Psicológica (2000) 21, 191-211.

A Bayesian Procedure in the Context of Sequential Mastery Testing

Hans J. Vos*
University of Twente, The Netherlands

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 continuing testing and administering another random item. The framework of Bayesian sequential decision theory is used; that is, optimal rules are obtained by minimizing the posterior expected losses associated with all possible decision rules at each stage of testing. The main advantage of this approach is that costs of testing can be taken explicitly 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. The paper concludes with a simulation study, in which the Bayesian sequential strategy is compared with other procedures that exist for similar classification decision problems in the literature.

Key words: sequential mastery testing, Bayesian sequential rules, binomial distribution, threshold loss, most efficient strategy.

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 this approach, 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

---

* Faculty of Educational Science and Technology. Department of Educational Measurement and Data Analysis. P.O. Box 217, 7500 AE Enschede, The Netherlands, Phone: +31 53 489 3628, Fax: +31 53 489 4239, E-mail: vos@edte.utwente.nl .
Acknowledgements: 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.

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. This prior probability represents our best prior beliefs concerning student's true level of functioning; that is, before any item yet has been administered.

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 testing and administering another random 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 Bayesian sequential decision theory (e.g., DeGroot, 1970; Lehmann, 1959). The main advantage of this approach is that costs of testing (i.e., administering another random item) can be taken explicitly into account.

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.

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 continue testing. 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.

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 testing 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 testing 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 procedure (Kingsbury and Weiss, 1983).

Recently, Lewis and Sheehan (1990), Sheehan and Lewis (1992), and Smith and Lewis (1995) have applied Bayesian sequential decision theory to SMT. In addition to a psychometric model and a loss function, cost of testing (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 testing. As far as the posterior expected loss associated with continue testing 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 testing using techniques of dynamic programming (i.e., backward induction). This technique starts by considering the final stage of testing and then works backward to the first stage of testing. 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 testing is optimal with respect to the entire sequential mastery testing procedure.

Lewis and Sheehan (1990) and Sheehan and Lewis (1992) modeled response behavior in the form of a three-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 continue testing 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.

Bayesian Sequential Principle Applied to SMT

In this section, it is indicated how the framework of Bayesian sequential decision theory in combination with the binomial distribution for modeling response behavior is applied to SMT. Also, a rationale is provided for why this approach should be applied to sequential mastery testing in comparison to other approaches that exist for the variable-length mastery problem (both of a sequential and adaptive character) in the literature.

Bayesian Sequential Decision Theory in Combination with the Binomial Model

In the present paper, as in Lewis and Sheehan's model (1990), the framework of Bayesian sequential decision theory will also be applied to SMT. As in Ferguson's (1969) approach, however, the binomial distribution instead of an IRT-model will be considered here for modeling response behavior. It will be shown later on 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.

Rationale for Applying the Bayesian Sequential Principle

As pointed out by Lewis and Sheehan (1990), an IRT-based adaptive item selection rule requires a pool of content-balanced test items such that its difficulty levels span the full range of ability levels in the population. These specialized pools are often difficult to construct. Random item selection, however, requires a pool of parallel items, that is, items from the same difficulty levels. Procedures for constructing such pools of parallel items are often available. In addition to the reasons of computational efficiency (i.e., no estimation of student's last ability required) and simplicity, therefore, Lewis and Sheehan (1990) decided to consider a random rather than adaptive item selection procedure.

Following the same line of reasoning as in the Lewis and Sheehan (1990) model, in the present paper also random rather than adaptive item selection is used. To comply with the requirement of administering the next item randomly from a pool of items from the same difficulty levels, following Ferguson (1969), the probability of a correct response for given true level of functioning will be modeled here by a binomial distribution.

For reasons given above, applying an IRT-based adaptive item selection procedure to the variable-length mastery problem is not considered in this paper. However, one might wonder why the Bayesian sequential principle should be preferred above the application of Wald's SPRT-framework. The main advantage of the Bayesian sequential strategy as compared to Wald's SPRT-framework is that cost per observation can explicitly been taken into account. In some real-life applications of variable-length mastery testing, costs associated with administering additional items might be quite large.

