Measurement and Research Department Reports 98-1

Item Selection in Adaptive Testing with the Sequential Probability Ratio Test

T.J.H.M. Eggen

Cito

Measurement and Research Department Reports 98-1

Item Selection in Adaptive Testing with the Sequential Probability Ratio Test

T.J.H.M. Eggen

Cito
Arnhem, 1998

Abstract

Computerized adaptive tests (CATs) were originally developed to obtain an efficient estimate of an examinee's ability. For classification problems, applications of the Sequential Probability Ratio Test (Wald, 1947) have been shown to be a promising alternative for testing algorithms which are based on statistical estimation. However, the method of item selection currently being used in these algorithms, which use statistical testing to infer on the examinees, is either random or based on a criterion which is related to optimizing estimates of examinees (maximum (Fisher) information). In this study, an item selection method based on Kullback-Leibler information is presented, which is theoretically more suitable for statistical testing problems and which can improve the testing algorithm for classification problems.

Simulation studies were conducted for two- and three-way classification problems, in which item selection based on Fisher information and Kullback-Leibler information were compared. The results of these studies showed that the performance of the testing algorithms with Kullback-Leibler information-based item selection are sometimes better and never worse than algorithms with Fisher information-based item selection.

Introduction

Efficient estimation of the ability of an examinee is often the purpose of computerized adaptive testing (CAT). But, if the goal of testing is to classify examinees in a limited number of categories, for example, pass/fail decisions on an exam or decisions regarding placement in courses on different levels, CATs can make use of algorithms that are based on statistical testing rather than statistical estimation. Studies by Reckase (1983), Lewis and Sheenan (1990), and Spray and Reckase (1994, 1996), for decisions in two categories, and Eggen and Straetmans (1996), for decisions in three categories, have shown that the Sequential Probability Ratio Test (SPRT) (Wald, 1947) can be successfully applied in adaptive testing using an item response theory (IRT) calibrated item bank.

An important part of a CAT algorithm is the item selection procedure, which determines, during testing, the choice of the items which are administered. In current adaptive tests using statistical testing in the algorithm, item selection is based on a criterion which is 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. In this article, item selection procedures will be proposed that are based on Kullback-Leibler information (Cover & Thomas, 1991). It will be shown that the item Kullback-Leibler information expresses the expected contribution of an item to the discriminatory power between two hypotheses. Conceptually, K-L information fits the statistical testing algorithm more closely than Fisher information. One of the questions addressed in this article is whether using K-L information has a positive impact on the performance of the adaptive tests with statistical testing for decision problems with two categories and with three categories. Bayesian item selection criteria are also in use in adaptive testing with estimation. These criteria, recently discussed by Van der Linden (1996), will not be considered in this article.

The first part of the article is an overview of the SPRT application in problems with two and three categories. Next, item selection based on both Fisher and Kullback-Leibler information will be presented. Finally, a comparison of the item selection procedures for both the two- and three- category problem will be made on the basis of simulation studies with an operational item bank.

Sequential Testing in the Testing Algorithm

In testing algorithms of adaptive tests, the likelihood function of an examinee's ability, θ, plays a central role in the inference on the examinee. Assuming that an IRT calibrated item bank is available, that is, the parameters of the items can be considered to be known, and given the scores on k items $(x_i, i = 1, \ldots, k)$, this function is

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

In this likelihood function, $p_i(\theta)$, the probability of answering item i correctly, is the item response function belonging to an IRT model. In this paper, the two-parameter logistic (2-PL) model is used:

$$p_i(\theta) = P(X_i = 1 \mid \theta) = \frac{\exp a_i(\theta - b_i)}{1 + \exp a_i(\theta - b_i)}. \quad (2)$$

The response to an item x_i is either correct (1) or incorrect (0). The probability of a correct response increases with the latent ability θ and depends on two item characteristics: the difficulty parameter, b_i, and the discrimination parameter, a_i.

In traditional adaptive tests, the ability is estimated after each item by maximizing the likelihood function with respect to θ. When statistical testing rather than estimation is used in the testing algorithm, the likelihood function is used somewhat differently, which will become clear in the following description of the statistical testing procedure.

Classification in Two Categories

On the latent ability scale, a decision or cutting point θ_0 between, for example, a master and non-master, or between an examinee who passes and an examinee who fails on an exam, is given. A small region on both sides of this point, a so-called indifference zone, is selected. The width of these regions, although they could differ from each other, is taken to be δ. 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 of the classification of the examinees who are that near to the cutting point.

Next, the statistical hypotheses are formulated:

$$\text{H0: } \theta \leq \theta_0 - \delta = \theta_1 \text{ against H1: } \theta \geq \theta_0 + \delta = \theta_2. \tag{3}$$

Acceptable decision error rates are specified as follows:

$$\text{P(accept H0}\mid \text{ H0 is true)} \geq 1 - \alpha, \text{ and P(accept H0}\mid \text{ H1 is true)} \leq \beta, \tag{4}$$

in which α and β are small constants. The test meeting these decision error rates can be carried out using the SPRT (Wald, 1947). The test statistic used is the ratio between the values of the likelihood function (Equation 1) under the alternative hypothesis and the null hypothesis:

$$LR_k(\underline{x}) = \frac{L_k(\theta_2 ; \underline{x})}{L_k(\theta_1 ; \underline{x})}. \tag{5}$$

