*Psicológica (2010), 31,*149-169.

Bayesian Item Selection in Constrained Adaptive Testing Using Shadow Tests

Bernard P. Veldkamp[1]

Research Center of Examination and Certification

University of Twente (The Netherlands)

Application of Bayesian item selection criteria in computerized adaptive testing might result in improvement of bias and MSE of the ability estimates. The question remains how to apply Bayesian item selection criteria in the context of constrained adaptive testing, where large numbers of specifications have to be taken into account in the item selection process. The Shadow Test Approach is a general purpose algorithm for administering constrained CAT. In this paper it is shown how the approach can be slightly modified to handle Bayesian item selection criteria. No differences in performance were found between the shadow test approach and the modified approach. In a simulation study of the LSAT, the effects of Bayesian item selection criteria are illustrated. The results are compared to item selection based on Fisher Information. General recommendations about the use of Bayesian item selection criteria are provided.

One of the basic ideas in CAT is to adapt item selection to the examinee's ability in order to measure as precise as possible. Several selection rules can be applied to guide this adaptive item selection process. The most well-known item selection criterion is maximum Fisher information. It is applied in almost all large-scale CAT programs. This criterion can be implemented straightforwardly. Moreover, it has been shown that the value of the Fisher Information for the whole test is asymptotically equal to the inverse of the variance of the ability estimator.

In the literature several alternatives have been described. The main reason for developing alternatives is that maximum Fisher information selects items that perform optimal at the current ability estimate. When the first items in the CAT procedure are presented to the examinee, the variability in the ability estimate is still considerable large. Therefore,

[1] Request for information can be send to: Bernard P. Veldkamp, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. E-mail: b.p.veldkamp@utwente.nl

maximum Fisher Information might select items that perform optimally at the wrong ability level during the first few iterations.

One way to overcome this problem is to apply an interval information measure. In Veerkamp & Berger (1997), Fisher Information was integrated over a small interval around the ability estimate in order to correct for the uncertainty in the estimate. Chang & Ying (1996) applied Kullback-Leibler information instead of Fisher information to select the next item. The performances of both alternatives were compared to the performance of Maximum Fisher information criterion both for dichotomous and polytomous item selection and only small improvements in measurement precision were obtained for very short tests. For larger tests, no significant differences were found.

In this paper, the focus is on Bayesian alternatives. These criteria generally take into account the posterior distribution of the examinee's ability for selecting the next item. In van der Linden (1998a), several Bayesian item selection criteria were introduced and promising results were obtained. Application of these criteria resulted in substantial improvement of both MSE and Bias. About the statistical properties of these criteria, it can be remarked that when an informative prior is used, 'inward bias' of estimators of the ability parameter often occurs for shorter tests. On the other hand, the use of an informative prior usually results in a favorable mean-squared error. Asymptotically, no differences between Bayesian criteria and maximum Fisher information exist. How the actual performance turns out in practice depends on many variables, for example the choice of prior, the test length, and the item bank.

The topic of this paper is how to apply Bayesian item selection criteria in the context of constrained adaptive testing. Therefore, several Bayesian item selection criteria are introduced. Application of these criteria in the context of constrained CAT with shadow tests is described. A modified shadow test approach is presented. Two simulation studies were carried out. General recommendations about the use of Bayesian item selection criteria are given.

BAYESIAN ITEM SELECTION CRITERIA

As mentioned before, Bayesian item selection criteria take the posterior distribution of the ability parameter into account. The posterior distribution can be calculated using Bayes theorem

$$f(\theta \mid \mathbf{u}) = L(\theta;\mathbf{u})\frac{f(\theta)}{f(\mathbf{u})}, \qquad (1)$$

where $L(\theta;\mathbf{u})$ is the likelihood associated with response vector $\mathbf{u}$, $f(\theta)$ is a prior distribution for θ, and $f(\mathbf{u})$ is the marginal probability of response vector $\mathbf{u}$ that serves as a normalizing constant in (1).

From the assumption of local independence it follows that after administration of (k-1) items

$$L(\theta;\mathbf{u}_{k-1}) = \prod_{i=1}^{k-1} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i}, \qquad (2)$$

where $P_i(\theta)$ is the probability of a correct response on item i, and $Q_i(\theta)$ is the probability of an incorrect response. As a prior distribution of θ a normal distribution is assumed. The normalization constant in (1) can be set equal to

$$f(\mathbf{u}_{k-1}) = \int L(\theta;\mathbf{u}_{k-1}) f(\theta) d\theta. \qquad (3)$$

Substitution of (2) and (3) in (1) gives an expression for the posterior distribution after (k-1) items have been administered. Some of the criteria that are introduced below are not based on the posterior, but on a point estimator that takes the posterior into account. The estimator of choice is the expected a posteriori (EAP) estimator. After (k-1) items, this estimator is found as

$$\hat{\theta}_{k-1} = E(\theta \mid \mathbf{u}_{k-1}) = \int \theta f(\theta \mid \mathbf{u}_{k-1}) d\theta. \qquad (4)$$

The same estimator is also used to report the final scores of the examinees. To evaluate the use of Bayesian item selection criteria for the problem of constrained adaptive testing, the following item selection criteria were applied.

