DOCUMENT RESUME

ED 424 235 TM 027 361

AUTHOR van der Linden, Wim J.
TITLE Bayesian Item Selection Criteria for Adaptive Testing. Research Report 96-01.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE 1996-10-00
NOTE 32p.
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; *Bayesian Statistics; Computation; Computer Assisted Testing; *Criteria; Equations (Mathematics); Error of Measurement; Estimation (Mathematics); *Selection; *Test Items

ABSTRACT

R. J. Owen (1975) proposed an approximate empirical Bayes procedure for item selection in adaptive testing. The procedure replaces the true posterior by a normal approximation with closed-form expressions for its first two moments. This approximation was necessary to minimize the computational complexity involved in a fully Bayesian approach, but is no longer necessary given the computational power currently available in adaptive testing. This paper suggests several item selection criteria for adaptive testing that are all based on the use of the true posterior. Some of the statistical properties of the ability estimator produced by these criteria are discussed and empirically characterized. An empirical study with 300 test items showed that the maximum predicted posterior expected information criterion had excellent mean-squared error for more extreme values of theta, and is the criterion elect for application in short adaptive tests. An appendix presents Owen's equations. (Contains 17 references.) (Author/SLD)

TM

Bayesian Item Selection Criteria for Adaptive Testing

Research Report 96-01

Wim J. van der Linden

Bayesian Item Selection Criteria for Adaptive Testing

Wim J. van der Linden

Bayesian Item Selection Criteria for Adaptive Testing, Wim J. van der Linden - Enschede: University of Twente, Faculty of Educational Science and Technology, December 1996. - 29 pages.

Abstract

Owen (1975) proposed an approximate empirical Bayes procedure for item selection in adaptive testing. The procedure replaces the true posterior by a normal approximation with closed-form expressions for its first two moments. This approximation was necessary to minimize the computational complexity involved in a fully Bayesian approach but is no longer necessary given the computational power currently available in adaptive testing. This paper suggests several item selection criteria for adaptive testing which are all based on the use of the true posterior. Some of the statistical properties of the ability estimator produced by these criteria are discussed and empirically characterized.

Introduction

Adaptive testing is based on the principle of selecting items to match the current estimate of the ability of the examinee. An important choice is how to translate this principle into a formal criterion of item selection implementable as a computer algorithm. Since the early days of adaptive testing, two item selection criteria have been popular: the maximum-information criterion and an approximate Bayesian criterion proposed by Owen (1975).

It is the purpose of this paper to introduce several new criteria for item selection in adaptive testing which are all Bayesian in the sense that they are based on the posterior distribution of the ability of the examinee. The criteria can be used as an alternative to Owen's criterion which is based on an approximate empirical Bayes approach to adaptive testing. The approximation was introduced at a time when the numerical complexity involved in fully Bayesian approach was a practical problem. However, for the computers currently in use in adaptive testing programs, this complexity is no longer a problem.

The paper is organized as follows: First, the maximum-information and Owen's criterion are reviewed. Subsequently, several Bayesian criteria for item selection are introduced. Next, some statistical properties of the final ability estimators for an adaptive test based on these criteria are discussed. The last section of the paper presents results from a simulation study run to characterize the properties of these estimators empirically.

Model

The two-parameter logistic model will be used as the response model under which the items in the pool have satisfactory fit. However, the results obtained in

this paper easily generalize to any (unidimensional) item response theory (IRT) model. To introduce the model, a random response variable U_i is defined to denote a correct (U_i=1) or an incorrect (U_i=0) response to item i. The model is given by the following equation for the probability of success on item i for an examinee with (fixed) ability $\theta \in (-\infty,\infty)$:

$$p_i(\theta) \equiv \text{Prob}\{U_i=1|\theta\} \equiv \frac{\exp[a_i(\theta-b_i)]}{1+\exp[a_i(\theta-b_i)]}. \quad (1)$$

Location parameter $b_i \in (-\infty,\infty)$ and scale parameter $a_i \in [0,\infty)$ in this model are commonly interpreted as the difficulty and discriminating power of item i, respectively.

Maximum-Information Criterion

To present the maximum information criterion, the following notation is needed. The items in the pool are denoted by i=1,...,I, For convenience, an adaptive test of fixed length n will be assumed. The rank of the items in the test is denoted by index k=1,...,n. It follows that i_k is the index of the item in the pool administered as the kth item in the test. Suppose k-1 items have already been selected. The indices of these items form the set $S_{k-1} \equiv \{i_1,...,i_{k-1}\}$. The remaining set of items in the pool is denoted as $R_k \equiv \{1,...,I\} \backslash S_{k-1}$.

