Developments in Social Science Methodology
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 18, Ljubljana: FDV, 2002

Applying the Minimax Principle to Sequential Mastery Testing

Hans J. Vos[1]

Abstract

The purpose of this paper is to derive optimal rules for sequential mastery tests. In a sequential mastery test, the decision is to classify a subject as a master, a nonmaster, or to continue sampling and administering another random item. The framework of minimax sequential decision theory (minimum information approach) is used; that is, optimal rules are obtained by minimizing the maximum expected losses associated with all possible decision rules at each stage of sampling. The main advantage of this approach is that costs of sampling can be explicitly taken into account. The binomial model is assumed for the probability of a correct response given the true level of functioning, whereas threshold loss is adopted for the loss function involved. Monotonicity conditions are derived, that is, conditions sufficient for optimal rules to be in the form of sequential cutting scores. The paper concludes with a simulation study, in which the minimax sequential strategy is compared with other procedures that exist for similar classification decision problems in the literature.

1 Introduction

Well-known examples of fixed-length mastery tests include pass/fail decisions in education, certification, and successfulness of therapies. The fixed-length mastery problem has been studied extensively in the literature within the framework of (empirical) Bayesian decision theory (e.g., De Gruijter & Hambleton, 1984; van der Linden, 1990). In addition, optimal rules for the fixed-length mastery problem

[1] University of Twente, Faculty of Educational Science and Technology, e-mail: vos@edte.utwente.nl.

The author is indebted to Wim J. van der Linden and Sebie J. Oosterloo for their valuable comments and to Frans Houweling for his computational support in developing the simulation study.

have also been derived within the framework of the minimax strategy (e.g., Huynh, 1980; Veldhuijzen, 1982).

In both approaches, the following two basic elements are distinguished: A psychometric model relating the probability of a correct response to student's (unknown) true level of functioning, and a loss structure evaluating the total costs and benefits for each possible combination of decision outcome and true level of functioning. Within the framework of Bayesian decision theory (e.g., DeGroot, 1970; Lehmann, 1959), optimal rules (i.e., Bayes rules) are obtained by minimizing the posterior expected losses associated with all possible decision rules. Decision rules are hereby prescriptions specifying for each possible observed response pattern what action has to be taken. The Bayes principle assumes that prior knowledge about student's true level of functioning is available and can be characterized by a probability distribution called the prior.

Using minimax decision theory (e.g., DeGroot, 1970; Lehmann, 1959), optimal rules (i.e., minimax rules) are obtained by minimizing the maximum expected losses associated with all possible decision rules. In fact, the minimax principle assumes that it is best to prepare for the worst and to establish the maximum expected loss for each possible decision rule (e.g., van der Linden, 1981). In other words, the minimax decision rule is a bit conservative and pessimistic (Coombs, Dawes, & Tversky, 1970).

The test at the end of the treatment does not necessarily have to be a fixed-length mastery test but might also be a variable-length mastery test. In this case, in addition to the actions declaring mastery or nonmastery, also the action of continuing sampling and administering another item is available. Variable-length mastery tests are designed with the goal of maximizing the probability of making correct classification decisions (i.e., mastery and nonmastery) while at the same time minimizing test length (Lewis & Sheehan, 1990). For instance, Ferguson (1969) showed that average test lengths could be reduced by half without sacrificing classification accuracy.

Generally, two main types of variable-length mastery tests can be distinguished. First, both the item selection and stopping rule (i.e., the termination criterion) are adaptive. Student's ability measured on a latent continuum is estimated after each response, and the next item is selected such that its difficulty matches student's last ability estimate. Hence, this type of variable-length mastery testing assumes that items differ in difficulty, and is denoted by Kingsbury and Weiss (1983) as adaptive mastery testing (AMT).

In the second type of variable-length mastery testing, the stopping rule only is adaptive but the item to be administered next is selected random. In the following, this type of variable-length mastery testing will be denoted as sequential mastery testing (SMT). The purpose of this paper is to derive optimal rules for SMT using the framework of minimax sequential decision theory (e.g., DeGroot, 1970; Lehmann, 1959). The main advantage of this approach is that costs of sampling (i.e., administering another random item) can be explicitly taken into account.

2 Review of existing procedures to variable-length mastery testing

In this section, earlier solutions to both the adaptive and sequential mastery problem will be briefly reviewed. First, earlier solutions to AMT will be considered. Next, it will be indicated how SMT has been dealt with in the literature.

2.1 Earlier solutions to adaptive mastery testing

In adaptive mastery testing, two item response theory (IRT)-based strategies have been primarily used for selecting the item to be administered next. First, Kingsbury and Weiss (1983) proposed the item to be administered next is the one that maximizes the amount of (Fisher's) information at student's last ability estimate.

In the second IRT-based approach, the Bayesian item selection strategy, the item that minimizes the posterior variance of student's last ability estimate is administered next. In this approach, a prior distribution about student's ability must be specified. If a normal distribution is assumed as a prior, an estimate of the posterior distribution of student's last ability, given observed test score, may be obtained via a procedure called restricted Bayesian updating (Owen, 1975). Also, posterior variance may be obtained via Owen's Bayesian scoring algorithm. Nowadays, numerical procedures for computing posterior ability and variance do also exist.

Both IRT-based item selection procedures make use of confidence intervals of student's latent ability for deciding on mastery, nonmastery, or to continue sampling. Decisions are made by determining whether or not a prespecified cut-off point on the latent IRT-metric, separating masters from nonmasters, falls outside the limits of this confidence interval.

As an aside, as pointed out by Chang and Stout (1993), it may be noted that the posterior variance converges to the reciprocal of the test information when the number of items goes to infinity. Therefore, the two methods of IRT-based item selection strategies should yield similar results when the item number is large.

2.2 Existing procedures to the sequential mastery problem

One of the earliest approaches to sequential mastery testing dates back to Ferguson (1969) using Wald's well-known sequential probability ratio test (SPRT), originally developed as a statistical quality control test for light bulbs in a manufacturing setting. In Ferguson's approach, the probability of a correct

response given the true level of functioning (i.e., the psychometric model) is modeled as a binomial distribution. The choice of this psychometric model assumes that, given the true level of functioning, each item has the same probability of being correctly answered, or that items are sampled at random.

As indicated by Ferguson (1969), three elements must be specified in advance in applying the SPRT-framework to sequential mastery testing. First, two values p_0 and p_1 on the proportion-correct metric must be specified representing points that correspond to lower and upper limits of true level of functioning at which a mastery and nonmastery decision will be made, respectively. Also, these two values mark the boundaries of the small region (i.e., indifference region) where we never can be sure to take the right classification decision, and, thus, in which sampling will continue. Second, two levels of error acceptance α and β must be specified, reflecting the relative costs of the false positive (i.e., Type I) and false negative (i.e., Type II) error types. Intervals can be derived as functions of these two error rates for which mastery and nonmastery is declared, respectively, and for which sampling is continued (Wald, 1947). Third, a maximum test length must be specified in order to classify within a reasonable period of time those students for whom the decision of declaring mastery or nonmastery is not as clear-cut.

Reckase (1983) has proposed an alternative approach to sequential mastery testing within an SPRT-framework. Unlike Ferguson (1969), Reckase (1983) did not assume that items have equal characteristics but allowed them to vary in difficulty and discrimination by using an IRT-model instead of a binomial distribution. Modeling response behavior by an IRT model, as in Reckase's (1983) model, Spray and Reckase (1996) compared Wald's SPRT procedure also with a maximum information item selection (MIIS) procedure (Kingsbury and Weiss, 1983). The results showed that under the conditions studied, the SPRT procedure required fewer test items than the MIIS procedure to achieve the same level of classification accuracy. This finding is consistent with Wald's (1947) conclusion that the SPRT was the uniformly most powerful test of simple hypotheses.