Owen's criterion

A Bayesian procedure for CAT is described in Owen (1975). Item selection is based on the EAP estimator of the ability parameter. The difficulty of the *k*-th item is chosen to satisfy

$$|b_k - \hat{\theta}_{k-1}| < \delta, \qquad (5)$$

where *δ* is an approximately small constant to be determined. In the procedure, bounds were set for the discrimination index *a*, and guessing parameter *c*. This criterion does not select a unique item. It might even happen that no item is selected at all. An approach that solves these problems is to minimize *δ*, such that the expression in Equation (5) is met. Bloxom & Vale (1987) generalized Owen's procedure to the case of multidimensional adaptive testing.

Maximum Posterior-Weighted Kullback-Leibler Information

Kullback-Leibler information can also be used for item selection. This information measure is based on the distance between two likelihoods over the same parameter space. When Kullback-Leibler information is applied in the context of adaptive item selection, the purpose is to select items that maximize the distance between the true ability θ^*, and any other value of the ability parameter *θ*. For a single item, Kullback-Leibler information can be written as

$$K_i(\theta,\theta^*) = E[\frac{L(\theta^* \mid \mathbf{u})}{L(\theta \mid \mathbf{u})}] = P_i(\theta^*)\ln\frac{P_i(\theta^*)}{P_i(\theta)} + (1-P_i(\theta^*))\ln\frac{1-P_i(\theta^*)}{1-P_i(\theta)}. \qquad (6)$$

Chang & Ying (1996) showed that for a test on *n* items Kullback-Leibler information is equal to the sum of Kullback-Leibler information of the individual items. Because of this, the next item to be selected has to maximize Kullback-Leibler item information. Since the true ability θ^* is generally unknown, and the ability parameter *θ* is unspecified in (6), the item information function can not be calculated directly. Chang & Ying propose to use posterior-weighted item information at the current ability estimate. The criterion for selecting the *k*-th item is defined as:

$$\max_i \int K_i(\theta \mid \hat{\theta}_{k-1}) f(\theta \mid \mathbf{u}_{k-1}) d\theta, \qquad (7)$$

where the index *i* runs over the items that have not been administered yet. Veldkamp & van der Linden (2002) applied the posterior-weighted Kullback-Leibler information criterion to several cases of constrained multidimensional adaptive testing.

Maximum Posterior-Weighted Information

The posterior-weighted information criterion is introduced in van der Linden (1998a). This criterion is based on the maximum observed information measure. When logistic IRT models are used, the observed information measure is equal to Fisher's information measure (Veerkamp, 1996). A Bayesian criterion is formulated by taking the expectation of the information measure across the posterior distribution. The criterion for selecting the *k*-th item is defined as

$$\max_i \int I_i(\theta) f(\theta \mid \mathbf{u}_{k-1}) d\theta, \qquad (8)$$

where $I_i(\theta)$ denotes Fisher's information measure, that is defined as

$$I_i(\theta) = -E\left\{ \frac{d^2}{d\theta^2} \ln L(\theta; \mathbf{u}_{k-1}) \right\}, \qquad (9)$$

and *i* runs over the items in the test that have not been administered yet.

Minimum Posterior Variance Criterion

For Fisher's information measure it can be shown that the reciprocal of this measure is asymptotically equal to the posterior variance. In small-sample applications it might be preferable to optimize posterior variance instead of the posterior-weighted information in a test. When the EAP estimator is applied, the posterior variance or uncertainty about this estimator can be expressed by

$$\text{var}(\hat{\theta}_{k-1} \mid \mathbf{u}_{k-1}) = \frac{\int \theta^2 f(\theta \mid \mathbf{u}_{k-1}) d\theta}{\int f(\theta \mid \mathbf{u}_{k-1}) d\theta} - E(\theta \mid \mathbf{u}_{k-1})^2. \qquad (10)$$

When the posterior variance criterion is applied, selecting the k-th item comes down to

$$\min_{i} \operatorname{var}(\hat{\theta} \mid \mathbf{u}_{k-1}; u_i) \qquad (11)$$

where i runs over the items that have not been administered yet, and u_i is the expected answer on item i based on $\hat{\theta}_{k-1}$.

Posterior Expected Information Measures.

The previous item selection criteria are based on observed information and a posterior distribution of the ability of the examinee. An alternative approach would be to take the probability distribution of unanswered items into account. In van der Linden (1998a), the maximum expected information measure is introduced. After responding (k-1) items, the probability distribution of the answers to the next item is described by the following posterior predictive probability:

$$P_k(U_k = u_k \mid \mathbf{u}_{k-1}) = \int P_k(U_k = u_k \mid \theta) f(\theta \mid \mathbf{u}_{k-1}) d\theta. \qquad (12)$$

The maximum expected information criterion selects the item that maximizes the expected observed information in the test, after administering the next item. The item is answered correctly $(u_{i_k} = 1)$ or incorrectly $(u_{i_k} = 0)$, and the EAP estimates are applied to calculate the observed information. This item selection criterion can be denoted as

$$\begin{aligned}\max_{i}\{&P_i(U_i = 0 \mid \mathbf{u}_{k-1}) I_{u_k}(\theta \mid u_1,\ldots,u_{k-1},u_{i_k} = 0) + \\ &P_i(U_i = 1 \mid \mathbf{u}_{k-1}) I_{u_k}(\theta \mid u_1,\ldots,u_{k-1},u_{i_k} = 1)\},\end{aligned} \qquad (13)$$

where $I_{u_i}(\theta \mid u_1,\ldots,u_{k-1},u_{i_k} = 0)$ denotes Fisher information of a k-item test where the k-th item is answered incorrectly. Several other criteria that take future responses into account for selecting the next item are introduced in van der Linden (1998a), but they showed identical performance. All these criteria turned out to result in estimates with smaller MSE and bias, when they were compared with item selection based on maximum Fisher

Information or maximum posterior-weighted Fisher Information (van der Linden, 1998a).