For responses $U_{i_1}=u_{i_1},...,U_{i_{k-1}}=u_{i_{k-1}}$ obtained on the first k-1 items, the likelihood function is

$$L(\theta|u_{i_1},...,u_{i_{k-1}}) \equiv \prod_{j=1}^{k-1} \frac{\{\exp[a_{i_j}(\theta-b_{i_j})]\}^{u_{i_j}}}{1+\exp[a_{i_j}(\theta-b_{i_j})]}. \quad (2)$$

An ML estimator (MLE) of θ based on these responses is a maximizer of (2) over

θ, that is,

$$\hat{\theta}^{ML}_{u_{i_1},\ldots,u_{i_{k-1}}} \equiv \max_\theta \{L(\theta \mid u_{i_1},\ldots,u_{i_{k-1}});\ \theta \in (-\infty,\infty)\}. \qquad (3)$$

Fisher's information about the unknown value of θ in the response variables associated with the k-1 items is defined as:

$$I_{U_{i_1},\ldots,U_{i_{k-1}}}(\theta) \equiv -E\left(\frac{\partial}{\partial\theta^2} \ln L(\theta \mid U_{i_1},\ldots,U_{i_{k-1}})\right) = \sum_{j=1}^{k-1} \frac{(p_{i_j}'(\theta))^2}{p_{i_j}(\theta)(1-p_{i_j}(\theta))}, \qquad (4)$$

where

$$p_{i_j}'(\theta) \equiv \frac{\partial}{\partial\theta} p_{i_j}(\theta)$$

and the last step in (4) is a well-known result for the model in (1) (see, for example, Hambleton & Swaminathan, 1985, sect. 6.3). Note that (4) is additive because of conditional independence of the response variables given θ.

The maximum-information criterion common in adaptive testing selects the kth item such that maximum information is obtained at $\theta = \hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}}$, i.e.,

$$i_k = \max_j \{I_{U_{i_1},\ldots,U_{i_{k-1}},U_j}(\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}});\ j \in R_k\}. \qquad (5)$$

Because the information measure is additive, the criterion is equivalent to

$$i_k = \max_j \{I_{U_j}(\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}});\ j \in R_k\}. \qquad (6)$$

Owen's Criterion

As an alternative to the maximum-information criterion, Owen (1975) proposed an approximate empirical Bayes procedure for adaptive testing based on the following three-parameter normal-ogive model:

$$p_i(\theta) \equiv c_i+(1-c_i)\Phi[a_i(\theta-b_i)], \tag{7}$$

where $\Phi(.)$ is the normal distribution function and c_i a lower asymptote to model the probability of guessing item i correctly.

For a vector of responses to the first k-1 items, the likelihood function is given by (2) with the logistic factor replaced by the normal ogive. Assuming a prior $g(\theta)$, the following expression for the posterior distribution of θ after k-1 items is obtained:

$$g(\theta|u_{i_1},...,u_{i_{k-1}}) = \frac{L(\theta|u_{i_1},...,u_{i_{k-1}}g(\theta)}{\int L(\theta|u_{i_1},...,u_{i_{k-1}})g(\theta)d\theta}. \tag{8}$$

Owen's procedure is based on (8) as an updating procedure for the posterior with the choice of a normal density for the prior $g(\theta)$. Item k is chosen to satisfy

$$|b_{i_k}-E(\theta|u_{i_1},...,u_{i_{k-1}})| < \delta, \tag{9}$$

for a small value of δ, where $E(\theta|u_{i_1},...,u_{i_{k-1}})$ is the expectation of θ over (8) now generally known as the Expected A Posteriori (EAP) estimator of θ. The procedure is stopped as soon as the variance of (8) is smaller than a prespecified threshold value.

It should be noted that the likelihood in (8) does not have the normal family as

class of conjugate distributions. Therefore, if a normal prior is chosen, the posterior is not normal, and repeated updating of the posterior using (8) soon leads to a posterior that could not be calculated in applications to real-time adaptive testing by the computers available in the 70s. Owen therefore proposed to replace the true posterior by a normal approximation with the same expected value and variance and presented closed-form expressions for the update of a normalized posterior (see Appendix 1). The approximation was motivated by showing that for mild bounds on δ the estimator $\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}} \equiv E(\theta | u_{i_1},\ldots,u_{i_{k-1}})$ goes to the true value of θ in mean square for $k \rightarrow \infty$ (Owen, 1975, Theorem 2).

Owen also referred to the criterion of minimization of the preposterior risk under a quadratic loss function as an alternative to (9). This criterion selects the item that minimizes the expected posterior variance. Computationally it is more involved than the criterion in (9) in combination with a normal approximation to the posterior, and for this reason the latter became widely popular as Owen's procedure of adaptive testing. An extensive simulation study of the statistical properties of Owen's procedure is reported in Weiss and McBride (1984). A generalization of the procedure to the case of multidimensional adaptive testing is discussed in Bloxom and Vale (1987). The criterion of minimum expected posterior variance will be returned to later in this paper.