Recently, Lewis and Sheehan (1990), Sheehan and Lewis (1992), and Smith and Lewis (1995) have applied Bayesian sequential decision theory (e.g., DeGroot, 1970; Lehmann, 1959) to SMT. In addition to a psychometric model and a loss function, cost of sampling (i.e., cost of administering one additional item) must be explicitly specified in this approach. Doing so, posterior expected losses associated with the nonmastery and mastery decisions can now be calculated at each stage of sampling. As far as the posterior expected loss associated with to continue sampling concerns, this quantity is determined by averaging the posterior expected losses associated with each of the possible future decision outcomes relative to the probability of observing those outcomes (i.e., the posterior predictive distributions).

Optimal rules (i.e., Bayesian sequential rules) are now obtained by choosing the action that minimizes posterior expected loss at each stage of sampling using

techniques of dynamic programming (i.e., backward induction). This technique starts by considering the final stage of sampling and then works backward to the first stage of sampling. Backward induction makes use of the principle that upon breaking into an optimal procedure at any stage, the remaining portion of the procedure is optimal when considered in its own right. Doing so, as pointed out by Lewis and Sheehan (1990), the action chosen at each stage of sampling is optimal with respect to the entire sequential mastery testing procedure.

Lewis and Sheehan (1990) and Sheehan and Lewis (1992), as in Reckase's approach, modeled response behavior in the form of a 3-parameter logistic (PL) model from IRT. The number of possible outcomes of future random item administrations, needed in computing the posterior expected loss associated with the continuing sampling option, can become very quick quite large. Lewis and Sheehan (1990), therefore, made the simplification that the number-correct score in the 3-PL model is sufficient for calculating the posterior predictive distributions rather than the entire pattern of item responses.

As an aside, it may be noted that Lewis and Sheehan (1990), Sheehan and Lewis (1992), and Smith and Lewis (1995) used testlets (i.e., blocks of items) rather than single items.

Vos (1999) also applied the framework of Bayesian sequential decision theory to SMT. As in Ferguson's (1969) approach, however, the binomial distribution instead of an IRT-model is considered for modeling response behavior. It is shown that for the binomial distribution, in combination with the assumption that prior knowledge about student's true level of functioning can be represented by a beta prior (i.e., its natural conjugate), the number-correct score is sufficient to calculate the posterior expected losses at future stages of item administrations. Unlike the Lewis and Sheehan (1990) model, therefore, no simplifications are necessary to deal with the combinatorial problem of the large number of possible decision outcomes of future item administrations.

3 Minimax sequential decision theory applied to SMT

In this section, the framework of minimax sequential decision theory (e.g., DeGroot, 1970; Lehmann, 1959) will be treated in more detail. Also, a rationale is provided for why this approach should be preferred above the Bayesian sequential principle.

3.1 Framework of minimax sequential decision theory

In minimax sequential decision theory, optimal rules (i.e., minimax sequential rules) are found by minimizing the maximum expected losses associated with all possible decision rules at each stage of sampling. Analogous to Bayesian

sequential decision theory, cost per observation is also explicitly been taken into account in this approach. Hence, the maximum expected losses associated with the mastery and nonmastery decisions can be calculated at each stage of sampling. The maximum expected loss associated with the continuing sampling option is computed by averaging the maximum expected losses associated with each of the possible future decision outcomes relative to the posterior predictive probability of observing those outcomes.

Unlike Bayesian sequential decision theory, specification of a prior is not needed in applying the minimax sequential principle. A minimax sequential rule, however, can be conceived of as a rule that is based on minimization of posterior expected loss as well (i.e., as a Bayesian sequential rule), but under the restriction that the prior is the least favorable element of the class of priors (e.g., Ferguson, 1967).

3.2 Rationale for preferring the minimax principle above the Bayesian principle

The question can be raised why minimax sequential decision theory should be preferred above the Bayesian sequential principle. As pointed out by Huynh (1980), the minimax (sequential) principle is very attractive when the only information is student's observed number-correct score; that is, no group data of 'comparable' students who will take the same test or prior information about the individual student is available. The minimax strategy, therefore, is sometimes also denoted as a *minimum information* approach (e.g., Veldhuijzen, 1982).

If group data of 'comparable' students or prior information about the individual student is available, however, it is better to use this information. Hence, in this situation it is better to use Bayesian instead of minimax sequential decision theory. Even if information in the form of group data of 'comparable' students or prior information about the individual student is available, it is sometimes too difficult a job to accomplish to express this information into a prior distribution (Veldhuijzen, 1982). In these circumstances, the minimax sequential procedure may also be more appropriate.

4 Notation

Within the framework of both minimax and Bayesian sequential decision theory, optimal rules can be obtained without specifying a maximum test length. In the following, however, a sequential mastery test is supposed to have a maximum test length n ($n \geq 1$). As pointed out by Ferguson (1969), a maximum test length is needed in order to classify within a reasonable period of time those students for whom the decision of declaring mastery or nonmastery is not as clear-cut.

Let the observed item response at each stage of sampling k ($1 \leq k \leq n$) for a randomly sampled student be denoted by a discrete random variable X_k, with realization x_k. The observed response variables $X_1,...,X_k$ are assumed to be independent and identically distributed for each value of k, and take the values 0 and 1 for respectively incorrect and correct responses to the k-th item. Furthermore, let the observed number-correct score be denoted by a discrete random variable $S_k = X_1 +...+ X_k$, with realization $s_k = x_1 +...+ x_k$ ($0 \leq s_k \leq k$).

Student's true level of functioning is unknown due to measurement and sampling error. All that is known is his/her observed number-correct score s_k. In other words, the mastery test is not a perfect indicator of student's true performance. Therefore, let student's true level of functioning be denoted by a continuous random variable T on the latent proportion-correct metric, with realization $t \in [0,1]$.

Finally, a criterion level t_c ($0 \leq t_c \leq 1$) on the true level of functioning scale T can be identified. A student is considered a true nonmaster and true master if his/her true level of functioning t is smaller or larger than t_c, respectively. The criterion level must be specified in advance by the decision-maker. Several methods for setting standards on the observed score level have been proposed in the literature (e.g., Angoff, 1971; Nedelsky, 1954). However, these standard setting methods do not apply to the true level of functioning T. The criterion level t_c on the true level of functioning T, therefore, must be set by content experts by indicating the minimal percentage of the total domain of items a student must be able to answer correctly in order to be declared mastery status.

Assuming $X_1 = x_1,...,X_k = x_k$ has been observed, the two basic elements of minimax sequential decision making discussed earlier can now be formulated as follows: A psychometric model $f(s_k \mid t)$ relating observed number-correct score s_k to student's true level of functioning t at each stage of sampling k, and a loss function describing the loss $l(a_i(x_1,...,x_k), t)$ incurred when action $a_i(x_1,...,x_k)$ is taken for the student whose true level of functioning is t. The actions nonmastery, mastery, and to continue sampling will be denoted as $a_0(x_1,...,x_k)$, $a_1(x_1,...,x_k)$, and $a_2(x_1,...,x_k)$, respectively.

5 Threshold loss

Generally speaking, as noted before, a loss function evaluates the total costs and benefits of all possible decision outcomes for a student whose true level of functioning is t. These costs may concern all relevant psychological, social, and economic consequences which the decision brings along. The Bayesian as well as minimax approach allows the decision-maker to incorporate into the decision process the costs of misclassifications (i.e., students for whom the wrong decision is made). As in Hambleton and Novick (1973), here the well-known threshold loss

function is adopted as the loss structure involved. The choice of this loss function implies that the "seriousness" of all possible consequences of the decisions can be summarized by possibly different constants, one for each of the possible classification outcomes.