BAYESIAN CRITERIA AND THE SHADOW TEST APPROACH

When Bayesian item selection criteria are applied in a practical context, different kinds of test specifications have to be met. Three kinds of specifications can be distinguished (van der Linden, 1998b, 2005). Some specifications deal with item content, item type, answer key, gender bias, or ethnical bias. These specifications are generally denoted as categorical criteria. Other criteria deal with quantitative specifications like word count or time limit. These criteria can be formulated as a function of the items involved. Psychometric aspects can also be described as test specifications. In fact, they are examples of quantitative test specifications. The aspects can be formulated as a function of the items involved. Sometimes, constraints that control for inter-item dependencies are imposed. Examples of this kind of criteria are enemy constraints, or constraints that deal with item sets.

In Veldkamp (1999), it is described how to deal with multiple criteria test assembly problems. One strategy is to define targets for each criteria and to minimize the weighted deviation of these targets. This strategy has been applied successfully by Stocking & Swanson (1993) when they developed their weighted deviation model (WDM). However, this method allows deviations from the targets, and as a consequence, it can't be guaranteed that all specifications are being met. In the shadow test approach (STA) one criterion is optimized and bounds are defined for all the others. In this way, the STA guarantees that all test specifications will be met, and test validity is increased. The STA is a general purpose algorithm for doing constrained adaptive testing. Other methods for handling test specifications in adaptive testing procedures have been proposed: Multi-stage testing (Lord, 1980, Adema, 1990, van der Linden, & Adema, 1998, and Luecht, & Nungester, 1998); testlet-based testing (Wainer & Kiely 1987, Wainer, Bradlow & Du, 2000, Glas, Wainer & Bradlow, 2000); or item pool partitioning (Kingsbury & Zara, 1991, Segall, Moreno, Bloxom, & Hetter, 1997). But these procedures are either are not fully adaptive, or are only able to handle categorical criteria.

The following pseudo-algorithm describes the STA:

1. Initialize the ability estimator.
2. Assemble a shadow test that meets the constraints and has maximum information at the current ability estimate.

3. Administer the item in the shadow test with maximum information at the ability estimate.
4. Update the ability estimate.
5. Return all unused items to the pool.
6. Adjust the constraints to allow for the attributes of the item administered.
7. Repeat Steps 2-6 until *m* items have been administered.

Where *m* denotes the test length. In Step 2 of the algorithm a shadow test is assembled. The assembly model optimizes the amount of information in the test, and bounds are set for the test specifications.

The STA requires the test assembly problem to be modeled as a linear programming model. Both the item selection criterion and the criteria that deal with different test attributes have be formulated as linear functions of the items. Standard commercial software packages for solving LP-models can be applied to assemble shadow tests in Step 2 of the algorithm, and in Step 3, the best item in the shadow test is selected to be presented to the examinee. For an overview of how to formulate specifications as linear functions of the items, see van der Linden (1998b, 2005).

Owen's Criterion

For some Bayesian item selection criteria, the requirement to be formulated as a linear function of the items can be easily fulfilled. Owen's criterion is very straightforward to apply. Instead of selecting the *k*-th item to meet the inequality in (5), all unadministered items in the shadow test have to meet this constraint. The LP formulation for applying Owen's criterion comes down to adding the next constraint to the test assembly model

$$|b_k x_i - \hat{\theta}_{k-1}| < \delta \quad \forall_i, \qquad (14)$$

where x_i is a decision variable that denotes whether item *i* is selected (x_i =1) or item *i* is not selected (x_i =0) for the test. A different implementation of Owen's criterion would be to minimize δ such that all items in the shadow test meet the inequality constraint defined in Equation (14).

Maximum Posterior-Weighted KL Information

Posterior-Weighted KL-information can also be handled by the STA The information measure can be formulated as a linear combination of the information of the items (Chang & Ying, 1996, Veldkamp & van der Linden, 2002):

$$\max \sum_{i=1}^{I} \int K_i(\theta, \hat{\theta}_{k-1}) f(\theta \mid \mathbf{u}_{k-1}) d\theta \cdot x_i. \qquad (15)$$

Maximum Posterior Weighted Information

For maximum posterior-weighted information, the following derivation shows how to formulate the criterion as a linear function of the items.

$$\begin{aligned}
&\max \int I(\theta) f(\theta \mid \mathbf{u}_{k-1}) d\theta \quad \Leftrightarrow \\
&\max \int (\sum_{i \in test} I_i(\theta)) f(\theta \mid \mathbf{u}_{k-1}) d\theta \quad \Leftrightarrow \\
&\max \int (\sum_{i=1}^{I} I_i(\theta) x_i) f(\theta \mid \mathbf{u}_{k-1}) d\theta \quad \Leftrightarrow \\
&\max \sum_{i=1}^{I} \int I_i(\theta) f(\theta \mid \mathbf{u}_{k-1}) d\theta \cdot x_i.
\end{aligned} \qquad (16)$$

From this derivation it can be concluded that, although calculating the integral in the last equation for all items might take some computation time, the maximum posterior-weighted information criterion can also be implemented in the STA for constrained adaptive testing.