Bayesian Criteria

A Bayesian approach to adaptive testing is loosely defined as any approach which uses a prior or posterior distribution to define rules for: (1) selecting the first item; (2) estimating θ; (3) selecting the next item; or (4) stopping the test. According to this convention, the use of an informative prior to select the first item

in an adaptive testing procedure is thus an example of a Bayesian adaptive testing procedure (van der Linden, 1996). Owen's procedure is adaptive in that the item selection criterion in (9) is based on the mean of the (approximate) posterior and the posterior variance is used to stop the test. However, his procedure does not base item selection on the full posterior which, in a Bayesian framework, is the best reflection of the uncertainty in the current ability estimate.

This section introduces several alternative criteria for item selection based on the full posterior. The first two criteria generalize the idea of maximum information in a Bayesian fashion. The next criterion is the one of minimum expected posterior variance also discussed in Owen (1975). The fourth criterion combines the ideas of posterior weighing and preposterior prediction underlying the first two criteria into a new one. Finally, some other Bayesian procedures of item selection are alluded to.

Maximum Posterior Expected Information

The first criterion reformulates the maximum information criterion in a Bayesian fashion by first choosing the appropriate information measure and then taking its expectation across the posterior distribution.

If the kth item is selected, responses to the first k-1 items are already known. Hence, these data can no longer be presented by random variables but only by the (fixed) values of their realizations. As a consequence, Fisher's information, defined as an expected value across random data, is no longer a valid measure. A typical Bayesian choice is to use the observed information measure

$$J_{u_{i_1},\ldots,u_{i_{k-1}}}(\theta) \equiv -\frac{\partial}{\partial\theta^2}\ln L(\theta \mid u_{i_1},\ldots,u_{i_{k-1}}) \qquad (10)$$

which reflects the relative curvature of the observed likelihood function at the value

of θ.

Though the distinction between the two information measures is important, it is easy to show that, under the model in (1), the second derivative in the right-hand side of (7) is the same for each possible response vector (for a derivation, see Veerkamp, 1996). Therefore, it holds for this model that

$$J_{u_{i_1},\ldots,u_{i_{k-1}}}(\theta) = I_{U_{i_1},\ldots,U_{i_{k-1}}}(\theta). \tag{11}$$

However, to obtain generality, the distinction between the two information measures will maintained.

The proposal is to select the next item to minimize the expected value of the observed information $J(\theta)$ over the posterior distribution of θ. The index of the kth item according to this criterion is:

$$i_k = \max_j \{\int J_{U_j}(\theta) g(\theta \mid u_{i_1},\ldots,u_{i_{k-1}}; j \in R_k\}, \tag{12}$$

where $g(\theta \mid u_{i_1},\ldots,u_{i_{k-1}})$ is the posterior update obtained from (8).

The criterion is a generalization of the likelihood weighted information criterion introduced in Luecht (1995) and Veerkamp and Berger (in press). The advantage of using the posterior for weighing the information is the possibility to incorporate prior knowledge about θ in the item selection procedure. Use of this possibility is recommended when data on background variables with a statistical relation to θ are available (van der Linden, 1996).

Maximum Predicted Expected Information

The following criterion predicts the probability distribution of the responses of the examinee on each item $j \in R_k$ and selects the item with maximum expected information over this probability distribution. More in particular, for each item $j \in R_k$ the distribution of U_j is given by the probabilities $\{p_j(\theta), 1-p_j(\theta)\}$. The best prediction of these probabilities is their evaluation at the current estimate $\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}}$. To maintain the Bayesian framework, $\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}}$ is chosen to be the maximum a posteriori (MAP) or expected a posterior (EAP) estimator throughout this paper. If at the next stage item j would be chosen and response $U_j=0$ obtained, the new estimate of θ would be $\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}},U_j=0}$ and observed information would be equal to $J_{u_{i_1},\ldots,u_{i_{k-1}},U_j=0}(\theta)$. Analogously, if $U_j=1$, the new estimate would become $\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}},U_j=1}$ and observed information would be equal to $J_{u_{i_1},\ldots,u_{i_{k-1}},U_j=1}(\theta)$.

The maximum predicted expected information criterion selects as the kth item:

$$i_k = \max_j \{(1-p_j(\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}})J_{u_{i_1},\ldots,u_{i_{k-1}},U_j=0}(\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}},U_j=0}) + p_{i_j}(\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}})J_{u_{i_1},\ldots,u_{i_{k-1}},U_j=1}(\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}},U_j=1}); j \in R_k\}. \quad (13)$$

Note that the criterion in (13) not only evaluates the observed information measure associated with $U_j=0$ and $U_j=1$ but that re-evaluation of the measure for $u_{i_1},\ldots,u_{i_{k-1}}$ is also implied. Further, since its two terms are evaluated at different values of θ, the expression in (13), though an expected value, is not an instance of Fisher's information defined in (4).