For the sequential mastery problem, a threshold loss function can be formulated as a natural extension of the one for the fixed-length mastery problem at each stage of sampling k as follows (see also Lewis & Sheehan, 1990):

Table 1: Table for threshold loss function at stage k ($1 \leq k \leq n$) of sampling.

True Level of Functioning Action	$T \leq t_c$	$T > t_c$
$a_0(x_1, ..., x_k)$	ke	$l_{01} + ke$
$a_1(x_1, ..., x_k)$	$l_{10} + ke$	ke

The value e represents the costs of administering one random item. For the sake of simplicity, following Lewis and Sheehan (1990), these costs are assumed to be equal for each classification outcome as well as for each sampling occasion. Of course, these two assumptions can be relaxed in specific sequential mastery testing applications. Applying an admissible positive linear transformation (e.g., Luce & Raiffa, 1957), and assuming the losses l_{00} and l_{11} associated with the correct classification outcomes are equal and take the smallest values, the threshold loss function in Table 1 was rescaled in such a way that l_{00} and l_{11} were equal to zero. Hence, the losses l_{01} and l_{10} must take positive values.

Note that no losses need to be specified in Table 1 for the continuing sampling action ($a_2(x_1,...,x_k)$). This is because the maximum expected loss associated with the continuing sampling option is computed at each stage of sampling as a weighted average of the maximum expected losses associated with the classification decisions (i.e., mastery/nonmastery) of future item administrations with weights equal to the probabilities of observing those outcomes.

The ratio l_{10}/l_{01} is denoted as the loss ratio R, and refers to the relative losses for declaring mastery to a student whose true level of functioning is below t_c (i.e., false positive) and declaring nonmastery to a student whose true level of functioning exceeds t_c (i.e., false negative).

The loss parameters l_{ij} ($i = 1,2$; $i \neq j$) associated with the incorrect decisions have to be empirically assessed, for which several methods have been proposed in the literature. Most texts on decision theory, however, propose lottery methods (e.g., Luce & Raiffa, 1957) for assessing loss functions empirically. In general, the consequences of each pair of actions and true level of functioning are scaled in

these methods by looking at the most and least preferred outcomes. But, in principle, any psychological scaling method can be used.

6 Psychometric model

As earlier remarked, here the well-known binomial model will be adopted for specifying the statistical relation between the observed number-correct score s_k and student's true level of functioning t. Its distribution $f(s_k \mid t)$ at stage k of sampling, given student's true level of functioning t, can be written as follows:

$$f(s_k \mid t) = \binom{k}{s_k} t^{s_k} (1-t)^{k-s_k} . \qquad (1)$$

If each response is independent of the other, and if the examinee's probability of a correct answer remains constant, the distribution function of s_k, given student's true level of functioning t, is given by Equation 1 (Wilcox, 1981). The binomial model assumes that the test given to each student is a random sample of items drawn from a large (real or imaginary) item pool (Wilcox, 1981). Therefore, for each student a new random sample of items must be drawn in practical applications of the sequential mastery problem.

7 Sufficient conditions for minimax sequential rules to be monotone

Linking up with common practice in mastery testing, minimax sequential rules in this paper are assumed to have monotone forms. Decision rules in practical situations in education and psychology usually take the form of selecting one or more cutting scores on the test. Decision rules of this form constitute a special subclass known as monotone rules (Ferguson. 1967, Sect. 6.1). In other words, a decision rule is monotone if cutting scores are used to partition the test scores into intervals for which different actions are taken. As a result, monotone sequential rules can be defined on the number-correct score metric in the form of sequential cutting scores. The restriction to monotone rules, however, is correct only if it can be proven that for any nonmonotone rule for the problem at hand there is a monotone rule with at least the same value on the criterion of optimality used (Ferguson, 1967, p. 55). Using a minimax sequential rule, as noted before, the minimum of the maximum expected losses associated with all possible decision rules is taken as the criterion of optimality at each stage of sampling.

As noted before, the maximum expected loss for continuing sampling is hereby determined by averaging the maximum expected losses associated with each of the possible future decision outcomes relative to the probability of observing those outcomes. Therefore, it follows immediately that the conditions sufficient for setting cutting scores for the fixed-length mastery problem are also sufficient for the sequential mastery problem at each stage of sampling.

Generally, conditions sufficient for setting cutting scores for the fixed-length mastery problem are given in Ferguson (1967). First, $f(s_k \mid t)$ must have a monotone likelihood ratio (MLR); that is, it is required that for any $t_1 > t_2$, the likelihood ratio $f(s_k \mid t_1) / f(s_k \mid t_2)$ is a nondecreasing function of s_k. MLR implies that the higher the observed number-correct score, the more likely it will be that the true level of functioning is high too. Second, the condition of monotonic loss must hold; that is, there must be an ordering of the actions such that for each pair of adjacent actions the loss functions possess at most one point of intersection.

In our example the binomial density function is chosen as the psychometric model $f(s_k \mid t)$. Since the binomial model belongs to the monotone likelihood ratio family (Ferguson, 1967, Chap. 5), it then follows that the condition of MLR is satisfied. Furthermore, by choosing $l_{00} = l_{11} = 0$ and assuming positive values for l_{01} and l_{10}, it follows that for each pair of adjacent actions the loss functions don't possess a point of intersection. Hence, it follows immediately that the condition of monotonic loss is also satisfied at each stage of sampling k.

8 Optimizing rules for the sequential mastery problem

In this section, it will be shown how optimal rules for SMT can be derived using the framework of minimax sequential decision theory. Doing so, given an observed item response vector $(x_1,...,x_k)$, first the minimax principle will be applied to the fixed-length mastery problem by determining which of the maximum expected losses associated with the two classification actions $a_0(x_1,...,x_k)$ or $a_1(x_1,...,x_k)$ is the smallest. Next, applying the minimax principle again, optimal rules for the sequential mastery problem are derived at each stage of sampling k by comparing this quantity with the maximum expected loss associated with action $a_2(x_1,...,x_k)$ (i.e., continuing sampling).

8.1 Applying the minimax principle to the fixed-length mastery problem

Given $X_1 = x_1,...,X_k = x_k$, as noted before, the minimax decision rule for the fixed-length mastery problem can be found by minimizing the maximum expected losses

associated with the two classification actions $a_0(x_1,...,x_k)$ and $a_1(x_1,...,x_k)$. It is assumed that there exists a cutting score on S_k, say $s_c(k)$ $(0 \le s_c(k) \le k)$, such that mastery is declared when $s_k \ge s_c(k)$ and that nonmastery is declared otherwise. Let $y = 0,1,...,k$ represent all possible values the number-correct score s_k can take after having observed k item responses, assuming the conditions of monotonicity are satisfied, it then can easily be verified from Table 1 and Equation 1 that mastery $(a_1(x_1,...,x_k))$ is declared when the maximum loss associated with the mastery decision is smaller than the maximum loss associated with the nonmastery decision, or, equivalently, when number-correct score s_k is such that

$$\sup_{t \le t_c} (l_{10} + ke) \sum_{y=s_k}^{k} \binom{k}{y} t^y (1-t)^{k-y} + \sup_{t > t_c} (ke) \sum_{y=0}^{s_k-1} \binom{k}{y} t^y (1-t)^{k-y} <$$

$$\sup_{t \le t_c} (ke) \sum_{y=s_k}^{k} \binom{k}{y} t^y (1-t)^{k-y} + \sup_{t > t_c} (l_{01} + ke) \sum_{y=0}^{s_k-1} \binom{k}{y} t^y (1-t)^{k-y}, \qquad (2)$$

and that nonmastery $(a_0(x_1,...,x_k))$ is declared otherwise. Since the cumulative binomial distribution function is decreasing in t, it follows that the inequality in (2) can be written as:

$$(l_{10} + ke) \sum_{y=s_k}^{k} \binom{k}{y} t_c^y (1-t_c)^{k-y} + (ke) \sum_{y=0}^{s_k-1} \binom{k}{y} t_c^y (1-t_c)^{k-y} <$$

$$(ke) \sum_{y=s_k}^{k} \binom{k}{y} t_c^y (1-t_c)^{k-y} + (l_{01} + ke) \sum_{y=0}^{s_k-1} \binom{k}{y} t_c^y (1-t_c)^{k-y}. \qquad (3)$$

Rearranging terms, it follows that mastery is declared when number-correct score s_k is such that:

$$\sum_{y=s_k}^{k} \binom{k}{y} t_c^y (1-t_c)^{k-y} < 1/(1+R), \qquad (4)$$

where R denotes the loss ratio (i.e., $R = l_{10}/l_{01}$). If the inequality in (4) is not satisfied, nonmastery is declared.