It will be clear that high values of this ratio indicate the examinee is more likely to have an ability above the cutting point, and small values support the decision that the examinee is below the cutting point. That is, the test meets the error rates if the following procedure is used:

Continue sampling if: $$\beta/(1-\alpha) < LR_k(\theta_2, \theta_1 ; \underline{x}) < (1-\beta)/\alpha; \tag{6}$$

accept H0 if: $$LR_k(\theta_2, \theta_1 ; \underline{x}) \leq \beta/(1-\alpha); \tag{7}$$

reject H0 if: $$LR_k(\theta_2, \theta_1 ; \underline{x}) \geq (1-\beta)/\alpha. \tag{8}$$

Equation 6 is called the critical inequality of the test. It can easily be shown that if the 2-PL model (Equation 2) is used, the critical inequality of this test can be written as follows:

$$\frac{\ln\beta/(1-\alpha) - C_{k\theta_1\theta_2}}{\theta_2 - \theta_1} < \sum_{i=1}^{k} a_i x_i < \frac{\ln(1-\beta)/\alpha - C_{k\theta_1\theta_2}}{\theta_2 - \theta_1}. \tag{9}$$

In Equation 9,

$$C_{k\theta_1\theta_2} = \sum_{i=1}^{k} \ln \frac{1+\exp a_i(\theta_1 - b_i)}{1+\exp a_i(\theta_2 - b_i)} = \sum_{i=1}^{k} \ln \frac{1-p_i(\theta_2)}{1-p_i(\theta_1)} = \sum_{i=1}^{k} \ln \frac{q_i(\theta_2)}{q_i(\theta_1)}, \tag{10}$$

which only depends on the item parameters and on constants in the statistical testing procedure, θ_1 and θ_2, that are chosen beforehand. The evaluation of the critical inequality is quite easy because it involves only the observed weighted score and known constants. Note that because $\theta_2 > \theta_1$, $q_i(\theta_2) / q_i(\theta_1) < 1$ and thus $C_{k\theta_1\theta_2} < 0$. Furthermore, if the indifference interval $2\delta = \theta_2 - \theta_1$ increases, the width of the critical interval gets smaller, which indicates that shorter tests can be used to make a decision.

Although Wald (1947) has shown that eventually a decision will be taken with probability 1 with the SPRT, in practice, a maximum test length, k_{max}, is usually specified. At this test length, a forced decision is taken. In that case, the most obvious decision is taken: H0 is rejected if the test statistic is larger than the midpoint of the critical inequality interval; otherwise it is accepted.

Classification in Three Categories

The above testing procedure is readily generalized to cases of classification in one of three categories. In this case, there are two cutting points, θ_1 and θ_2, and three different levels of ability are distinguished. See Figure 1.

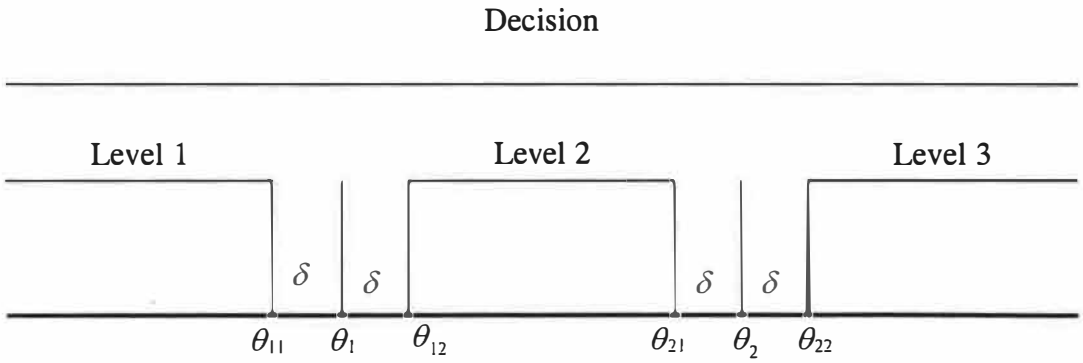

Figure 1

Schematic representation of the classification problem with three categories

After selecting the indifference zones, all taken to be δ, two pairs of hypotheses are formulated:

H0_1: $\theta \leq \theta_{11} = \theta_1 - \delta$ (level 1) H0_2: $\theta \leq \theta_{21} = \theta_2 - \delta$ (lower than 3) (11)

H1_1: $\theta \geq \theta_{12} = \theta_1 + \delta$ (higher than 1); H1_2: $\theta \geq \theta_{22} = \theta_2 + \delta$ (level 3). (12)

The SPRT test described in the preceding section is applied for each pair of hypotheses. In the specification of the acceptable decision errors, as in Equation 4, the small constants α_1 and β_1, α_2 and β_2, respectively are used.