Minimum Predicted Expected Variance

If the information measures in (13) are replaced by the predicted posterior variances of θ for $U_j=0$ and $U_j=1$, the following criterion is obtained:

$$i_k = \max_j \{(1-p_j(\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}})\mathrm{Var}(\theta|u_{i_1},\ldots,u_{i_{k-1}},U_j=0) + p_j(\hat{\theta}^{k-1})\mathrm{Var}(\theta|u_{i_1},\ldots,u_{i_{k-1}},U_j=1); j\in R_k\}. \quad (14)$$

Though the use of information measures for item selection is a well-established practice in IRT, the reciprocal of the information measure is only a large-sample approximation to the true variance of the posterior. Therefore, from a Bayesian point of view, the criterion in (14) should be preferred over the one in (13). As already noted, the same criterion was proposed as an alternative to (9) in Owen (1975). Reviews of the criterion can be found, for example, in Thissen and Mislevy (1990) and Weiss (1982).

Maximum Predicted Posterior Expected Information

In the criterion in (12), observed information is predicted for wrong and correct responses and the expectation is taken over these predictions. However, rather than evaluating observed information at predicted point estimates, its expectation over predicted posteriors could also be used. Let $g(\theta|u_{i_1},\ldots,u_{i_{k-1}},U_j=0)$ and $g(\theta|u_{i_1},\ldots,u_{i_{k-1}},U_j=1)$ be the posterior of θ after a wrong and correct response to item j, respectively. The following proposal is to select as the kth item:

$$i_k = \max_j \{(1-p_j(\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}})\int J_{u_{i_1},\ldots,u_{i_{k-1}},U_j=0}(\theta)g(\theta|u_{i_1},\ldots,u_{i_{k-1}},U_j=0)d\theta$$

$$+ p_j(\hat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}})J_{u_{i_1},\ldots,u_{i_{k-1}},U_j=1}(\theta)g(\theta|u_{i_1},\ldots,u_{i_{k-1}},U_j=1)d\theta; j\in R_k\}. \quad (15)$$

Note that this criterion combines the ideas underlying the criteria in (12)-(13): As in (12), observed information is weighted by a posterior density, but at the same time the criterion shares the idea of preposterior prediction with (13).

Additional Criteria

The above criteria do not constitute an exhaustive set of posterior-based criteria for item selection. For example, it is an easy step to generalize the criteria in (13)-(15) to predictions two or more items ahead. However, for larger item pools the combinatorial complexity of such criteria would quickly exceed the possibility of application to real-time adaptive testing but for small pools the idea seems attractive. Chang (1996) proposes to replace Fisher's information by the Kullback-Leibler measure. The same substitution could easily be made for the criteria (12)-(13) and (15). Analogous to (15), the predicted probabilities in (14) could be replaced by expectations over predicted posterior distributions. Finally, maximum posterior variance between groups who score the item correct and incorrect was proposed as an item selection criterion by Wainer, Lewis, Kaplan, and Braswell (1992) (for an empirically comparison with the maximum-information criterion, see Schnipke and Green, 1995).

Large-Sample Equivalence

As is well known in Bayesian statistics, for $k\rightarrow\infty$ the posteriors $g(\theta|u_{i_1},\ldots,U_{i_k}=0)$ and $g(\theta|u_{i_1},\ldots,U_{i_k}=1)$ converge to a common (degenerate)

distribution. Consequently, $\hat{\theta}_{u_{i_1},\ldots,U_{i_k}=0}$ and $\hat{\theta}_{u_{i_1},\ldots,,U_{i_k}=1}$ tend to the same value, and asymptotically the criteria in (12)-(13) and (15) lead to the selection of the same items as the maximum-information criterion in (6). A comparable statement can be made for the criterion in (14) since both posterior variances converge to the reciprocal of Fisher's information. Thus, the choice of a Bayesian item selection criterion in adaptive testing can not be expected to lead to improved ability estimation in adaptive testing for large tests.

Statistical Properties of Criteria

For conventional linear tests, "inward" bias of the EAP and MAP estimator of θ is a typical Bayesian result. More precisely, if the prior is centered at $\theta=0$, estimators of positive values of θ are negatively biased whereas estimators of negative values show a positive bias. However, this bias is usually offset by a favorable mean-squared error, in particular if the prior is informative. Of course, the precise magnitudes of these effect depends on the choice of prior and the length of the test.