8.2 Derivation of minimax sequential rules

Let $d_k(x_1,...,x_k)$ denote the action $a_0(x_1,...,x_k)$ or $a_1(x_1,...,x_k)$ yielding the minimum of the maximum expected losses associated with these two classification actions,

and let the maximum expected loss associated with this minimum be denoted as $V_k(x_1,...,x_k)$. These notations can also be generalized to the situation that no observations have been taken yet; that is, $d_0(x_0)$ denotes the action $a_0(x_0)$ or $a_1(x_0)$ which yields the smallest of the maximum expected losses associated with these two actions, and $V_0(x_0)$ denotes the smallest maximum expected loss associated with $d_0(x_0)$.

Minimax sequential rules can now be found by using the following backward induction computational scheme: First, the minimax sequential rule at the final stage of sampling n is computed. Since the continuing sampling option is not available at this stage of sampling, it follows immediately that the minimax sequential rule is given by $d_n(x_1,...,x_n)$; its associated maximum expected loss is given by $V_n(x_1,...,x_n)$.

Subsequently, the minimax sequential rule at the next to last stage of sampling (n-1) is computed by comparing $V_{n-1}(x_1,...,x_{n-1})$ with the maximum expected loss associated with action $a_2(x_1,...,x_{n-1})$ (i.e., continuing sampling). As noted before, the maximum expected loss associated with taking one more observation, given a response pattern $(x_1,...,x_{n-1})$, is computed by averaging the maximum expected losses associated with each of the possible future decision outcomes at the final stage n relative to the probability of observing those outcomes (i.e., backward induction).

Let $P(X_n = x_n \mid x_1,...,x_{n-1})$ denote the distribution of X_n, given the observed item response vector $(x_1,...,x_{n-1})$, then, the maximum expected loss associated with taking one more observation after (n-1) observations have been taken, $E[V_n(x_1,...,x_{n-1}, X_n) \mid x_1,...,x_{n-1}]$, is computed as follows:

$$E[V_n(x_1,...,x_{n-1},X_n) \mid x_1,...,x_{n-1}] = \sum_{x_n=0}^{x_n=1} V_n(x_1,...,x_n)P(X_n = x_n \mid x_1,...,x_{n-1}), \quad (5)$$

Generally, $P(X_k = x_k \mid x_1,...,x_{k-1})$ is called the posterior predictive distribution of X_k at stage (k-1) of sampling. It will be indicated later on how this conditional distribution can be computed.

Given a response pattern $(x_1,...,x_{n-1})$, the minimax sequential rule at stage (n-1) of sampling is now given by: Take one more observation if $E[V_n(x_1,...,x_{n-1}, X_n) \mid x_1,...,x_{n-1}]$ is smaller than $V_{n-1}(x_1,...,x_{n-1})$, and take action $d_{n-1}(x_1,...,x_{n-1})$ otherwise. If $E[V_n(x_1,...,x_{n-1}, X_n) \mid x_1,...,x_{n-1}]$ and $V_{n-1}(x_1,...,x_{n-1})$ are equal to each other, it does not matter whether or not the decision-maker takes one more observation.

To compute the maximum expected loss associated with the continuing sampling option, it is convenient to introduce the risk at each stage of sampling k, which will be denoted as $R_k(x_1,...,x_k)$. Let the risk at stage n of sampling be defined

as $V_n(x_1,...,x_n)$. Generally, given a response pattern $(x_1,...,x_{k-1})$, the risk at stage $(k-1)$ is then computed inductively as a function of the risk at stage k as follows:

$$R_{k-1}(x_1,...,x_{k-1}) = min\{V_{k-1}(x_1,...,x_{k-1}), E[R_k(x_1,...,x_{k-1}, X_k) \mid x_1,...,x_{k-1}]\}. \quad (6)$$

The maximum expected loss associated with taking one more observation after $(n-2)$ observations, $E[R_{n-1}(x_1,...,x_{n-2}, X_{n-1}) \mid x_1,...,x_{n-2}]$, can then be computed as the expected risk at stage $(n-1)$ as follows:

$$E[R_{n-1}(x_1,...,x_{n-2},X_{n-1}) \mid x_1,...,x_{n-2}] =$$

$$\sum_{x_{n-1}=0}^{x_{n-1}=1} R_{n-1}(x_1,...,x_{n-1})P(X_{n-1}=x_{n-1} \mid x_1,...,x_{n-2}). \quad (7)$$

Given $(x_1,...,x_{n-2})$, the minimax sequential rule at stage $(n-2)$ of sampling is now given by: Take one more observation if $E[R_{n-1}(x_1,...,x_{n-2}, X_{n-1}) \mid x_1,...,x_{n-2}]$ is smaller than $V_{n-2}(x_1,...,x_{n-2})$; otherwise, action $d_{n-2}(x_1,...,x_{n-2})$ is taken. In the case of equality between $V_{n-2}(x_1,...,x_{n-2})$ and $E[R_{n-1}(x_1,...,x_{n-2}, X_{n-1}) \mid x_1,...,x_{n-2}]$, it does not matter again whether or not the decision-maker takes one more observation.

Following the same computational backward scheme as in determining the minimax sequential rules at stages $(n-1)$ and $(n-2)$, the minimax sequential rules at stages $(n-3),...,1,0$ are computed. The minimax sequential rule at stage 0 denotes the decision whether or not to take at least one observation.

9 Computation of posterior predictive probabilities

As can be seen from (5) and (7), the posterior predictive distribution $P(X_k = x_k \mid x_1,...,x_{k-1})$ is needed for computing the maximum expected loss associated with taking one more observation at stage $(k-1)$ of sampling. From Bayes' theorem, it follows that:

$$P(X_k = x_k \mid x_1,...,x_{k-1}) = P(X_1 = x_1,...,X_k = x_k)/P(X_1 = x_1,...,X_{k-1} = x_{k-1}) \quad (8)$$

For the binomial distribution as the psychometric model involved and assuming the beta distribution $B(\alpha, \beta)$ as prior with parameters α and β $(\alpha, \beta > 0)$, it is known (e.g., Keats & Lord, 1962) that the unconditional distribution of $(X_1,...,X_k)$ is equal to:

$$P(X_1 = x_1,...,X_k = x_k) = [\Gamma(\alpha+\beta)\Gamma(\alpha+s_k)\Gamma(\beta+k-s_k)]/[\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+k)] \quad (9)$$

where Γ is the usual gamma function. From (8)-(9) it then follows that the posterior predictive distribution of X_k, given a response pattern $(x_1,...,x_{k-1})$, can be written as:

$$P(X_k=x_k \mid x_1,...,x_{k-1})=[\Gamma(\alpha+s_k)\Gamma(\beta+k-s_k)\Gamma(\alpha+\beta+k-1)]/[\Gamma(\alpha+s_{k-1})\Gamma(\beta+k-1-s_{k-1})\Gamma(\alpha+\beta+k)]. \quad (10)$$