If the 2-PL model is used, the critical inequalities of the tests are

$$L_1 = \frac{\ln \frac{\beta_1}{1-\alpha_1} - \sum_{i=1}^{k} \ln \frac{q_i(\theta_1+\delta)}{q_i(\theta_1-\delta)}}{2\delta} < \sum_{i=1}^{k} a_i x_i < \frac{\ln \frac{1-\beta_1}{\alpha_1} - \sum_{i=1}^{k} \ln \frac{q_i(\theta_1+\delta)}{q_i(\theta_1-\delta)}}{2\delta} = U_1 \quad (13)$$

$$L_2 = \frac{\ln \frac{\beta_2}{1-\alpha_2} - \sum_{i=1}^{k} \ln \frac{q_i(\theta_2+\delta)}{q_i(\theta_2-\delta)}}{2\delta} < \sum_{i=1}^{k} a_i x_i < \frac{\ln \frac{1-\beta_2}{\alpha_2} - \sum_{i=1}^{k} \ln \frac{q_i(\theta_2+\delta)}{q_i(\theta_2-\delta)}}{2\delta} = U_2. \quad (14)$$

It can easily be checked that $\partial (\ln q_i(\theta + \delta)) / (q_i(\theta - \delta)) / \partial\theta < 0$ for all θ, which means it is decreasing in θ. A consequence of this is that the lower bound of test 1, L_1, can never be larger than the upper bound of test 2, U_2. So, by combining the decisions of the simultaneously applied two SPRTs, unequivocal decisions can be made in the three-way classification problem.

Decisions based on a combination of two SPRTs:

	decision test 1	
decision test 2	1	2 or 3
1 or 2	1	2
3	impossible	3

This generalization of the SPRT, known as Sobel and Wald's (1949) combination procedure, performs as well as Armitage's (1950) combination procedure, which is applied by Spray (1993) for classification in three and even k categories.

The procedure for the combined test with the 2-PL model is as follows:

take decision 1 if	$\sum_{i=1}^{k} a_i x_i \leq L_1;$	(15)
take decision 2 if	$U_1 \leq \sum_{i=1}^{k} a_i x_i \leq L_2;$	(16)
take decision 3 if	$\sum_{i=1}^{k} a_i x_i \geq U_2;$	(17)
continue testing if	else.	(18)

A sketch of the procedure is given in Figure 2:

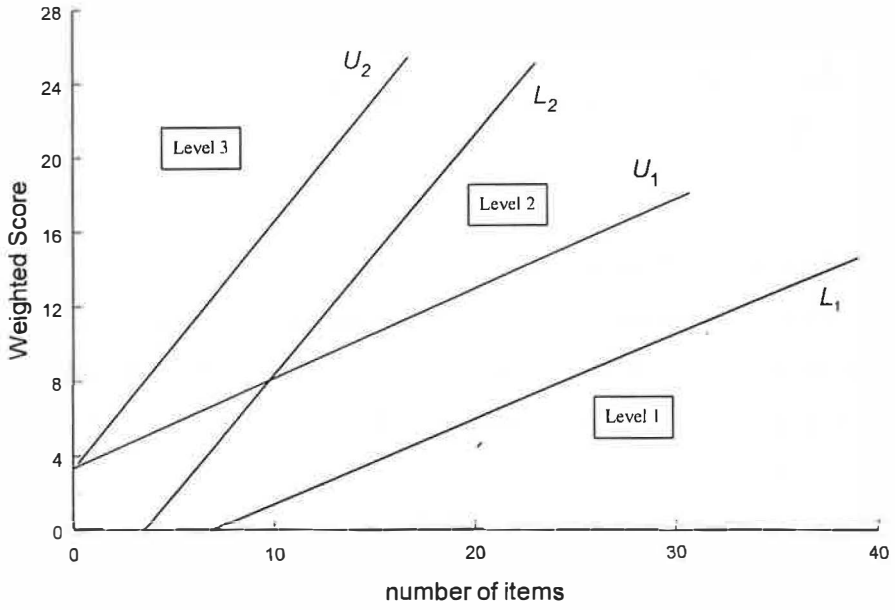

Figure 2

Sketch of the statistical test procedure with three levels

Note that, of course, $L_1 < U_1$ and $L_2 < U_2$, but it generally requires some items before $U_2 < L_1$ can be true and decision 2 can be taken.

Item Selection

In the testing algorithm, the item selection procedure chooses items from the item bank. In connection with the use of the SPRT, random selection is a possibility, but it is well known that the efficiency is much greater when a maximum information selection strategy is applied (see, for example, Eggen & Straetmans, 1996). Information usually means Fisher information which will be described first.

In adaptive testing in which the aim is estimating the ability of an examinee and items are selected to have maximum information at the current ability estimate, Chang and Ying (1996) introduced an item information measure which is not based on Fisher information but on Kullback-Leibler information (K-L information). In this section, an information measure will be introduced that is also based on Kullback-Leibler information but which is more suitable to be used in connection with adaptive testing with the statistical testing with the SPRT.

Some item selection procedures for both Fisher information and Kullback-Leibler information will be given for both the two- and three-categories problem.

Fisher Information

Current maximum information item selection procedures are almost all based on the Fisher item information, which for an item i is defined as

$$\mathrm{I}_i(\theta) = \mathscr{E}\left[\frac{\frac{\partial}{\partial\theta}L(\theta\,;x_i)}{L(\theta\,;x_i)}\right]^2. \tag{19}$$

In the 2-PL, the expression is given by

$$\mathrm{I}_i(\theta) = a_i^2\, p_i(\theta)\, q_i(\theta). \tag{20}$$

For a test consisting of k items, the test information is the sum of the information of the items in the test: $\mathrm{I}(\theta) = \Sigma_{i=1}^{k}\,\mathrm{I}_i(\theta)$. Selecting items with maximum information maximizes the contribution to the test information. The usefulness of this is readily understood if an estimate of the ability of an examinee is wanted, especially when the maximum likelihood estimator (MLE) is used. In this case, $\hat{\theta}_k$, the MLE after taking k items, follows from $\max_{\theta} \prod_{i=1}^{k} L(\theta\,;x_i)$ and the standard error of this estimator is estimated by $\mathrm{se}(\hat{\theta}_k) = 1 / \sqrt{\mathrm{I}(\hat{\theta}_k)}$. So, it can be seen that by selecting items having

maximum information, the contribution to the decrease of the standard error is greatest. Furthermore, from the definition in Equation 19, it can be seen that maximizing the information is the same as maximizing the contribution of an item to the expected relative rate of change of the likelihood function. As Chang and Ying (1996) point out, the greater this change rate at a given value of θ, the better it can be distinguished from points near to this value, and the better this value can be estimated.

