therefore, this procedure has often been truncated (TSPRT) in order to be of practical use. However, the TSPRT is suboptimal regarding test length. In some cases it presents an item for which the information provided by that item and its response cannot further change the classification decision. Algorithms were added to the TSPRT by using stochastic curtailment (Finkelman, 2003, 2004, 2008) in order to be able to classify an examinee in an earlier stage of testing. The resulting stochastically curtailed SPRT (SCSPRT) ceases testing if a classification change for an examinee is possible, but improbable.

To be able to test the additional stopping rules of the SCSPRT in a somewhat more realistic setting, non-statistical constraints must also be considered. Constraints are, for example, content balancing, answer key balancing, conflicting items, and a special case is item exposure control. In this study, the performance of the SPRT was compared with the performance of the SCSPRT for three methods of exposure control: the Sympson-Hetter method, the progressive method, and alpha-stratified testing. Methods for content balancing that were compared in this study are the Kingsbury and Zara (1989, 1991) approach and the weighted deviation method (WDM) of Stocking and Swanson (1993). These methods were compared by using simulated examinees (simulees) and simulated data.

Basic IRT

Before discussing the (SC)SPRT, some of the basic IRT used in this study must be introduced. The items are dichotomously scored, so u_{ij} is 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} student'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. Then under the two parameter logistic (2PL) model (Birnbaum, 1968)

$$p_j(\theta) = P(U_j = 1|\theta) = \frac{\exp a_j(\theta - b_j)}{1 + \exp a_j(\theta - b_j)}. \qquad (1)$$

where a_j is the discrimination parameter and b_j is the difficulty parameter. The ability estimator used to estimate the maximum likelihood of θ is weighted maximum likelihood (WML) (Warm, 1989). In the 2PL model this WML estimator is

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

This 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 ,

$$I_j(\theta) = a_j^2 p_j(\theta)(1 - p_j(\theta)). \tag{3}$$

The (Stochastically Curtailed) SPRT

The truncated SPRT. The TSPRT is a sequential testing procedure in which the likelihoods of a statistical hypothesis and an alternative are compared. A cut point, θ_0, is set along the θ scale to separate mastery ($\theta_j \geq \theta_0$) from non-mastery ($\theta_j \leq \theta_0$). Subsequently, indifference regions δ.. are assigned around this cut point (see Eggen & Straetmans, 2000). The basic TSPRT rationale is the evaluation of the ratio of two likelihoods. The test statistic is the ratio between the likelihoods of the hypotheses of mastery ($\theta_0 + \delta$) and non-mastery ($\theta_0 - \delta$). As can be seen in Eggen (1999), the TSPRT stops testing and accepts level 0 (non-mastery) if

$$\sum_{j=1}^{k} a_j u_j \leq \frac{ln(\frac{\beta}{1-\alpha}) - \sum_{j=1}^{k} ln(\frac{1-p_j(\theta_0+\delta_2)}{1-p_j(\theta_0-\delta_1)})}{\delta_1 + \delta_2}. \tag{4}$$

The TSPRT stops testing and accepts level 1 (mastery) if

$$\frac{ln(\frac{1-\beta}{\alpha}) - \sum_{j=1}^{k} ln(\frac{1-p_j(\theta_0+\delta_2)}{1-p_j(\theta_0-\delta_1)})}{\delta_1 + \delta_2} \leq \sum_{j=1}^{k} a_j u_j. \tag{5}$$

Otherwise, the TSPRT continues testing and administers another item. At $k = N$, testing is ceased and a classification decision is forced. α and β are constants representing the allowed decision error rates of the two statistical tests. The decision rules at $k = N$ are the ratio tests that are evaluated against the weighted sum. These decision rules are the same as the stopping rules at $k < N$, but without the $ln[\beta/(1 - \alpha)]$ or $ln[(1 - \beta)/\alpha]$ terms. Usually, in CAT, it is optimal to choose items that provide maximum information at the current θ estimate. As a result, the test is adapted to the difficulty that is appropriate for an examinee. In mastery testing, it is optimal to choose items that provide maximum information at the cut point (Eggen & Straetmans, 2000).