Using the well-known identity $\Gamma(j+1) = j\Gamma(j)$ and the fact that $s_k = s_{k-1}$ and $s_k = s_{k-1}+1$ for $x_k = 0$ and 1, respectively, it follows from (10) that:

$$P(X_k = x_k \mid x_1,...,x_{k-1}) = \begin{cases} (\beta+k-1-s_{k-1})/(\alpha+\beta+k-1) & \text{if } x_k = 0 \\ (\alpha+s_{k-1})/(\alpha+\beta+k-1) & \text{if } x_k = 1. \end{cases} \quad (11)$$

10 Illustration of computing the appropriate action

To illustrate the computation of the appropriate action (i.e., mastery, nonmastery, continuation) outlined above, suppose that the maximum test length is 25 (i.e., n = 25). First, the appropriate classification decision (i.e., declare mastery or nonmastery) and its associated maximum expected loss at the final stage of testing, $d_{25}(x_1,\ldots,x_{25})$ and $V_{25}(x_1,\ldots,x_{25})$, are then computed for all possible values of s_{25} (i.e., s_{25} = 0,…,25). More specifically, mastery is declared for those values of s_{25} for which the inequality in (4) holds, while nonmastery is declared otherwise.

Likewise, the appropriate classification decision and its associated maximum expected loss are computed after 24 items have been administered (i.e., $d_{24}(x_1,\ldots,x_{24})$ and $V_{24}(x_1,\ldots,x_{24})$) for s_{24} = 0,…,24. Next, the maximum expected loss associated with administering one more random item after 24 items have been administered, $E[R_{25}(x_1,\ldots,x_{24}, X_{25} \mid x_1,\ldots,x_{24}]$, is computed using (5) and (11) for s_{24} = 0,…,24. Another random item is administered if these values are smaller than $V_{24}(x_1,\ldots,x_{24})$, and otherwise classification decision $d_{24}(x_1,\ldots,x_{24})$ is taken.

For computing the appropriate action after 23 items have been administered, in addition to computing $d_{23}(x_1,\ldots,x_{23})$ and $V_{23}(x_1,\ldots,x_{23})$ for s_{23} = 0,…,23, the risk $R_{24}(x_1,\ldots,x_{24})$ at stage 24 of testing is computed using (6) for s_{24} = 0,…,24. The maximum expected loss associated with administering one more random item after 23 items have been administered, $E[R_{24}(x_1,\ldots,x_{23}, X_{24} \mid x_1,\ldots,x_{23}]$, can then be computed as the expected risk using (7) and (11) for s_{23} = 0,…23. One more random item is now administered if these values are smaller than $V_{23}(x_1,\ldots,x_{23})$; otherwise, classification decision $d_{23}(x_1,\ldots,x_{23})$ is taken. Similarly, the appropriate action is determined at stage 22 until stage 0 of testing.

11 Determination of the least favorable prior

To be able to compute the posterior predictive distribution $P(X_k = x_k \mid x_1,...,x_{k-1})$, the form of the assumed beta prior $B(\alpha, \beta)$ must be specified more specifically, that is, the numerical values of its parameters α and β ($\alpha, \beta > 0$) must be determined. In the present paper the least favorable prior will be taken for $B(\alpha, \beta)$, as will be shown in this section, which results if for β the value 1 is taken and if α is taken sufficiently small. It should be noted, however, that other forms of the beta prior (e.g., the uniform prior with $\alpha = \beta = 1$) might also be considered in computing the posterior predictive distribution.

Let $I_p(r,s)$ denote the incomplete beta function with parameters r and s ($r, s > 0$). It has been known for some time that

$$\sum_{x=m}^{n} \binom{n}{x} p^x (1-p)^{n-x} = I_p(m, n-m+1). \qquad (12)$$

Hence, the inequality in (4) can be written as:

$$I_{t_c}(s_k, k-s_k+1) \; < \; 1/(1+R). \qquad (13)$$

Within the framework of Bayesian decision theory, given a response pattern $(x_1,...,x_k)$, it can easily be verified from Table 1 that mastery is declared for the fixed-length mastery problem if number-correct score s_k is such that

$$(l_{10}+ke)P(T \le t_c / s_k) + (ke)P(T > t_c / s_k) \; < \; (ke)P(T \le t_c / s_k) + (l_{01}+ke)P(T > t_c / s_k), \qquad (14)$$

and that nonmastery is declared otherwise. Rearranging terms, it can easily be verified from (14) that mastery is declared if

$$P(T \le t_c / s_k) \; < \; 1/(1+R), \qquad (15)$$

and that nonmastery is declared otherwise.

Assuming a beta prior, it follows from an application of Bayes' theorem that under the assumed binomial model from (1), the posterior distribution of T will be a member of the beta family again (the conjugacy property, see, e.g., Lehmann, 1959). In fact, if the beta function

$B(\alpha, \beta)$ with parameters α and β ($\alpha, \beta > 0$) is chosen as the prior distribution and student's observed number-correct score is s_k from a test of length k, then the posterior distribution of T is $I_t(\alpha + s_k, k - s_k + \beta)$.

Hence, assuming a beta prior, it follows from (15) that mastery is declared if:

$$I_{t_C}(\alpha+s_k,\ k-s_k+\beta) \ < \ 1/(1+R), \qquad (16)$$

and that nonmastery is declared otherwise.

Thus, comparing (13) and (16) with each other, it can be seen that the least favorable prior for the minimax solution is given by a beta prior $B(\alpha, \beta)$ with $\beta = 1$ and α sufficiently small. It should be noted that the parameter $\alpha > 0$ can not be chosen equal to zero, because otherwise the prior distribution for T should be improper; that is, the prior does not integrate to 1 but to infinity.

12 Simulation of different strategies for variable-length mastery testing

In a Monte Carlo simulation the minimax sequential strategy will be compared with other existing approaches to both sequential and adaptive mastery testing. More specifically, four variable-length mastery testing strategies described in detail in Kingsbury and Weiss (1983) (see also, Weiss & Kingsbury, 1984) will be used here as a comparison in terms of average test length (i.e., the number of items that must be administered on the average before a mastery/nonmastery decision is made), correspondence between the simulated students' true mastery status and his/her estimated mastery status as indexed by the Loevinger's coefficient *H*, and coefficient *H* as a function of average test length.

12.1 Description of the testing strategies used for comparison

The first comparison will be made with a conventional fixed-length test (CT) in which student performance was recorded as proportion of correct answers (CT/PC). The student was declared a master for answering 60% or more items correctly after completion of the test, whereas nonmastery was declared otherwise.

In order to determine whether the scoring method possibly accounts for differences between a Bayesian-scored AMT algorithm and the CT/PC procedure, the second comparison will be made with a conventional test where item responses were converted by Owen's Bayesian scoring procedure (CT/B) to a latent ability on an IRT-metric, assuming a standard normal prior $N(0,1)$. Mastery was declared if the final posterior estimate of student's latent ability was higher than the prespecified cut-off point on the latent IRT-metric corresponding to 60% correct; otherwise nonmastery was declared. The cut-off point on the latent IRT-metric was hereby determined by transforming the proportion-correct of 0.6 through the use of the test response function (TRF), that is, the mean of the item response functions for all items in the pool.

The third comparison will be made with Wald's SPRT procedure. The limits of the indifference region in which sampling will continue were set at proportion-correct values p_0 and p_1 of 0.5 and 0.7, respectively, whereas values of Type I and Type II error rates (i.e., α and β) were each set equal to 0.1. According to the SPRT procedure, after k items have been administered with s_k of them being answered correctly, mastery was now declared if the likelihood ratio $L(x_1,\ldots,x_k \mid p_1) / L(x_1,\ldots,x_k \mid p_0) = [(0.7)^{s_k} (0.3)^{k-s_k} / (0.5)^{s_k} (0.5)^{k-s_k}]$ was smaller than $\alpha/(1-\beta)$, nonmastery if this likelihood ratio was larger than $(1-\alpha)/\beta$, and otherwise sampling was continued. For those students who could not be classified as either a master or nonmaster before the item pool was exhausted, a classification decision was made in the same way as in the CT/PC procedure, using a mastery proportion-correct value of 0.6.