Some selection procedures based on Fisher information

F1 In adaptive tests in which the examinee's ability is to be estimated, the most popular item selection method is to select the item that has maximum information at the current ability estimate:
Select the item i for which: $\max_i I_i(\hat{\theta}_k)$.

F2 Spray and Reckase (1994) have shown that in a classification problem with two categories for which the SPRT procedure is being used, it is more efficient to select the items which have maximum information at the cutting point θ_0 rather than at the current ability estimate:
Select the item i for which: $\max_i I_i(\theta_0)$.

F3 In a three-way classification problem for which the generalized SPRT described in the preceding section is being used, an alternative selection method could be to select the item which maximizes the information at the cutting point nearest to the current estimate:
Determine $\min(|\theta_1 - \hat{\theta}_k|, |\theta_2 - \hat{\theta}_k|)$ and choose the item with maximum information at the cutting point with the minimum.

F4 For the three-way classification problem with the SPRT in which no use is made of estimates of abilities, Eggen and Straetmans (1996) propose selecting the item which maximizes the information at the cutting point corresponding to the midpoint of the critical inequality interval which is closest to the current examinee's score. If the 2-PL is the IRT model, the midpoints follow from the Equations 19 and 20; after determining the minimum of $|\sum_i a_i x_i - (U_1 + L_1)/2|$ and $|\sum_i a_i x_i - (U_2 + L_2)/2|$, the item with maximum information at the corresponding cutting point will be selected.

Kullback-Leibler Information

The item selection methods described in the preceding section all use a criterion related to Fisher information that has a strong relation to optimizing estimates. Although these

selection methods can also be used in adaptive testing with the SPRT, one could wonder whether making use of Fisher information is optimal in this case. An alternative could be to base the item selection process on the relative entropy or Kullback-Leibler information (Cover & Thomas, 1991), which is an information concept as strongly related to statistical testing as Fisher information is to statistical estimation. The relative entropy is a measure of the discrepancy between two distributions:

$$\mathrm{K}(f_1 \,||\, f_0) = \mathscr{E}_{f_1} \log \left[\frac{f_1(x)}{f_0(x)} \right], \tag{21}$$

which is the expected information in x for discrimination between the two hypotheses H0: $f(x) = f_0(x)$ and H1: $f(x) = f_1(x)$. The larger this information, the more efficient the statistical test will be.

The definition in Equation 21 can be directly applied to the SPRT application in adaptive testing: H0 is the hypothesis that we have a distribution (likelihood) with parameter value $\theta = \theta_1$ and under H1 the distribution has parameter $\theta = \theta_2$. And

$$\mathrm{K}(\theta_2 \,||\, \theta_1) = \mathscr{E}_{\theta_2} \log \left[\frac{L_k(\theta_2 ; \underline{x})}{L_k(\theta_1 ; \underline{x})} \right], \tag{22}$$

is the Kullback-Leibler test information (k items), which can be written as the sum of the Kullback-Leibler information of the items:

$$\mathrm{K}(\theta_2 \,||\, \theta_1) = \sum_{i=1}^{k} \mathrm{K}_i(\theta_2 \,||\, \theta_1) = \sum_{i=1}^{k} \mathscr{E}_{\theta_2} \log \left[\frac{L(\theta_2 ; x_i)}{L(\theta_1 ; x_i)} \right]. \tag{23}$$

The K-L item information $\mathrm{K}_i(\theta_2 \,||\, \theta_1)$ is defined for any pair θ_2 and θ_1 and is a positive real number and, consequently, an eligible item information index. The usefulness of applying an item selection procedure based on maximum K-L information can be understood, since this procedure will maximize the contribution to the K-L test information. When the K-L test information is maximized, it is expected that the difference between the likelihood under both hypotheses is maximized, which is, in

turn, expected to minimize the number of items needed to take a decision because the test statistic is the likelihood ratio (see Equation 5).

K-L information is also the basis for an index proposed by Chang and Ying (1996) for estimation problems as a more global information index in contrast to the local Fisher information. They consider, for any θ, the K-L item information to the true ability θ_0: $\mathrm{K}_i(\theta_0 \,||\, \theta)$, which is then, of course, a function of θ. They define their information index which is used in item selection as an integral of this function over an interval depending on the current MLE, $\hat{\theta}_k$, and a expression, δ_k, which is decreasing in the number of items (k):

$$K_i\left(\hat{\theta}_k\right) = \int_{\hat{\theta}_k - \delta_k}^{\hat{\theta}_k + \delta_k} \mathrm{K}_i\left(\hat{\theta}_k \,||\, \theta\right) d\theta. \tag{24}$$

Chang and Ying's (1996) claim is that their information measure is a good alternative, especially in the beginning of the test, when the ability of an examinee is poorly estimated. It should be noted that information indices like the one given in Equation 24, but then based on Fisher information, were also proposed by Veerkamp and Berger (1997). But, because these indices are not expected to be useful alternatives for item selection in the case of the SPRT, they will not be discussed further in this article.