The SCSPRT. The TSPRT is inefficient to the extent that there are cases in which it presents another item although the item cannot further change the classification decision about the examinee (Finkelman, 2008). The SCSPRT is an extension of the SPRT (Finkelman 2003, 2004, 2008; Wouda & Eggen, 2009). It adds stochastic curtailment in the form of two extra stopping rules per level. In the cases in which the probability of a change of classification is impossible, the SCSPRT intervenes and halts testing; this 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; this is called *stochastic curtailment.*

This Study

In this study, the performance of the TSPRT and the SCSPRT was compared for different non-statistical constraints. Different item exposure and content balancing methods were explored and combined in order to determine which method or combination of methods would result in the largest performance gain for the SCSPRT.

Method

The Mathematics Item Bank

The different methods for handling constraints and item exposure were evaluated with a simulation study. Data were sampled from a realistic θ distribution, using real item parameters (Eggen & Straetmans, 2000). The items in the bank belonged to one of three content subdomains: mental arithmetic/estimating (A), measuring/geometry (B), and the other elements of the curriculum (C). The items were shown to fit a one-dimensional IRT model. These item parameters originated from a mathematics item bank consisting of 250 calibrated items (48, 49, and 153 belonging to the subdomains A, B, and C, respectively). To satisfy the necessary scaling constraints, the geometric mean of the estimated discrimination indices was fixed at $\left(\prod \hat{a}_j\right)^{1/250}$ = 3.10, and the sum of the estimated difficulty parameters was fixed at $\sum_{j=1}^{N} \hat{b}_j$ = 0. Per constraint, 5,000 simulees were sampled from a normal distribution (mean = 0.294, standard deviation = 0.522) to test the SCSPRT against the SPRT. The selection method was always based on cut scores.

Item Exposure

Three item exposure methods were explored. The Sympson-Hetter method gives a lower probability of administration if an item has a probability of administration that is too high. This method ensures a maximum exposure rate for every item. The alpha-stratified method selects a set of *a* parameters randomly from a distribution of *a* parameters, making sure that items with low *a* parameters will be chosen early in the test. The progressive method adds a random component (which decreases as the test progresses) to every item probability, to ensure that most items get a minimum exposure rate.

Sympson-Hetter method. The general idea behind the Sympson-Hetter (1985) method is that every item gets a maximum exposure rate value and a probability of being administered, given that it is selected. Consequently, the maximum exposure rate is controlled, but the minimum exposure rate is not. This method was implemented by simulating θ values from the distribution as described above, simulating data, and administering the items (following the respective procedures) to simulees. Then we checked which items were administered more than 35% of the time. If items were administered more than 35% of the time their probability of selection was set to 0.2. After 19 iterations with 1,000 simulees, no more items were selected more than 35% of the time, and the Sympson-Hetter procedure was set.

Alpha-stratified method. The alpha-stratified (Chang & Ying, 1999) adaptive testing method ensures that items are chosen according to their discrimination parameters. This way, also items with low discrimination parameters will be chosen early in the test. In the mathematics item bank used in this study, the distribution of *a* parameters was not equal across the possible values (see Table 1). Therefore, in order to prevent a situation in which an *a* parameter was chosen from an "empty" value category (for instance from 6 or 7 when all items from those value categories were already used earlier in the test), we grouped the items with values 2–3 and the items with values 4–7. This way, items could only be selected from one of two groups, either from a group with low *a* parameters, or from a group with high *a* parameters. Items had an equal probability of being selected from either group.

Table 1. Frequencies of Different Values for *a* Parameters in the Mathematics Item Bank

	a Value					
	2	3	4	5	6	7
Freq.	52	115	64	14	4	1

Progressive method. The progressive method (Revuelta & Ponsoda, 1998) adds a random component to the maximum Fisher information. The contribution of this random component is important at the beginning of the test and becomes increasingly less influential as the test progresses. This was implemented by first generating a random value (R) for each item between 0 and $I_{\max}$, the maximum Fisher information. Then the relative serial position of the current item was determined, by $s_j = k/k_{\max}$. It must be noted that k is not the real position of the current item in the test, because the total number of items used in a test is not known prior to the test. Then the weight per item j was computed by means of the formula

$$W_j = (1 - s_j)R + s_j I_j. \tag{6}$$

Subsequently, the weights were ordered descendingly after which the first item in line was administered to the respective simulee.

Content Balancing

Two content balancing methods were explored. The Kingsbury-Zara method chooses the next item in a test according to the largest difference in percentage of content domain use between desired and current. The weighted deviation method treats constraints by choosing items according to the minimal weighted sum of deviations.