The fourth comparison will be made with an AMT strategy using a maximum information item selection strategy with a symmetric Bayesian confidence interval of 90% and using Owen's Bayesian scoring algorithm for a point estimation of student's latent ability on an IRT-metric. Like in the CT/B procedure, a standard normal prior $N(0,1)$ was assumed for the Bayesian scoring of the adaptive test. Also, like in the CT/B procedure, the prespecified cut-off points on the latent IRT-metric (i.e., the mastery levels) in each of the 100-item pools corresponding to 60% correct were determined from the TRF.

In order to make a fair comparison of the minimax sequential strategy with the four strategies described above, the criterion level t_c was set equal to 0.6. Furthermore, the losses l_{01} and l_{10} associated with the incorrect classification decisions were assumed to be equal corresponding to the assumption of equal error rates in Wald's SPRT procedure. On a scale in which one unit corresponded to the cost of administering one item (i.e., $e = 1$), l_{01} and l_{10} were each set equal to 200 reflecting the fact that costs for administering another random item were assumed to be rather small relative to the costs associated with incorrect classification decisions. Finally, the parameter α of the beta distribution $B(\alpha,1)$ as least favorable prior was set equal to 10^{-9}.

Using the backward induction computational scheme discussed earlier, for given maximum test length n, a computer program called MINIMAX was developed to determine the appropriate action (i.e., nonmastery, mastery, continuing sampling) for the minimax sequential strategy at each stage of sampling k for different number-correct score s_k. The recurrent relation:
$\binom{k+1}{y+1} = \binom{k}{y} + \binom{k}{y+1}$, in combination with $\binom{n}{n} = \binom{n}{0} = 1$, was hereby used for computing the binomial coefficients in (4). A copy of the program MINIMAX is available from the author upon request.

12.2 Item pools

In the simulation study by Kingsbury and Weiss (1983), the simulations were conducted using four 100-item pools generated to reflect different types of item pools.

Pool 1 (uniform pool) consisted of items that were perfect replications of each other. More specifically, each item had discrimination *a* of 1, difficulty *b* of 0, and lower asymptote *c* (pseudo-guessing level) of 0.2. This item pool reflected the SPRT procedure's assumption that all items have equal difficulty. As noted before, this assumption also reflects the choice of the binomial distribution for modeling response behavior in the minimax sequential procedure.

Pool 2 (*b*-variable pool) varied from the uniform pool only in that the difficulties *b* differed across a range of values and reflected the 1-parameter IRT model (i.e., Rasch model).

Pool 3 (*a*- and *b*-variable pool) varied from the *b*-variable pool only in that the discriminations *a* differed across a range of values and was designed to simulate the 2-parameter IRT model.

Pool 4 (*a*-, *b*-, and *c*-variable pool) varied from the *a*- and *b*-variable pool only in that the lower asymptotes *c* were allowed to spread across a range of values and simulated the 3-parameter IRT model.

For a more detailed description of the four different item pools, refer to Kingsbury and Weiss (1983).

12.3 Maximum test lengths

Conventional tests (CTs) of three different lengths (10, 25, and 50 items) were randomly drawn from each of the four item pools. Doing so, the 10-item test served as the first portion of the 25-item test and the 25-item test in turn served as the first portion of the 50-item test. These 12 CTs served as subpools from which the SPRT, AMT, and minimax sequential procedures drew items during the simulations.

It is important to notice that this random sampling from a larger domain of items implies that the binomial model assumed in both Wald's SPRT and the minimax sequential procedure holds. Thus, not only for the uniform pool but also for the *b*-variable, *a*- and *b*-variable, and *a*-, *b*-, and *c*-variable pool, the assumed binomial model holds in these two testing strategies.

12.4 Item response generation

Item responses for 500 simulated students, drawn from a $N(0,1)$ distribution, were generated for each item in each of the four item pools. For known ability of the

simulated student and given item parameters, first the probability of a correct answer was calculated using the 3-PL model. Next, this probability was compared with a random number drawn from a uniform distribution in the range from 0 to 1. The item administered to the simulated student was scored correct and incorrect if this randomly selected number was less and greater than the probability of a correct answer, respectively.

Furthermore, a simulated student was supposed to be a "true" master if his/her ability used to generate the item responses was higher than a prespecified cut-off point on the $N(0,1)$ ability metric. Since a value of 0.6 on the proportion-correct metric of each of the four item pools corresponded after conversion with a value of 0 on the $N(0,1)$ ability metric, the cut-off point on the $N(0,1)$ ability metric was set equal to 0.

13 Results of the Monte Carlo simulation

In this section, the results of the Monte Carlo simulations will be compared for the different variable-length mastery testing strategies in terms of average test length, correspondence with true mastery status (i.e., classification accuracy), and correspondence as a function of average test length (i.e., efficiency of testing strategy).

13.1 Average test lengths

Table 2 shows the average number of items required by each of the variable-length mastery testing strategies before a mastery/nonmastery decision can be made. The minimax sequential testing strategy is hereby denoted as MINI.

As can be seen from Table 2, the MINI strategy resulted in considerably average test length reductions for each combination of item pool and maximum test length (MTL). Table 2 also shows that, except for the *a*-, *b*-, and *c*-variable pool by the SPRT strategy at the 50-item MTL level, the MINI procedure resulted in a greater reduction of average test lengths than the conventional, AMT, and SPRT strategies for each item pool at all MTL levels. Finally, like under the other strategies, it can be inferred from Table 2 that for each item pool the reduction in average test length increased under the MINI strategy with increasing MTL. For the uniform pool, the average test length was reduced by 36%, 54%, and 71% for the 10-item MTL, 25-item MTL, and 50-item MTL, respectively. For the *b*-variable pool, *a*- and *b*-variable pool, and *a*-, *b*-, and *c*-variable pool, these percentages in average test length reduction were (25%; 44%; 61%), (41%; 57%; 68%), and (28%; 50%; 65%), respectively. Hence, under the MINI strategy, the greatest reductions in average test length were achieved by the *a*- and *b*-variable pool and uniform pool.

Table 2: Mean number of items administered to each simulee for four mastery testing strategies using each item pool, at three maximum test lengths.

Item pool and testing strategy	Maximum test length 10	25	50
Uniform pool			
Conventional	10.00	25.00	50.00
AMT	9.03	15.99	23.00
SPRT	8.75	13.12	15.39
MINI	6.41	11.47	14.49
b-variable pool			
Conventional	10.00	25.00	50.00
AMT	9.43	18.09	27.17
SPRT	9.62	16.79	21.41
MINI	7.55	14.08	19.48
a- and *b*-variable pool			
Conventional	10.00	25.00	50.00
AMT	8.55	15.78	24.07
SPRT	9.41	15.78	18.55
MINI	5.86	10.86	15.96
a-, *b*-, and *c*-variable pool			
Conventional	10.00	25.00	50.00
AMT	8.73	16.35	23.39
SPRT	8.62	13.42	15.70
MINI	7.18	12.61	17.27

13.2 Classification accuracy

Kingsbury and Weiss (1983) and Weiss and Kingsbury (1984) used phi correlations between true classification status (i.e., true master or true nonmaster) and estimated classification status (i.e., declaring mastery or nonmastery) as indicators of the quality/validity of the classification decisions. Therefore, these authors denoted the phi correlations as measures of classification accuracy. However, the phi coefficient is not appropriate for the assessment of classification accuracy. The reason is that phi is sensitive to unequal proportions of true and declared masters (see, Lord & Novick, 1968, sect. 15.9). Van der Linden and Mellenbergh (1978) proposed coefficient delta for the assessment of classification accuracy, which is not sensitive to unequal proportions of true and declared masters. They showed that delta reduces to the well-known Loevinger's coefficient

H if the threshold loss function is: $l_{00} = l_{11} = 0$, $l_{01} = l_{10}$. Since the losses for the correct classification decisions were assumed to be equal to zero and the losses for the incorrect classification decisions were both set equal to 200, it follows thus that coefficient *H* applies to our simulation study. Coefficient *H* is defined as phi/phi(max), where phi(max) is the maximum of phi given the marginal distributions of the 2 x 2 table. Although coefficient delta is not always in the interval from 0 to 1, however, it has been shown (van der Linden & Mellenbergh, 1978) that coefficient *H* is in this interval. A value of 0 signifies that the test is worthless, and a value of 1 signifies that the test is perfect for the decision situation.