Posterior Expected Information Measures and Posterior Variance

Application of other Bayesian item selection criteria that either incorporate the posterior variance or expectations of future answers is more complicated. When posterior expected information measures were described (13), only one future answer was taken into account. When an optimal shadow test is assembled in Step 2 of the STA, contributions of all remaining items have to be added. As a consequence, all possible answer patterns to the remaining items and their probabilities have to incorporated in the item selection process. When the maximum expected information criterion is applied, the objective function of the LP model for selecting all remaining (n-k) items can be formulated as

$$\max \int \sum_{U_i \in \{0,1\}} \sum_{i=1}^{I} I_i(\hat{\theta}_n) x_i \cdot P_i(U_{i_{k+1}}, ..., U_{i_n} \mid \mathbf{u}_k) d\theta. \qquad (17)$$

This function could be rewritten into a linear function of the items. But after administering k items, the number of possible answer patterns for the remaining items is 2^{n-k}. Especially during the first few stages of adaptive testing, the resulting objective function becomes very complex, hard to manage, and time consuming to calculate. Since the time for selecting the next item is restricted, assembly of a shadow test in this way does not seem to be a suitable approach.

An additional problem occurs when the minimum posterior variance criterion is applied. Because of the square in the equation, this criterion results in an information measure that is a non-linear function of the items. Therefore, Linear-Programming techniques can not be applied to assemble shadow tests that perform optimally with respect to the information measure. Since assembly of an optimal shadow test is essential in the STA, application of the approach becomes problematic. A rather straightforward way to overcome this problem would be to linearize the objective function. However, linear approximations of the objective function only provide good results, when the first order Taylor approximation is close to the function, which is not the case for this criterion. In Veldkamp (2002), it is illustrated what happens when the approximation oversimplifies the problem.

A strategy to prevent these problems is to adapt the STA in order to handle non-linear and complicated objective functions.

STA FOR HANDLING COMPLICATED NON-LINEAR OBJECTIVES

A modification of the STA would be to select a single item in every iteration of the CAT algorithm that performs optimal with respect to the information measure, and to make sure that for this item a feasible shadow test exists that meets the constraints. Instead of assembling an optimal shadow test, the modified approach only selects the optimal next item. An heuristical algorithm for this approach is to select the item that provides most information, and to make an attempt to assemble a shadow test. When the attempt succeeds, the item will be administered. When the attempt fails, the item that performs next best is selected, and the procedure is repeated until an item is found for which a feasible shadow test exists. Items that have been rejected, have to be temporarily removed from the item pool, just to make sure that they won't be in the list again for selection subsequent items. After the candidate completed the test, the rejected items can be returned to the pool.

A pseudo algorithm for this Modified version of the Shadow Test Approach (MSTA) is

```
1. Initialize the ability estimator.
2. Order the unadministered items with respect to their contribution to
   the Bayesian criterion at the current ability estimate.
3. Iteratively
       a. Select the subsequent item in the list.
       b. Try to assemble a shadow test that contains this item.
       c. When a feasible shadow test is found, proceed to Step 4.
       d. Else remove the item temporarily from the item pool and
          return to Step 3(a).
4. Administer the selected item.
5. Update the ability estimate.
6. Adjust the constraints.
7. Repeat Steps 2 - 6 until $n$ items have been administered.
```

The purpose of constrained CAT is to assemble an adaptive test that meets all the constraints and performs optimal with respect to the information measure at the true ability level of the examinee. The check in Step 3(b) guarantees that items are selected that meet the constraints. With respect to optimal measurement precision at the true ability level, the MSTA differs slightly from the STA. For the STA, optimal assembly of a full length shadow test in every iteration can be proven to converge to the optimal value for the information measure (van der Linden, 2000). The MSTA heuristic only performs optimal for selecting the next item at the given ability estimate.

Technical implementation

When the pseudo algorithm for the MSTA is implemented, the steps that need special consideration are Step 2 and Step 3(b). In Step 2 of the pseudo-algorithm, the items are ordered with respect to their contribution to the information measure. When items are selected based on a Bayesian information measure, calculation is required of posterior distributions or of EAP estimates. Evaluation is needed of integrals for which no expression in closed form is available. In this paper, the prior distribution of the ability parameters is chosen to follow a normal distribution, so the method of Gauss-Hermite Quadrature could be applied.

In step 3(b) of the algorithm, an attempt has to be made to assemble a shadow test. Since this shadow test has only to be feasible and need not to perform optimal with respect to the Bayesian criterion, the objective function can be chosen to be an arbitrary linear function of the items. The constraints in the model for assembly of the modified shadow test are almost similar to those of the shadow test. The only difference is that a constraint is added to make sure that the item proposed for selection is in the shadow test; that is, if item i^* is evaluated, the constraint (x_{i*}=1) is added. As a result, both objective function and the constraints are linear functions and 0-1 LP-techniques can be applied. The LP model can be imported in optimization software packages like CPLEX (ILOG, 2003). These packages employ efficient implementations of implicit enumeration algorithms, like the Branch-and-Bound (BAB) algorithm, to find feasible solutions to the test assembly problems.

LP-model for modified shadow test

To formulate the LP model, the following notation will be used:

$i = 1,\ldots,I$	items in the pool,
$k = 1,\ldots,n$	items in the adaptive test,
S_{k-1}	set of items previously administered,
S_c	set of items with categorical attribute *c,*
S_q	set of items with quantitative attribute *q,*
S_e	set *e* of mutually exclusive items.
x_i	decision variable to denote whether the item is in the test ($x_i = 1$).