If an IRT model for dichotomously scored items is used, the K-L item information index can be written as:

$$\mathrm{K}_i(\theta_2 \,||\, \theta_1) = p_i(\theta_2) \log \frac{p_i(\theta_2)}{p_i(\theta_1)} + q_i(\theta_2) \log \frac{q_i(\theta_2)}{q_i(\theta_1)}, \tag{25}$$

which with the 2-PL model specializes to:

$$\mathrm{K}_i(\theta_2 \,||\, \theta_1) = a_i(\theta_2 - \theta_1) p_i(\theta_2) + \ln \frac{q_i(\theta_2)}{q_i(\theta_1)}. \tag{26}$$

Note that the index is linear in the discrimination parameter, whereas the Fisher item information is quadratic in a_i, which means that the weight of the discrimination parameter in the selection is less, which can be favorable in the beginning of the test.

If the K-L information is computed in $\theta_2 = \theta_0 + \delta$ and $\theta_1 = \theta_0 - \delta$, Equation 26 becomes

$$K_i(\theta_0 + \delta \,||\, \theta_0 - \delta) = \frac{2a_i\delta \exp a_i(\theta_0 + \delta - b_i)}{1 + \exp a_i(\theta_0 + \delta - b_i)} + \ln \frac{1 + \exp a_i(\theta_0 - \delta - b_i)}{1 + \exp a_i(\theta_0 + \delta - b_i)}, \qquad (27)$$

which is a monotone increasing function of δ. This means that for any fixed item i, the K-L item information increases if the width of the indifference zones increases. This property illustrates that the K-L item information expresses the contribution of an item to contribute to the capability to distinguishing between two hypotheses, which is larger when δ is larger. However, this property does not imply that the order of the item K-L information over items is the same for each δ.

Some selection procedures based on K-L information

K1 In the case of a classification problem in two categories, the K-L item information can be used directly in a straightforward way in item selection. The K-L item information will be computed in two points symmetric around the cutting point:
Select the item i for which: $\max_i K_i(\theta_0 + \delta \,||\, \theta_0 - \delta)$.

K2 In the three-way classification for K-L item selection, there are more possibilities. One possibility is to select the item which maximizes the K-L information at two fixed points. Possible choices are (see Figure 1): a. $K_i(\theta_{21} \,||\, \theta_{12})$, b. $K_i(\theta_2 \,||\, \theta_1)$ and c. $K_i(\theta_{22} \,||\, \theta_{11})$, which have in common that 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.

K3 One way to deal with this problem is, as with Fisher information (see F3 before), is to look for the nearest cutting point and to select the items with maximum K-L information around this cutting point. The nearest cutting point is, as in F4, determined without estimation by comparison of the score with the midpoints of the critical intervals of the tests.

K4 An alternative approach is to look more precisely at the progress of hypothesis testing: as long as none of the pairs of hypotheses have led to a decision, items are chosen with maximum K-L information between the two cutting points θ_1

and θ_2; if one of the pairs of hypotheses has led to a decision while the other has not, items will be chosen which have maximum K-L information around the cutting point corresponding to the test which has not yet led to a decision. If the 2-PL model is used, the following selection procedure is used:

$$\text{if: } \sum_{i=1}^{k} a_i x_i \geq U_1: \qquad \max_i \mathrm{K}_i(\theta_{22} \,||\, \theta_{21}) \tag{28}$$

$$\text{if: } \sum_{i=1}^{k} a_i x_i \leq L_2: \qquad \max_i \mathrm{K}_i(\theta_{12} \,||\, \theta_{11}) \tag{29}$$

$$\text{else: } \max_i \mathrm{K}_i(\theta_2 \,||\, \theta_1). \tag{30}$$

A small variation on this procedure could be made in case no decision has been taken yet: instead of the expression in Equation 30, a narrower interval is chosen:

$$\max_i \mathrm{K}_i(\theta_{21} \,||\, \theta_{12}). \tag{31}$$

Comparison of Item Selection Procedures

The performance of the item selection procedures were compared by means of a simulation study drawing on an operational item bank. The bank contains 250 mathematics items which are used in adult education to place students in one of three course levels and to measure progress at these levels. The items were calibrated with the 2-PL model. On the scale fixed by restrictions on the item parameters, the distribution of the ability θ in the population was estimated to be normal with a mean of .294 and a standard deviation of .522. More details on the scaling can be found in Eggen and Straetmans (1996).

The simulations were conducted as follows. An ability of a simulee θ_v was randomly drawn from $N(.294, .522)$. Three relatively easy starting items were selected; subsequent items were selected using one of the item selection methods. 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 decision error rates, the item selection

procedures were compared on the mean number of items required to make a decision and the classification accuracy, the percentages of correct decisions.

Classification in Two Categories

The cutting point in the simulation was $\theta_0 = .1$, and the maximum test length was: $k_{max} = 40$. The procedure was conducted for three different error rates: $\alpha = \beta$ were .05, .075 and .1 and varying indifference zone: $.1 \leq \delta \leq .23$, in steps of .01.

The following item selection procedures were compared:

F1 Maximum Fisher information at the current estimate.

F2 Maximum Fisher information at the cutting point.