Table 3: Loevinger's coefficients *H* between observed mastery status and true mastery status for each mastery testing strategy, using each type of item pool, at three maximum test lengths.

Item pool and testing strategy	Maximum test length		
	10	25	50
Uniform pool			
CT/PC	0.862	0.901	0.923
CT/B	0.813	0.887	0.919
AMT	0.873	0.912	0.920
SPRT	0.856	0.908	0.916
MINI	0.612	0.824	0.721
b-variable pool			
CT/PC	0.614	0.722	0.837
CT/B	0.609	0.691	0.848
AMT	0.649	0.749	0.879
SPRT	0.607	0.698	0.788
MINI	0.578	0.709	0.689
a- and *b*-variable pool			
CT/PC	0.691	0.801	0.832
CT/B	0.696	0.799	0.834
AMT	0.691	0.803	0.829
SPRT	0.683	0.789	0.801
MINI	0.671	0.765	0.792
a-, *b*-, and *c*-variable pool			
CT/PC	0.389	0.781	0.841
CT/B	0.413	0.817	0.889
AMT	0.408	0.809	0.878
SPRT	0.378	0.689	0.678
MINI	0.748	0.892	0.948

Table 3 shows Loevinger's coefficients *H* for the present simulation study. As can be seen from Table 3, the MINI strategy resulted only for the *a*-, *b*-, and *c*-variable pool in higher coefficients *H* than the other four testing strategies at all MTL levels. In particular, for the 10-item MTL the coefficients *H* were considerably higher. For both the *b*-variable and *a*- and *b*-variable pool, the other four testing strategies generally yielded somewhat higher coefficients *H*. For the uniform pool, however, the other four testing strategies yielded considerably higher coefficients *H*.

Furthermore, Table 3 shows that the coefficients *H* for both the 25-item and 50-item MTL were higher than for the 10-item MTL by each pool type under the MINI strategy. For both the *a*- and *b*-variable pool and *a*-, *b*-, and *c*-variable pool, under the MINI strategy, the 50-item MTL yielded higher coefficients *H* than the 25-item MTL, whereas the opposite did hold for both the uniform and *b*-variable pool.

13.3 Most efficient testing strategy

Kingsbury and Weiss (1983) depicted graphically the phi correlation as a function of the average number of items administered by each testing strategy for each item pool (see also Weiss and Kingsbury, 1984). In other words, they matched the average test length on the classification accuracy. From these graphs conclusions were derived concerning which testing strategy was most efficient. A testing strategy was hereby said to be most efficient if it results in the combination of highest phi correlation and shortest average test length. Following Kingsbury and Weiss (1983) and Weiss and Kingsbury (1984), a testing strategy will be called most efficient in this paper if it results in the combination of highest Loevinger's coefficient *H* and shortest average test length.

As is immediately clear from Tables 2 and 3, the MINI strategy was the most efficient of all testing procedures for the (realistic) *a*-, *b*-, and *c*-variable pool, since it generally yielded both the highest coefficients *H* and shortest average test lengths at each MTL level. Although the SPRT strategy required at the 50-item MTL level, on the average, somewhat fewer items for reaching a mastery/nonmastery decision than the MINI strategy (i.e., 15.70 versus 17.27), however, the coefficient *H* for the SPRT strategy was much lower compared to the MINI strategy (i.e., 0.678 versus 0.948). For an average test length of 15.70 (interpolating from the data in Tables 2 and 3), the MINI strategy would result in a coefficient *H* of 0.804.

For the *a*- and *b*-variable pool, as can been from Tables 2 and 3, the MINI strategy yielded shorter mean test lengths than all other strategies, whereas the coefficients *H* were generally somewhat lower at each MTL level. The MINI strategy resulted in a coefficient *H* of 0.792 at a mean test length of 15.96 (the longest mean test length observed at the 50-item MTL level). Interpolating data

from Tables 2 and 3, it can easily be verified that the SPRT procedure would need to administer 16.47 items to achieve this same coefficient H of 0.792, the AMT procedure would need 15.07 items, the CT/B procedure would need 24.63 items, and the CT/PC procedure would need 21.48 items. Hence, for the a- and b-variable pool, the MINI procedure was considerably more efficient than both the CT/PC and CT/B strategies, whereas the MINI procedure was somewhat more efficient than the SPRT procedure. Compared to the AMT procedure, however, the MINI procedure was somewhat less efficient.

For the b-variable pool, Tables 2 and 3 show that at the longest mean test length observed for the MINI procedure (i.e., 19.48 at the 50-item MTL level), this strategy resulted in a coefficient H of 0.689. Interpolating data from Tables 2 and 3, it follows that the SPRT procedure would need to administer 16.08 items to achieve this same coefficient H of 0.689, the AMT procedure would need 12.89 items, the CT/B procedure would need 23.95 items, and the CT/PC procedure would need 20.42 items. Hence, for the b-variable pool, it can be concluded that the MINI procedure was considerably more efficient than the CT/B procedure and somewhat more efficient than the CT/PC procedure. On the other hand, however, the MINI procedure was somewhat less efficient than the SPRT procedure and considerably less efficient than the AMT procedure.

Finally, it can be inferred from Tables 2 and 3 that the MINI strategy resulted for the uniform pool in a coefficient H of 0.721 at the longest mean test length observed (i.e., 14.49 at the 50-item MTL level). It follows immediately from Tables 2 and 3 that each of the four other testing strategies would need to administer less than 10 items to achieve this same coefficient H of 0.721. Hence, for the (unrealistic) uniform pool, it can be concluded that the MINI procedure is considerably less efficient than the four other testing strategies.

14 Discussion

Optimal rules for the sequential mastery problem (nonmastery, mastery, and to continue sampling) were derived using the framework of minimax sequential decision theory. The binomial distribution was assumed for modeling response behavior, whereas threshold loss was adopted for the loss function involved. The least favorable prior, used in the present paper for computing the posterior predictive distributions, turned out to be the beta distribution with parameter β equal to 1 and parameter α sufficiently small.

In a Monte Carlo simulation, the minimax sequential procedure (MINI) was compared with other procedures that exist for both sequential and adaptive mastery testing in the literature. Maximum test length (MTL) varied from 10 to 50 items, and different types of item pools were considered by changing the values of the item parameters.

The results of the simulation study indicated that, compared to the other testing strategies examined in the literature, the MINI strategy was most efficient (i.e., combination of highest Loevinger's coefficient *H* between true and estimated mastery status and shortest average test length) for item pools reflecting the (realistic) 3 PL-model at each MTL level. Also, except for the AMT strategy, the MINI strategy turned out to be most efficient for item pools reflecting the 2 PL-model at each MTL level. For item pools reflecting the 1 PL-model (i.e., the Rasch model), the MINI strategy appeared to be more efficient than the two conventional fixed-length methods (i.e., employing proportion correct and a Bayesian scoring method for making mastery/nonmastery decisions) but less efficient than both the AMT and SPRT procedure at each MTL level. For the (unrealistic) uniform item pools, however, it turned out that the MINI strategy was less efficient than the other testing strategies at each MTL level.