For the MLE of θ, the opposite holds. These estimators are "outwardly" biased and typically have a larger mean-squared error than Bayesian estimators. For the logistic model in (1) Lord (1983) and Samejima (1983) derive the following approximation to the bias function of the MLE:

$$\beta(\theta) \equiv E(\hat{\Theta} - \theta) \approx \frac{1}{I(\theta)^2} \sum_{j=1}^{n} a_j I_j(\theta)[p_j(\theta)-0.5], \qquad (16)$$

which is accurate to the order n^{-1}. As $I(\theta)^{-2}$ is always positive, the term $a_j I_j(\theta)[p_j(\theta)-0.5]$ in (16) allows for an evaluation of the sign of the contribution of

an individual item to the bias of the MLE. For an item with $\theta=b_i$, it holds that $p_i(\theta)=0.5$, and the corresponding term in (16) vanishes. However, if $\theta>b_i$, a positive contribution to the bias is obtained, whereas the opposite occurs for $\theta<b_i$. As a consequence, for a full test, the MLE of θ is always "biased away from where the items are".

For a conventional linear test, the bias function of the chosen ability estimator is fixed by design. In an adaptive test, however, the choice of the next item matches the current ability estimate, and, as a consequence, the bias added to the final estimator by the item is dependent on the bias already present in the current estimator. For an adaptive test with ML estimation of ability and item selection according to the criterion of maximum information in (6), the dependency creates a perfect negative feedback mechanism, as is shown by the following argument.

For the response model in (1), the maximum-information criterion in (6) implies

$$b_{i_k} = \hat{\theta}^{ML}_{u_{i_1},\ldots,u_{i_{k-1}}}. \tag{17}$$

Therefore, it follows from (16) that

$$\begin{aligned} & E(\hat{\theta}^{ML}_{U_{i_1},\ldots,U_{i_{k-1}}}-\theta) > 0 \\ & \Rightarrow E(b_{i_k}-\theta) < 0 \\ & \Rightarrow E(p_{i_k}(\theta)-0.5) < 0 \\ & \Rightarrow E(\hat{\theta}^{ML}_{U_{i_1},\ldots,U_{i_k}}-\theta) < E(\hat{\theta}^{ML}_{U_{i_1},\ldots,U_{i_{k-1}}}-\theta), \end{aligned} \tag{18}$$

whereas the opposite occurs for $E(\hat{\theta}^{ML}_{U_{i_1},\ldots,U_{i_{k-1}}} - \theta) < 0$. (In the second and third step, the expectations are taken over random selection of the kth item.) It can be concluded that each time the current estimator of θ has a negative or positive error, the next item is selected automatically to counteract the error.

As is well known, for a flat prior a Bayesian procedure with the maximum a posteriori estimator is identical to ML estimation. If the posterior is symmetric, the same holds for the EAP estimator. For an informative prior, the behavior can nicely be illustrated using Owen's equations (Appendix, Eqs. A.3-A.6). Suppose the prior is located at $\theta=0$, the difficulty of the first item has the same value, but the examinee has ability $\theta>0$. The probability of a correct response is larger than the probability of an incorrect response, and Equation A.4 shows that the EAP estimator tends to increase by a positive amount which is smaller, the more informative the prior (i.e., the smaller variance in this equation). For ability $\theta<0$, Equation 4.5 shows that the same holds in the opposite direction. Thus, for a noninformative prior the feedback mechanism above can be expected to hold (provided the posterior is symmetric), but as the prior becomes more informative the process changes and the value of the estimator can be expected to move gradually from the location of the prior to the true value of the ability parameter. If the prior strongly dominates the length of the test, large bias in the final estimator can be expected, unless the prior is located exactly at the ability of the examinee. However, larger bias does not imply a larger MSE, since it may be offset by a smaller variance of the estimator due to the information in the prior.

Whether or not a Bayesian adaptive procedure actually outperforms one with maximum-information item selection and ML estimation of ability depends on the choice of such quantities as the prior, the initial item, and the length of the test. Further, for an actual item pool, both procedures may be hampered differently by

the fact that the item parameters are not spread densely enough in certain ability intervals. Therefore, to get a more quantitative evaluation of the statistical properties of the Bayesian item selection criteria relative to the maximum-information criterion, a simulation study was run. The results for the item selection criteria in this paper complement earlier results reported for Bayesian and maximum-likelihood ability estimation for conventional linear tests (Kim & Nicewander, 1993; Warm, 1989) and adaptive tests (Warm, 1989).