Observe that the test assembly model in Step 3(b) is almost equal to the shadow test assembly model (van der Linden & Reese, 1998). The same set of constraints has to be met. The differences are that a constraint has to be added to make sure that the selected item i^* is in the shadow test, and that the objective function can be chosen arbitrarily. Therefore, when the item pool does not contain item sets, the model comes down to:

$$\max f(\cdot) \qquad (18)$$

Subject to

$$\begin{aligned}
&\sum_{i \in S_c} x_i \leq n_c \\
&\sum_{i \in S_q} q_i x_i \leq n_q \\
&\sum_{i \in S_e} x_i \leq 1 \\
&\sum_{i \in S_{k-1}} x_i = k-1 \\
&x_{i^*} = 1 \\
&x_i \in \{0,1\} \qquad (19)
\end{aligned}$$

where *f*(.) is an arbitrary function. Equations (19a)-(19c) describe generic constraints for specifying categorical attributes, quantitative attributes and inter-item dependencies of the test. In (19d), values of the decision variables for the administered items are fixed to one. Constraint (19e) guarantees that the selected item i^* is in the shadow test.

Computational Complexity

From theory of linear programming it is known that the solution time for 0-1 LP problems is not bounded by a polynomial function of the problem size (Nemhauser & Wolsey, 1988). On the other hand, it should be mentioned that the time needed for solving the problem highly depends on how the model is built (Williams,1999). When the general LP formulations of the models for the STA and the modified STA are compared, it turns out that the model for the modified STA is less complicated. The objective is not to find an optimal solution, but to find a feasible solution. Because of this, computation time will be reduced.

On the other hand, the STA requires one LP model to be solved, whereas the number of LP models to be solved for the modified STA is unknown in advance. It depends on the quality of the item bank whether a feasible shadow test exists for the subsequent items in the row defined in Step 2 of the algorithm.

A rather pragmatic approach to reduce the number of LP models to be solved when the modified STA heuristic is applied, is to limit the number of items that are checked. When no feasible test is found within a time limit, for example a time limit based on observed response times, just take the next best item from a previous shadow test. The consequence is that the selected item might be different from the optimal item.

NUMERICAL EXAMPLES

CAT software developed at the University of Twente was applied. The software makes calls to the solver in the CPLEX 9.0 package (ILOG, 2003) to assemble the shadow tests. In the first example, the performance of the STA and the modified STA were compared. In the second example different criteria were applied to an adaptive version a high-stakes admission test.

Comparison STA and MSTA

An item pool from the ACT Assessment Program was used to simulate adaptive test administrations following the STA and the modified STA. The item pool consisted of 176 items calibrated under a two-dimensional version of 2-Parametric Logistic Model (Reckase, 1985).

$$P_i(\theta_j) = \frac{(\exp^{(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i})}{(1 + \exp^{(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i})}, \tag{20}$$

where $P_i(\theta_j)$ is the probability that a person j gives correct answer to item i, a_i is the vector of slope parameters of item i along the dimensions of the ability vector θ_j, and d_i is a scalar denoting the easiness of the item. The calibration of the item parameters was carried out using the program NOHARM (Fraser & McDonald, 1988). The items in the pool were classified according to the five content and three skill categories used in the ACT Assessment Program to formulate the test specifications: Pre-Algebra (PA); Elementary Algebra (EA); Coordinate Geometry (CG); Trigonometry (T); Intermediate Algebra (IA); Basic Skills (BS); Application (AP); and Analysis (AN). The test consisted of 25 items.

Maximum Fisher information criterion for multiple dimensions (Segall, 1996) was applied to select the items (See also Veldkamp, 2002). This criterion can be applied for both the STA and the modified STA. Both approaches were compared with respect to MSE of the resulting ability estimators.

A grid of nine points, {-1,0,1 }×{ -1,0,1 }, was applied to discretize the two-dimensional space. For each point, 1000 examinees were simulated. The resulting MSE's are shown in Figure 1.

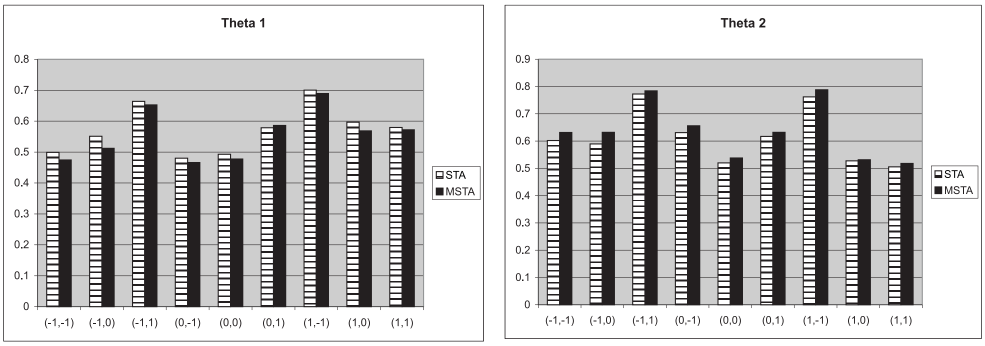

Figure 1: MSE’s for both ability estimates.