It is important to notice, however, that the MINI strategy is especially appropriate when costs of testing can be assumed to be quite large. For instance, when testlets rather than single items are considered. Also, the MINI strategy might be appropriate in psychodiagnostics. Suppose that a new treatment (e.g., cognitive-analytic therapy) must be tested on patients suffering from some mental health problem (e.g., anorexia nervosa). Each time after having exposed a patient to the new treatment, it is desired to make a decision concerning the effectiveness/ineffectiveness of the new treatment or testing another patient. In such clinical situations, costs of testing generally are quite large and the MINI approach might be considered as an alternative to other testing strategies, such as SPRT, AMT, or conventional fixed-length tests.

An issue that still deserves some attention is why in the present paper, somewhat counter to the current trend in applied measurement, a random rather than IRT-based adaptive item selection procedure is preferred. As noted before, IRT-based item selection strategies assume that a calibrated pool of items exists which differ in their particular characteristics (i.e., levels of difficulty and discrimination). For random item selection strategies, such as Wald's SPRT procedure and the minimax sequential procedure advocated in this paper, however, the existence of a pool of parallel items only is required. Such pools of parallel items often are easier to construct than pools of items, which do differ in their IRT characteristics.

In case a calibrated pool of items does exist, however, an IRT-based adaptive strategy that selects items for administration based on their particular characteristics is preferred rather than to randomly select items from a pool. A promising approach, in which the strong point of the minimax and Bayesian sequential procedures, that is, taking cost per observation explicitly into account, is combined with an IRT-based adaptive item selection strategy might be the following. The item to be administered next is the one that maximizes information or minimizes posterior variance at student's last ability estimate on an IRT-metric. At each stage of sampling, the action declaring mastery, declaring nonmastery, or

to continue sampling is then chosen which minimizes the posterior or maximum expected losses associated with all possible decision rules (see also Vos & Glas, 2000).

A final note is appropriate. Following the same line of reasoning as in the present paper, optimal rules derived here can easily be generalized to the situation where three or more mutually exclusive classification categories can be distinguished. In Weiss and Kingsbury (1984), it is indicated how the AMT procedure can be employed in the context of allocating students to more than two grade classes (i.e., adaptive grading test). Spray (1993) has shown how a generalization of Wald's SPRT procedure (i.e., Armitage's (1950) combination procedure) can be applied to multiple categories, whereas Bayesian sequential decision theory is applied in Vos (1999) to SMT in case the three classification decisions declaring nonmastery, partial mastery, and mastery are open to the decision-maker (see also Smith & Lewis, 1995).

References

[1] Angoff, W.H. (1971): Scales, norms and equivalent scores. In R.L. Thorndike (Ed.): *Educational Measurement* (2nd ed.), 508-600. Washington, D.C.: American Council on Education.

[2] Armitage, P. (1950): Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society*, **12**, 137-144.

[3] Chang, H. and Stout, W.F. (1993): The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, **58**, 37-52.

[4] Coombs, C.H., Dawes, R.M., and Tversky, A. (1970): *Mathematical Psychology: An Elementary Introduction*. Englewood Cliffs, New Jersey: Prentice-Hall Inc.

[5] DeGroot, M.H. (1970): *Optimal Statistical Decisions*. New York: McGraw-Hill.

[6] De Gruijter, D.N.M. and Hambleton, R.K. (1984): On problems encountered using decision theory to set cutoff scores. *Applied Psychological Measurement*, **8**, 1-8.

[7] Ferguson, R.L. (1969): *The Development, Implementation, and Evaluation of a Computer-Assisted Branched Test for a Program of Individually Prescribed Instruction*. Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh PA.

[8] Ferguson, T.S. (1967): *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.

[9] Hambleton, R.K. and Novick, M.R. (1973): Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, **10**, 159-170.

[10] Huynh, H. (1980): A nonrandomized minimax solution for passing scores in the binomial error model. *Psychometrika*, **45**, 167-182.

[11] Keats, J.A. and Lord, F.M. (1962): A theoretical distribution of mental test scores. *Psychometrika*, **27**, 59-72.

[12] Kingsbury, G.G. and Weiss, D.J. (1983): A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D.J. Weiss (Ed.): *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, 257-283: New York: Academic Press.

[13] Lehmann, E.L. (1959): *Testing Statistical Hypotheses* (3rd ed.). New York: Macmillan.

[14] Lewis, C. and Sheehan, K. (1990): Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, **14**, 367-386.

[15] Lord, F.M. and Novick, M.R. (1968): *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley.

[16] Luce, R.D. and Raiffa, H. (1957): *Games and Decisions*. New York: John Wiley and Sons.

[17] Nedelsky, L. (1954): Absolute grading standards for objective tests. *Educational and Psychological Measurement*, **14**, 3-19.

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

[19] Reckase, M.D. (1983): A procedure for decision making using tailored testing. In D.J. Weiss (Ed.): *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*, 237-257: New York: Academic Press.

[20] Sheehan, K. and Lewis, C. (1992): Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement*, **16**, 65-76.

[21] Smith, R.L. and Lewis, C. (1995, April): *A Bayesian Computerized Mastery Model with Multiple Cut Scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

[22] Spray, J.A. (1993): *Multiple-Category Classification Using a Sequential Probability Ratio Test* (Research Rep. No. 93-7). Iowa City, IA: American College Testing.

[23] Spray, J.A. and 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.

[24] van der Linden, W.J. (1981): Decision models for use with criterion-referenced tests. *Applied Psychological Measurement*, **4**, 469-492.

[25] van der Linden, W.J. (1990): Applications of decision theory to test-based decision making. In R.K.

[26] Hambleton and J.N. Zaal (Eds.), *New Developments in Testing: Theory and Applications*, 129-155. Boston: Kluwer.

[27] van der Linden, W.J. and Mellenbergh, G.J. (1977): Optimal cutting scores using a linear loss function. *Applied Psychological Measurement*, **1**, 593-599.

[28] van der Linden, W.J. and Mellenbergh, G.J. (1978): Coefficients for tests from a decision theoretic point of view. *Applied Psychological Measurement*, **2**, 119-134.

[29] van der Linden, W.J. and Vos, H.J. (1996): A compensatory approach to optimal selection with mastery scores. *Psychometrika*, **61**, 155-172.

[30] Veldhuijzen, N.H. (1982): Setting cutting scores: A minimum information approach. In W.J. van der Linden (Ed.): *Aspects of Criterion-Referenced Measurement. Evaluation in Education: An International Review Series*, **5**, 141-148.

[31] Vos, H.J. (1997a): Simultaneous optimization of quota-restricted selection decisions with mastery scores. *British Journal of Mathematical and Statistical Psychology*, **50**, 105-125.

[32] Vos, H.J. (1997b): A simultaneous approach to optimizing treatment assignments with mastery scores. *Multivariate Behavioral Research*, **32**, 403-433.

[33] Vos, H.J. (1999): Applications of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Statistics*, **24**, 271-292.

[34] Vos, H.J. and Glas, C.A.W. (2000): Testlet-based adaptive mastery testing. In W.J. van der Linden and C.A.W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice*, 289-309.

[35] Wald, A. (1947): *Sequential Analysis*. New York: Wiley.

[36] Weiss, D.J. and Kingsbury, G.G. (1984): Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, **21**, 361-375.

[37] Wilcox, R.R. (1981): A review of the beta-binomial model and its extensions. *Journal of Educational Statistics*, **6**, 3-32.