K1a Maximum K-L information at $\theta_1 = .05$ and $\theta_2 = .15$.

K1b Maximum K-L information at $\theta_1 = .00$ and $\theta_2 = .20$.

K1c Maximum K-L information at $\theta_1 = -.05$ and $\theta_2 = .25$.

The results for a typical indifference zone $\delta = .15$ are given in Table 1.

Table 1

Mean number of required items and percentage of correct decisions in a decision problem with one cutting point $\theta_0 = .1$ and with $\delta = .15$

	Selection method									
	F1		F2		K1a		K1b		K1c	
Error rate	k	%	k	%	k	%	k	%	k	%
$\alpha=\beta=.05$	16.0	95.6	16.3	94.7	16.1	95.4	16.3	94.6	15.9	95.5
$\alpha=\beta=.075$	14.9	95.0	14.0	95.2	13.9	94.8	13.9	95.2	13.9	95.6
$\alpha=\beta=.10$	13.2	94.9	12.7	94.8	13.2	94.8	12.7	95.3	12.9	94.8

Most notable is that there are almost no differences in Table 1. For the three error rates and all five item selection methods, the percentage of correct decisions are about 95%. A consistent difference between these error rates over selection methods can be seen in the mean number of required items: the lower the rates, the more items are needed. There are hardly any differences between the three variants of K-L information selection. There seems to be a slight tendency, at least when the error rates are .075

and .10, for the selection of items with maximum Fisher information at the cutting point (F2) to be better than the selection of items with maximum Fisher information at the current estimate (F1). This is in line with the findings of Spray and Reckase (1996). Furthermore, K-L information selection (K1) seems to be as good as selecting with maximum Fisher information at the cutting point (F2).

These results are also found if the indifference zone δ is varied. The results for $\alpha = \beta = .1$ for the three selection procedures F1, F2 and K1 are given in Figures 3 and 4.

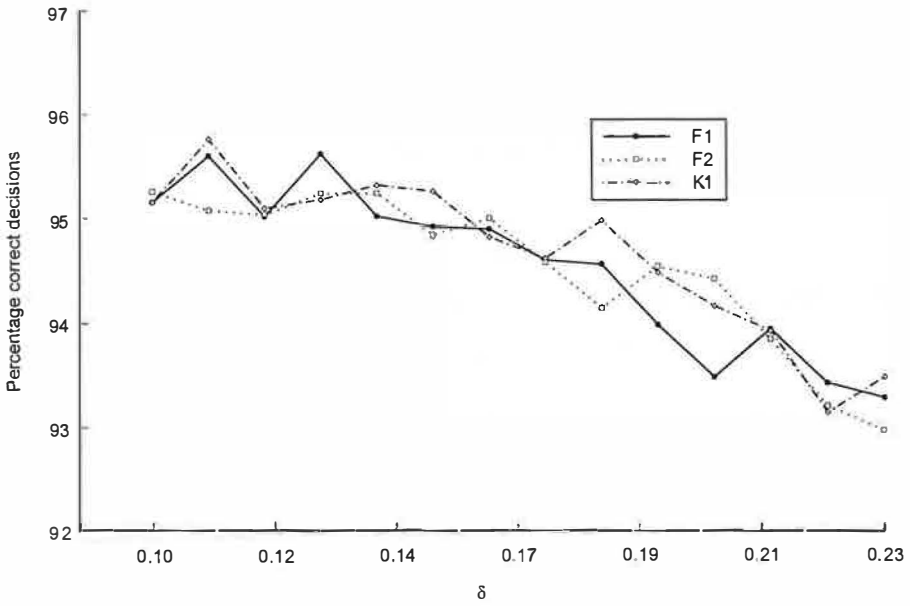

Figure 3

Percentage correct decisions for three item selection procedures in the two category problem as a function of δ

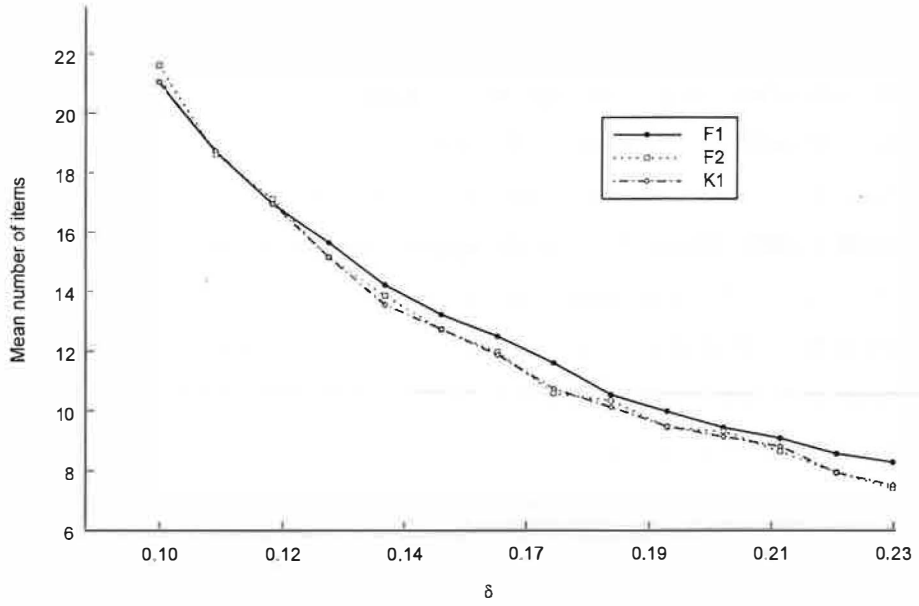

Figure 4

Mean number of required items for three item selection procedures in the two category problem as a function of δ

A slight decrease in the percentages of correct decisions with increasing δ is seen in Figure 3, but there are hardly differences between the item selection methods.