The modified approach performed slightly better for the first dimension, whereas the STA performed better for the second dimension. An explanation can be found in the value of the discrimination parameters for the different dimensions. The discrimination parameters of the items for the first dimension were on average higher than the discrimination parameters for the second dimension. The MSTA strictly focuses on selecting the most informative item in each iteration. Therefore, it focuses more on the first dimension and better MSEs are obtained there. Overall, both approaches seem to perform equally well.

The average time for selecting the next item was smaller for the modified approach then for the STA. However, the maximum time needed for the modified approach was larger. The maximum time was obtained for selecting the last item. At the end of the test, many items were rejected before a feasible test was found.

Application of Bayesian Criteria to the LSAT

To compare the different Bayesian criteria, they were applied to a condensed adaptive version of the LSAT. The item pool for the admission test contained 753 items, that fitted the 3-Parameter Logistic Model (3PLM). So the probability that examinee *j* answers item *i* correctly can be described as

$$P_i(\theta_j) = c_i + (1 - c_i)\frac{\exp^{(a_i\theta_j - b_i)}}{1 + \exp^{(a_i\theta_j - b_i)}}, \qquad (21)$$

where a_i denotes the discrimination parameter, b_i the difficulty parameter, and c_i the guessing parameter of item *i*. In this simulation study, item set constraints were ignored. Restrictions about content, item type, minority orientation, gender orientation, word count and answer key were present. The test contained 50 items. The total number of constraints was equal to 94. The initial estimate of the ability parameter was set equal to $\hat{\theta} = 0$. Both MSE and Bias were recorded after 10, 20, 30, and 50 items.

The method with probabilistic item-ineligibility constraints (van der Linden & Veldkamp, 2004, 2007) was used to control for over- exposure of the items. The target value for the exposure rates was set at .25.

The Bayesian criteria were compared with respect to (1) MSE and Bias of the resulting ability estimates, and (2) The number of items used. Selecting items based on Fisher information was applied to judge the performance of the selection rules.

In Figure 2, the bias functions are shown. All selection rules show inward bias. The bias decreases when test length increases. No differences in performance were found between selection of items based on Fisher information or on one of the Bayesian selection rules, besides Owen's criterion. In this simulation, all criteria performed equally well with respect to bias. For small tests (test length = 10), it was possible to distinguish between the different item selection rules, but for test lengths larger than 20 items, hardly any differences were found. It turned out that the different selection criteria selected mainly the same group of items. Just, the order in which the items were selected differed.

In Figure 3, the MSE functions are shown. Owen's criterion is performing worse than the other criteria. For small tests, test length equal to ten, small differences in performance can be observed. No persistent differences between the criteria were found, and maximum Fisher Information criterion performed comparable to the Bayesian criteria.

The second criteria dealt with item pool usage. The exposure rates of the items are shown in Figure 4. The maximum exposure rates slightly exceed the target of .25. This is because of the probabilistic nature of exposure control methods. The only criterion that showed different behavior is Owen's criterion. This criterion increased the number of items used. All other criteria performed comparably with respect to item pool usage.

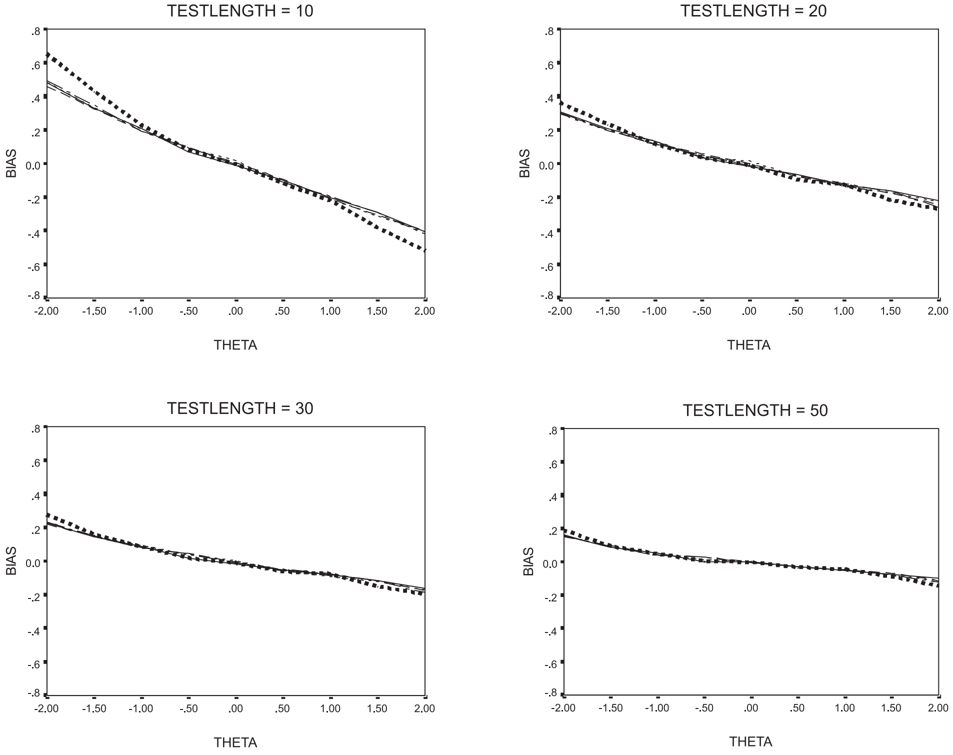

Figure 2: BIAS for different test lengths. Fisher information (solid line), Owen's criterion (big dotted line), maximum posterior-expected KL information (dashed line), maximum posterior weighted information (small dotted line), and maximum expected information (dashed-dotted line).