Empirical Results

The item pool consisted of 300 items with response functions following the two-parameter logistic model in (1) and values for their item parameters drawn from $a_i \sim U(0.5,1.5)$ and $b_i \sim U(-4.0,4.0)$. The following procedures were compared:

1. maximum information, with $\theta \sim U(-4.0,4.0)$ as prior;
2. maximum posterior expected information, with $N(\beta_{\theta x}, \sigma^2_{\theta|x})$ as prior;
3. maximum predicted expected information, with $N(\beta_{\theta x}, \sigma^2_{\theta|x})$ as prior;
4. minimum predicted expected variance, with $N(\beta_{\theta x}, \sigma^2_{\theta|x})$ as prior;
5. maximum predicted posterior expected information, with $N(\beta_{\theta x}, \sigma^2_{\theta|x})$ as prior,

where x is assumed to be a background variable with know linear regression on θ. For estimation of empirical priors from response data and data on background variables, see van der Linden (1996).

The standard setup consisted of $\beta_{\theta x}=0.50$, $\sigma_{\theta|x}=0.87$ for the last four procedures, MAP estimation of θ during the test, and ML estimation to obtain a

final estimate of θ. The choice of ML estimation for the final estimator of θ for all procedures was motivated by the fact that background information can be used to improve the choice of the initial item but is generally not accepted as a source of data for ability estimation. Note that the combination of MAP estimation and a uniform prior for θ in the first procedure yields an MLE for θ. The bias and mean-squared error (MSE) functions of the final estimator of θ were studied as a function of test length (n=5, 10, 20, 30). Estimates of the functions were calculated for θ=-4.0(.50)4.0 on the basis of 300 replications of the procedures for each θ value.

The results are given in Figure 1. For a test length of n=5 the differences between

[Figure 1 about here]

the bias and MSE functions for the maximum posterior expected information, maximum predicted expected information, and minimum predicted expected variance criteria in (12)-(14) were negligible for all practical purposes. However, the MSE function of the maximum predicted posterior expected information criterion in (15) clearly outperformed all other criteria and was remarkably flat over the entire range of θ values. The criterion also had the best bias function for this test length. The maximum-information criterion in (6) had the worst MSE function of all criteria but yielded a bias function that was second best. Note that, unlike for conventional linear tests, this function showed inward bias, just as the functions for the Bayesian criteria. This reversal of bias is assumed to be the result of the negative feedback mechanism discussed earlier. The function also showed more favorable results than those for the Bayesian criteria in (12)-(14) for the more

extreme values of θ. However, as shown by the plot of the MSE functions, the price paid for these results by the maximum-information criterion was higher instability of the ability estimator.

For n=10, the gain in bias for the maximum-information criterion was already lost completely and both the bias and MSE functions did not show any systematic differences between the five criteria. For n=20 and n=30, the results for the maximum predicted expected information criterion fell behind but those four the other four criteria improved and did not show any systematic differences.

Table 1 gives the average CPU time per selected item for the four Bayesian

[Table 1 about here]

criteria for a 20-item test on a PC with Pentium 133 processor. The maximum predicted posterior expected information criterion took most time but a waiting time between items of 1.38 secs/item should be no problem for application of the criterion in real-life adaptive testing.

Conclusion

Application of item selection criteria for adaptive testing based on the full posterior distribution of the ability parameter is no longer hampered by costs of computing. The results in this paper show that for short tests these criteria are clearly superior to the maximum-information criterion with ML estimation of the ability parameter. The maximum predicted posterior expected information criterion showed excellent mean-squared error for extremer values of θ, and is the criterion elect for application in short adaptive tests, for example, adaptive routing tests in a

multi-stage testing format. However, as soon the test length becomes longer than 10 items, the statistical differences between the two classes of criteria are likely to be negligible for all practical purposes. This pattern is in accordance with the asymptotic equivalence of the procedures discussed earlier in the paper, albeit that the limit of 10 items is lower than was anticipated.

Appendix

Owen's (1975) equations for updating the mean and variance of a (normalized) posterior are not found in the adaptive testing literature. Using the notation in this paper, let ξ_{i_k} and ζ_{i_k} be defined as

$$\xi_{i_k} \equiv (b_{i_k} - E(\theta|u_{i_1},...,u_{i_{k-1}})/(a_{i_k}^{-2} + Var(\theta|u_{i_1},...,u_{i_{k-1}}))^{1/2} \qquad \textbf{(A.1)}$$

and

$$\zeta_{i_k} \equiv c_{i_k} + (1-c_{i_k})\Phi(\xi_{i_k}). \qquad \textbf{(A.2)}$$