Figure 4 shows that, as expected, the mean number of items required decreases as the indifference zone increases for all three selection procedures. Furthermore, it can be seen that the K-L information item selection and maximum Fisher information selection at the cutting point is as good as and, from $\delta > .12$, a bit better than selection with Fisher information at the current estimate.

Classification in Three Categories

The cutting points in the simulation were $\theta_0 = -.13$ and $\theta_0 = .33$, and the maximum test length was: $k_{max} = 25$. The procedure was conducted for three different sets of error rates: $\alpha_1 = \beta_2 = 2\beta_1 = 2\alpha_2$ were .05, .075, and .1. Halving β_1 and α_2 compared to β_2 and α_1 has the effect that it is expected that all three decisions will have the same error rate. The width of the indifference zones was also varied: $.10 \leq \delta \leq .20$, in steps of .01. No δ larger than .2 was considered, as the zones of both hypotheses would then overlap.

The following item selection procedures were compared:

F1 Maximum Fisher information at the current estimate.

F3 Maximum Fisher information at the cutting point nearest to the current estimate.

F4 Maximum Fisher information at the nearest cutting point.

K2a Maximum K-L information at $\theta_1 = -0.03$ and $\theta_2 = 0.23$.

K2b Maximum K-L information at $\theta_1 = -0.13$ and $\theta_2 = 0.33$.

K2c Maximum K-L information at $\theta_1 = -0.23$ and $\theta_2 = 0.43$.

K3 Maximum K-L information at the nearest cutting point.

K4a Maximum K-L information at varying points: Equations 28, 29, and 30.

K4b Maximum K-L information at varying points: Equations 28, 29, and 31.

The results for $\delta = .13$ are given in Table 2. It is seen that for every selection method, there is an expected decrease in the mean number of required items if the acceptable error rates are increased. Increasing the error rates has little effect on the percentages of correct decisions.

Table 2

Mean number of required items and percentage of correct decisions in a decision problem with two cutting points $\theta_1 = -.13$, $\theta_2 = .33$ and with $\delta = .13$

	error rates					
	.05		.075		.1	
selection	k	%	k	%	k	%
F1	16.7	89.9	15.6	89.2	14.6	89.1
F3	21.8	87.0	20.5	87.7	19.4	87.4
F4	16.8	89.6	15.6	90.0	14.3	88.5
K2a	18.7	89.5	17.4	89.5	16.3	88.1
K2b	18.4	88.4	17.0	88.0	16.3	88.6
K2c	18.7	87.9	17.1	88.2	16.4	88.6
K3	16.8	90.1	15.3	89.2	14.2	89.4
K4a	17.0	89.2	15.6	89.2	14.4	89.4
K4b	17.0	89.2	15.5	89.7	14.2	89.1

A comparison of the selection methods shows that the differences between them are consistent over the different error rates. Next, it is noted that with the K-L selection methods K2 and K4, varying the exact pair of points between which the K-L information is computed has no impact on the performance of the adaptive test. Clearly, the worst performing selection method is the one in which items are selected with maximum Fisher information at the cutting point nearest to the current estimate (F3). It needs more items and has a lower percentage of decisions. This finding, confirming those of Eggen and Straetmans (1996), may be explained by the fact that the current estimate of the ability, especially in the beginning of the test, is so inexact that it is sometimes nearer to the wrong cutting point than the cutting nearest to the true value of the ability. It is also clear that in decision problems with three categories, item selection which maximizes the K-L information at two fixed points (K2) is worse than other methods. There seem to be four (F1, F4, K3, K4) selection methods in the three-category problem that perform almost equally well.

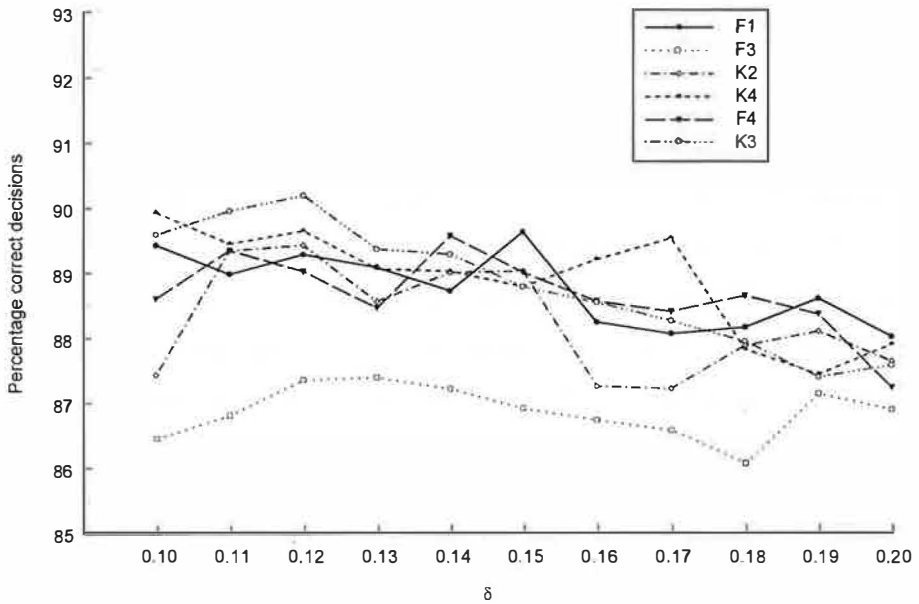

Figure 5

Percentage correct decisions for six item selection procedures in the two category problem as a function of δ