The reason why Owen's criterion is behaving differently is the way it is implemented. As mentioned before, Owen's criterion in Equation (5) does not select an optimal item. It just selects an item that fulfils the inequality. Therefore, Bias and MSE are larger, and item pool usage is increased.

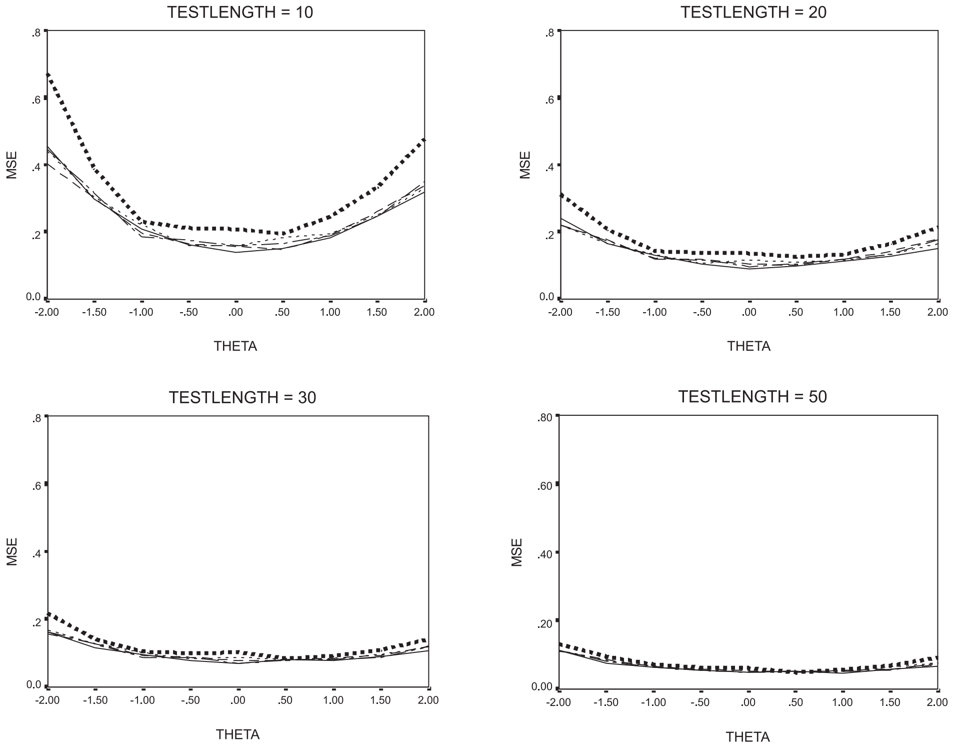

Figure 3: MSE for different test lengths. Fisher information (solid line), Owen's criterion (big dotted line), maximum posterior-expected KL information (dashed line), maximum posterior weighted information (small dotted line), and maximum expected information (dashed-dotted line).

DISCUSSION

Bayesian item selection criteria turned out to be formulated by rather complex functions. Owen's criterion, maximum posterior-weighted Kullback-Leibler information, and maximum posterior-weighted information could be transformed into linear equations rather straightforwardly. Other criteria, like minimum posterior variance and posterior-expected information measures, either resulted in very complicated linear or in non-linear equations. In general, Bayesian criteria that take expectations over a posterior result in linear equations, whereas criteria that take posterior predictive probabilities of future responses into account result in equations that are difficult to handle.

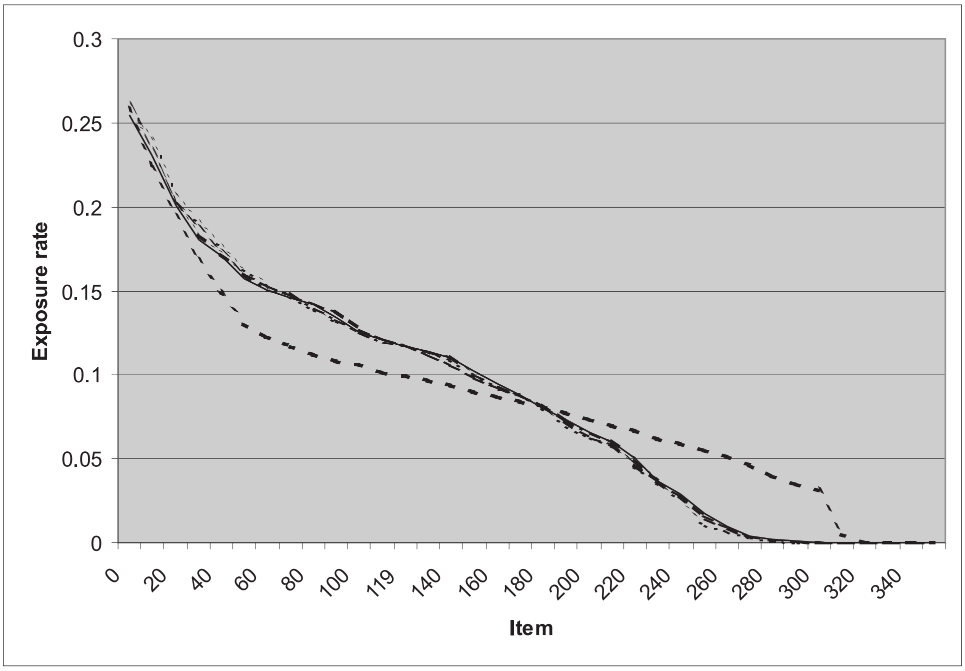

Figure 4: Exposure rates of items. Fisher information (solid line), Owen's criterion (big dotted line), maximum posterior-expected KL information (dashed line), maximum posterior weighted information (small dotted line), and maximum expected information (dashed-dotted line).