Kingsbury-Zara method. With the content balancing method by Kingsbury and Zara (1989, 1991) the idea is that every next item in a CCT is chosen from the item content group with the largest difference in percentage between desired and current. In the current research, there were three content groups (sub-domains), A, B, and C, with respective proportions of 19.2%, 19.6%, and 61.2%. The desired proportions were respectively 16%, 20%, and 64%. The first two items were chosen to be maximally informative at the cut point. Subsequently, the proportions of all content groups were calculated. After that, the difference in proportion between desired and current was calculated, and the most informative item from the content group with the largest difference between desired and current was chosen.

Weighted deviation method. The weighted deviation method (WDM; Stocking & Swanson, 1993; Swanson & Stocking, 1993) handles constraints by choosing items according to the minimal weighted sum of deviations, which was based on a constraint relevancy matrix.

Constraints are seen as desirable properties that do not have to be met strictly. The goal of WDM is to minimize the sum of the weighted deviations. Selecting an item involves three steps (Bakker & Chang, 2009). First, compute the deviation for each of the constraints if the item were added to the test for every item not already in the test. Then sum the weighted deviations across all constraints. Finally, select the item with the smallest weighted sum of deviations.

This was implemented by first subtracting half of the maximum Fisher information $I_{\max}$ from all items' I_j not in the content subdomain with the largest difference in percentage between desired and current. Subsequently the same was done for the other constraint (answer key balancing). Finally all resulting numbers were sorted in descending order and the first item of that row was administered to the respective simulee.

Results

As can be seen in Table 2, from the tested methods for exposure control, with the Sympson-Hetter method the largest gain in efficiency was established for the SCSPRT as opposed to the SPRT (2.27 items). There was no difference found in efficiency gain between the two content balancing methods. When exposure control and content balancing methods were combined, the largest gain in efficiency was found in the combination of the Kingsbury-Zara method and the Sympson-Hetter method (3.07 items). With respect to the use of the item bank, the combination of the weighted deviation method with the progressive method outperformed all other methods. The progressive method used 89% of the item bank (*N* = 250), but the combination of WDM with PM used 100% of the items. However, this result comes at an efficiency cost, because the SPRT as well as the SCSPRT used more items to arrive at a classification decision than with all other explored methods.

Table 2. Results of Different Item Selection Methods With the SPRT and SCSPRT

	SCSPRT		SPRT		Gain	% of Item
Item Selection Method	% Correct	*N*	% correct	*N*	Δ	Bank Used
No constraints	94.90	13.13	94.90	13.46	0.33	16
Sympson-Hetter (SH)	94.00	14.63	93.95	16.90	2.27	42
α stratified (*a*)	94.56	14.30	94.56	16.07	1.77	44
Progressive (PM)	95.54	14.43	95.26	15.01	0.58	89
Kingsbury-Zara (KZ)	95.24	13.16	95.18	13.68	0.52	16
Weighted deviation (WD)	95.00	14.12	95.00	14.88	0.76	32
KZ + SH	94.24	14.33	94.20	17.40	3.07	51
KZ + PM	95.13	15.06	95.15	16.19	1.13	95
WDM + α	95.20	15.89	94.85	17.04	1.15	44
WDM + SH	94.06	16.61	94.16	18.27	1.66	58
WDM + PM	94.95	17.54	94.85	18.28	0.74	100

Conclusions

The Sympson-Hetter method had the largest gain of the three explored exposure control methods. The two content balancing methods performed almost the same. When the two methods were combined, the largest gain was found in the combination Kingsbury-Zara with Sympson-Hetter. The progressive method used most of the item bank, measured over all simulees; however, with this method it remains unclear whether there are large differences in selection frequency. This might be an important issue for item bank security.

References

Bakker, M., & Chang, H. (2009). *Constrained item selection in computerized classification testing.* Unpublished master's thesis, University of Amsterdam.

Chang, H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23,* 211–222.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 2* , 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. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33,* 442–463.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359–375.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4,* 241–261.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 311–327.

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.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17,* 151–166.

Sympson, J. B., & Hetter, R. D. (1985). *Controlling item-exposure rates in computerized adaptive testing.* In Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973–977). San Diego, CA.

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

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

Wouda, J. T. (2008). *Stochastic curtailment to truncate the SPRT for use in computerized classification testing in three categories*. Unpublished master's thesis, University of Amsterdam.