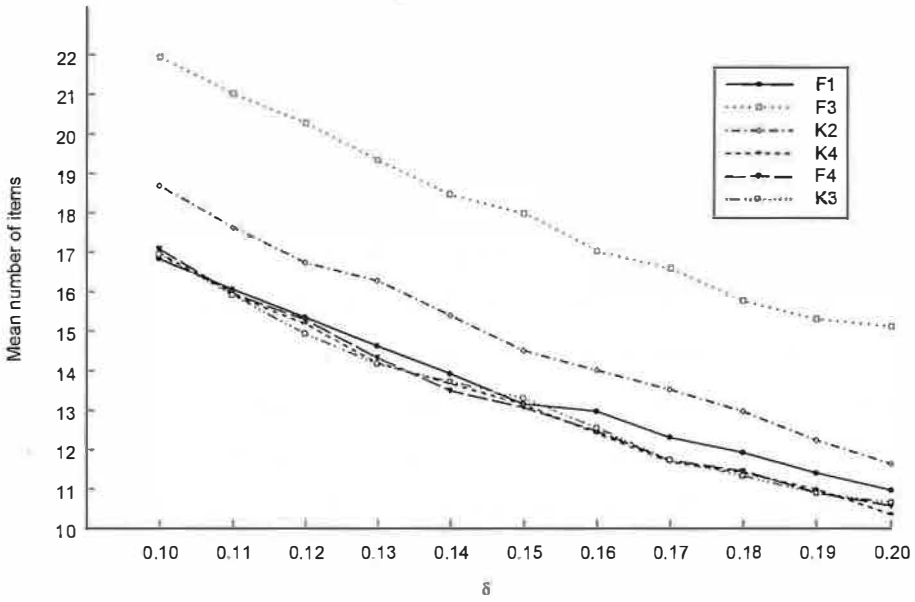

Figure 6

Mean number of required items for six item selection procedures in the two category problem as a function of δ

In Figures 5 and 6, the simulation results are shown as a function of the indifference zone. The results mentioned before are independent of the width of the indifference zone. In Figure 5, there are systematically lower percentages of correct decisions if the selection is based on maximizing the Fisher information at the cutting point nearest to the current estimate (F3). In Figure 6, the expected decrease in the mean number of required items with increasing δ is again seen for all selection methods. On this aspect, the F3 method and the K2 methods (choosing items with maximum K-L information at two fixed points), clearly performed worse than the other four methods. Of these four methods, selecting items which maximum Fisher information at the current estimate (F1), with some indifference zones δ, needs, on average, slightly more items than the K4, F3, and K3 methods, which could be a reason to prefer one of these methods of item selection. Of these three methods, F4 and K3 have in common the way the nearest cutting point is sought: the current weighted score (in the 2-PL) is compared with the midpoints of the critical interval of the two tests. This is an ad hoc criterion which is not based on a clear concept. Nevertheless, the performance of these selection methods is as good as the conceptually better grounded K4 selection method.

Conclusion

The results of the present study indicate that when the sequential probability ratio test is applied in adaptive testing, item selection methods can be defined which are based on an information concept which has a natural relation with hypothesis testing. These item selection methods are based on Kullback-Leibler information or relative entropy, which expresses the power of an item to discriminate between two hypotheses. For decision problems in two and three categories, item selection methods based on K-L information were given as an alternative for item selection methods which are based on the 'estimation-related' Fisher information.

The comparison of the performance of the item selection methods in the decision problem with two categories, showed there was no difference between maximizing K-L information around the cutting point and maximizing Fisher information at the cutting point, but both are slightly better than maximizing Fisher information at the current estimate of an examinee. In the decision problem with three categories, one of the best performing item selection methods was the selection method which maximizes the K-L information between two varying hypotheses. The hypotheses to be considered depend on the progress of the testing thus far: if testing one of the two pairs of hypotheses has led to a decision, items are chosen with maximum information at the other pair of hypotheses; if none has reached a decision, the information is maximized between the two pairs of hypotheses.

In SPRT adaptive testing item selection based on the conceptually strongly related K-L information is generally preferred to Fisher information-based methods. In both the two- and three-category decision problem, the item selection based on K-L information never performed worse and sometimes better than Fisher information-based selection in the simulation study. Moreover, in some of the Fisher information-based item selection methods, an estimate of the current ability is needed. This is never the case in K-L information item selection which is computationally much easier.

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.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

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

Eggen, T.J.H.M., & Straetmans, G.J.J.M. (1996). *Computerized adaptive testing for classifying examinees into three categories*. Measurement and Research department reports, 96-3. Arnhem: Cito.

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* (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.

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

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. (1996). *Bayesian item selection criteria for adaptive testing*. (Research Report 96-01). Enschede: University of Twente, Faculty of Educational Science and Technology.

Veerkamp, W.J.J., & Berger, M.P.F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-226.

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