In order to deal with all possible criteria the modified shadow test was introduced. In this method, the item selection steps of the STA are modified. The modified STA is able to handle both linear and non-linear objective functions, because it only focuses on optimal selection of the next item.

About the performance of the modified STA heuristic some remarks can be made. The search procedure in the pseudo algorithm of the MSTA is rather naive. For example, it does not take into account any information gathered about why selection of a certain item does not result in a feasible shadow test. Infeasibility analysis (Huitzing, Veldkamp, & Verschoor, 2005, ILOG, 2003) might reveal which features of the item cause infeasibility. Items in the list that have the same features can be skipped in the search process. This approach would speed up the algorithm considerably.

In the second study, no differences in performance between selecting items based on maximum Fisher information and Bayesian item selection criterion were observed. This result was quite surprising, because in van der Linden (1998a), Bayesian criteria outperformed maximum Fisher information with respect to MSE even for 30 item tests. An explanation for this difference is believed to lie in the regression of the prior on background variables used in van der Linden (1998a). Apparently, this regression accounts for the differences found.

REFERENCES

Adema, J.J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement, 27*, 241-253.

Bloxom, B.M., & Vale, C.D. (1987, june). *Multidimensional adaptive testing: A procedure for sequential estimation of the posterior centroid and dispersion of theta*. Paper presented at the annual meeting of the Psychometric Society, Montreal.

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

Fraser, C., & McDonald, R.P. (1988). *NOHARM II: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: University of New England, Centre for Behavioral Studies.

Glas, C.A.W., Wainer, H., & Bradlow, E.T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W.J. van der Linden, and C.A.W. Glas (Eds.) *Computerized adaptive testing: Theory and practice*, (p. 271-288). Boston, MA: Kluwer Academic Publishers.

Huitzing, H.A., Veldkamp, B.P., & Verschoor, A.J. (2005). Infeasibility in automated test assembly models: A comparison study of different methods. *Journal of Educational Measurement, 42,* 223-243.

ILOG inc. (2003). *CPLEX 9.0* [Computer program and manual]. Incline Village, NV: ILOG.

Kingsbury, G.G., & Zara, A.R. (1991). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 4*, 359-375.

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

Luecht, R.M., & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 229-249.

Nemhauser, G.L., & Wolsey, L.A. (1988). *Integer and combinatorial optimization*. New York, NY: John Wiley & Sons Inc.

Owen, R.J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of American Statistical Association, 70*, 351-356.

Reckase, M.D. (1985). The difficulty of test items that measure more than one dimension. *Applied Psychological Measurement, 9,* 401-412.

Segall, D.O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354.

Segall, D.O., Moreno, K.E., Bloxom, B.M. & Hetter, R.D. (1997). Psychometric procedures for administering CAT-ASVAB. In W.A. Sands, B.K. Waters, & J.R.

McBride (Eds.) *Computerized adaptive testing: From inquiry to operation.* (p. 131-140). Washington D.C.: American Psychological Association.

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

van der Linden, W.J. (1998a). Bayesian item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-226.

van der Linden, W.J. (1998b). Optimal assembly of educational and psychological tests. *Applied Psychological Measurement, 22*, 195-211.

van der Linden, W.J. (2000). Constrained adaptive testing with shadow tests. In W.J. van der Linden, and C.A.W. Glas (Eds.) *Computerized adaptive testing: Theory and practice,* (p. 27-53). Boston, MA: Kluwer Academic Publishers.

van der Linden, W.J. (2005). *Linear models for optimal test design.* New York: Springer Verlag.

van der Linden, W.J., & Adema, J.J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement, 35*, 185-198.

van der Linden, W.J., & Reese, L.M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259-270.

van der Linden, W.J. & Veldkamp, B.P. (2004). Constraining item exposure rates in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29,* 273-291.

van der Linden, W.J. & Veldkamp, B.P. (2007). Conditional item exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32,* 398-417.

Veerkamp, W.J.J. (1996). *Statistical inference for adaptive testing*. (Internal Report). Enschede, The Netherlands: University of Twente, Department of educational measurement and data-analysis.

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.

Veldkamp, B.P (1999). Multiple objective test assembly problems. *Journal of Educational Measurement, 36*, 253-266.

Veldkamp, B.P. (2002). Multidimensional constrained test assembly. *Applied Psychological Measurement, 26*, 133-146.

Veldkamp, B.P., & van der Linden, W.J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*, 575-588.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H., Bradlow, E.T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W.J. van der Linden, and C.A.W. Glas (Eds.) *Computerized adaptive testing: Theory and practice*, (p. 245-270). Boston, MA: Kluwer Academic Publishers.

Williams, H.P. (1999). *Model building in mathematical programming*. (4th ed.). Chichester, England: John Wiley & Sons Ltd.

(Manuscript received: 5 November 2007; accepted: 24 February 2009)