Note that ξ_{i_k} can be interpreted as a standardized difference between the difficulty of item i and the value of the EAP estimator based on the previous k-1 items. If the posterior $g(\theta|u_{i_1},...,u_{i_{k-1}})$ is normal, then, for the 3-parameter normal-ogive model in (7), its expectation, $E(\theta|u_{i_1},...,u_{i_{k-1}})$, and variance, $Var(\theta|u_{i_1},...,u_{i_{k-1}})$, have updates for $U_{i_k}=0$ and $U_{i_k}=1$ that can be written as

$$E(\theta|u_{i_1},...,u_{i_{k-1}},U_{i_k}=0) = E(\theta|u_{i_1},...,u_{i_{k-1}}) - Var(\theta|u_{i_1},...,u_{i_{1-k}}).$$

$$[a_{i_k}^{-2} + Var(\theta|u_{i_1},...,u_{i_{k-1}})]^{-1/2}\phi(\xi_{i_k})\Phi(\xi_{i_k}); \qquad \textbf{(A.3)}$$

$$E(\theta|u_{i_1},...,u_{i_{k-1}},U_{i_k}=1) = E(\theta|u_{i_1},...,u_{i_{k-1}}) + (1-c_{i_k})Var(\theta|u_{i_1},...,u_{i_{1-k}}).$$

$$[a_{i_k}^{-2} + Var(\theta|u_{i_1},...,u_{i_{k-1}})]^{-1/2}\phi(\xi_{i_k})[c_{i_k} + (1-c_{i_k})\Phi(-\xi_{i_k})]; \qquad \textbf{(A.4)}$$

$$\mathrm{Var}(\theta|u_{i_1},\ldots,u_{i_{k-1}},U_{i_k}=0) = \mathrm{Var}(\theta|u_{i_1},\ldots,u_{i_{k-1}})\{1-(1+a_{i_k}^{-2}\mathrm{Var}(\theta|u_{i_1},\ldots,u_{i_{k-1}})^{-1})^{-1}\cdot$$

$$\phi(\xi_{i_k})[\phi(\xi_{i_k})/\Phi(\xi_{i_k})+\xi_{i_k}]/\Phi(\xi_{i_k})\}; \qquad \textbf{(A.5)}$$

$$\mathrm{Var}(\theta|u_{i_1},\ldots,u_{i_{k-1}},U_{i_k}=1) = \mathrm{Var}(\theta|u_{i_1},\ldots,u_{i_{k-1}})\{1-(1-c_i)(1+a_{i_k}^{-2}\mathrm{Var}(\theta|u_{i_1},\ldots,u_{i_{k-1}})^{-1})^{-1}\cdot$$

$$\phi(\xi_{i_k})[(1-c_{i_k})\phi(\xi_{i_k})/\zeta_{i_k}-\xi_{i_k}]/\zeta_{i_k}, \qquad \textbf{(A.6)}$$

where $\phi(.)$ and $\Phi(.)$ are the normal density and distribution function, respectively. If the item pool is dense enough ξ_{i_k} can be set to zero and ζ_{i_k} takes the value $0.5(1+c_{i_k})$, whereupon the equations simplify drastically. Note that the first two equations show that the next EAP estimate is equal to the previous estimate plus a correction which is negative for a wrong ($U_{i_k}=0$) and positive for a correct response ($U_{i_k}=1$). The size of the correction depends on the variance of the estimator. The last two equations show that at each update the posterior variance decreases, but that the decrease for a correct response is smaller due to the presence of the factor $1-c_{i_k}$ which accounts for the probability of guessing the correct response under the 3-parameter normal-ogive model.

References

Bloxom, B., & Vale, C.D. (1987, June). Multidimensional adaptive testing: An approximate procedure for updating. Paper presented at the annual meeting of the Psychometric Society, Montreal, Canada.

Chang, H.-H. (1996, April). A global information approach to computerized adaptive testing. Paper presented at the annual meeting of the National Council for Measurement in Education, New York City, NY.

Hambleton, R.K., & Swaminathan, H. (1985), Item response theory: Principles and applications. Boston: Kluwer-Nijhof.

Kim, J.K., & Nicewander, W.A. (1993). Ability estimation for conventional tests. Psychometrika, 58, 587-599.

Lord, F.M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. Psychometrika, 48, 233-246.

Luecht, R.M. (1995). Some alternative CAT item selection heuristics (Internal report). Philadelphia, PA: National Board of Medical Examiners.

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

Samejima, F. (1993). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. Psychometrika, 58, 195-210.

Schnipke, D.L., & Green, B.F. A comparison of item selection routines in linear and adaptive tests. Journal of Educational Measurement, 32, 227-242.

Thissen, E., & Mislevy, R.J. (1990). In H. Wainer (Ed.), Computerized adaptive testing: A primer. Hillsdale, NJ: Erlbaum.

van der Linden, W.J. (1996). A procedure for empirical initialization of adaptive testing algorithms (Research Report 96-3). Enschede, The Netherlands: University of Twente, Department of Educational Measurement and Data Analysis.

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. (in press). Some new item selection criteria for adaptive testing. Journal of Educational and Behavioral Statistics.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory with tests of finite length. Psychometrika, 54, 427-450.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 4, 473-492.

Weiss, D.J., & McBride, J.R. (1984). Bias and information of Bayesian adaptive testing. Applied Psychological Measurement, 8, 273-285.

Wainer, H., Lewis, C., Kaplan, B., & Braswell, J. (1991). Building algebra testlets: A comparison of hierarchical and linear structures. Journal of Educational Measurement, 28, 311-323.

Table 1

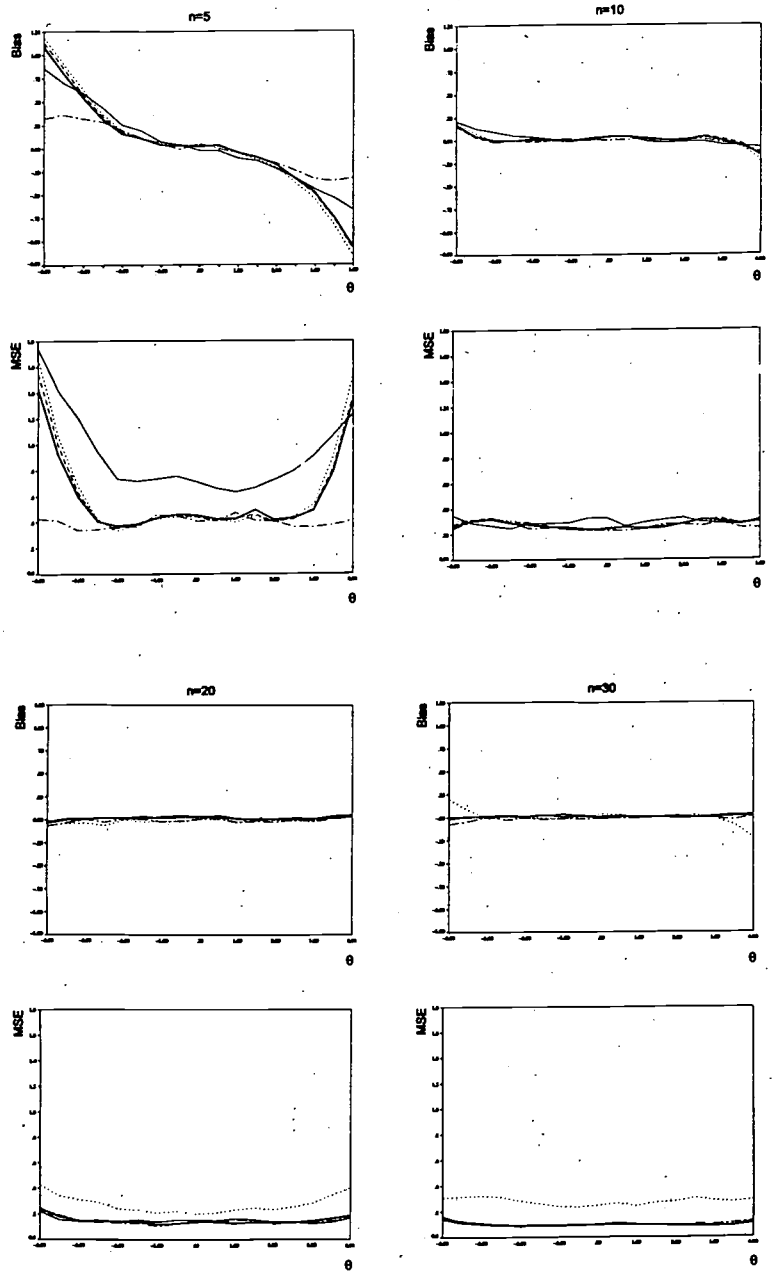
Average CPU time per selected item (pool size: 200; n=20)

Criterion	CPU Time
Maximum posterior expected information	0.19
Maximum predicted expected information	0.89
Minimum predicted expected variance	0.60
Maximum predicted posterior expected information	1.38

Note. CPU time in seconds per item

Figure Caption

Figure 1. Estimated bias and MSE functions for five item selection criteria (Maximum Information: solid; Maximum Posterior Expected Information; dashed; Maximum Predicted Expected Information: dotted; Maximum Predicted Expected Variance: bold solid; Maximum Predicted Posterior Expected Information: dashed-dotted).



Author Note

Portions of this paper were presented at the 60th annual meeting of the Psychometric Society, Minneapolis, Minnesota, June, 1995. The author is indebted to Wim M.M. Tielen for his computational support. Correspondence should be sent to: W.J. van der Linden, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email; vanderlinden@edte.utwente.nl

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.

TM027361

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC

NOTICE

REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").