Notation

In the following, as in Ferguson's approach (1969), a sequential mastery test is supposed to have a maximum length of n ($n \geq 1$). Let the observed item response at each stage of testing k ($1 \leq k \leq n$) for a randomly sampled student be denoted by x_k, which can take the values 0 and 1 for respectively incorrect and correct responses to the k-th item. Furthermore, let $s_k = x_1 + \ldots + x_k$ ($0 \leq s_k \leq k$) denote the observed number of correct responses after k items have been administered. 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 (unknown) true level of functioning be denoted by t ($0 \leq t \leq 1$). Finally, a criterion level t_c ($0 \leq t_c \leq 1$) must be specified in advance by the decision-maker using methods of standard-setting (e.g., Angoff, 1971; Nedelsky, 1954). 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.

Assuming an observed response pattern $(x_1,\ldots,x_k)$, the two basic elements of Bayesian sequential decision making discussed earlier can now be formulated as follows: A psychometric model in the form of a probability distribution, $\text{Prob}(s_k \mid t)$, relating observed number-correct score s_k to student's true level of functioning t at each stage of testing k, and a loss function describing the loss L(m,t) or L(n,t) incurred when mastery or nonmastery is declared for given t, respectively.

Threshold Loss and Costs of Testing

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 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 Lewis and Sheehan (1990), 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 testing k as follows (see also Lewis & Sheehan, 1990):

Table 1. Table for threshold loss function at stage k (1 £ k £ n) of testing

True level of functioning Decision	$t \leq t_c$	$t > t_c$
Declaring nonmastery	ke	l_{01} + ke
Declaring mastery	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 testing 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 continue testing option. This is because the posterior expected loss associated with the continue testing option is computed at each stage of testing as a weighted average of the posterior 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} ($i = 0,1$; $i \neq j$) 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.

An obvious disadvantage of the threshold loss function is that, as can be seen from Table 1, it assumes constant loss for students to the left or to the right of t_c, no matter how large their distance from t_c. In practice, however, errors in classification are sometimes considered to be more serious, the further a student is from the criterion level t_c. For instance, a student who is declared a nonmaster with true level of functioning just above t_c gives the same loss as a misclassified true nonmaster with true level of functioning far above t_c. It seems more realistic to suppose that for misclassified true nonmasters the loss is a strictly increasing function of t. Moreover, the threshold loss function shows a "threshold" at the point t_c, and this discontinuity also seems unrealistic in many cases. In the neighborhood of this point, the losses for correct and incorrect decisions should change smoothly rather than abruptly (van der Linden, 1981).

To overcome these shortcomings, van der Linden and Mellenbergh (1977) proposed a continuous loss function for the fixed-length mastery problem which is a linear function of student's true level of functioning (see also van der Linden & Vos, 1996; Vos, 1997a, 1997b, 1999). Although a linear loss function is probably more appropriate for the sequential mastery problem, following Lewis and Sheehan (1990), in the present paper a threshold loss function is adopted for reasons of simplicity and computational efficiency. Another reason for using threshold rather than linear loss is that a linear loss function may be more appropriate in the neighborhood of t_c indeed but that the further away from t_c, however, the losses can be assumed to take more and more the same constant values again.

Psychometric Model

Following Wald (1947) and Ferguson (1969), as noted before, in the present paper the well-known binomial model will be adopted for the probability that after k items have been administered, s_k of them have been answered correctly. Its distribution at stage k of testing for given student's true level of functioning t, $\text{Prob}(s_k \mid t)$, can be written as follows:

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

If each response is independent of the other, and if student's probability of a correct answer remains constant, the distribution function of s_k, given 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.

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 Bayesian sequential decision theory. Doing so, given an observed item response vector $(x_1,\ldots,x_k)$, first the Bayesian principle will be applied to the fixed-length mastery problem by determining which of the posterior expected losses associated with the two classification decisions is the smallest. Next, applying the Bayesian principle again, optimal rules for the sequential mastery problem are derived at each stage of testing k by comparing this quantity with the posterior expected loss associated with the continue testing option.

Applying the Bayesian Principle to the Fixed-Length Mastery Problem

As noted before, the Bayesian decision rule for the fixed-length mastery problem can be found by minimizing the posterior expected losses associated with the two classification decisions of declaring mastery or nonmastery. In doing so, the posterior expected loss is taken with respect to the posterior distribution of t. Let $E[L(m, t) \mid s_k]$ and $E[L(n, t) \mid s_k]$ denote the posterior expected losses associated with these two classification decisions, respectively, given number-correct score s_k. It then follows that mastery is declared when number-correct score s_k is such that

$$E[L(m, t) \mid s_k] \; < \; E[L(n, t) \mid s_k], \qquad (2)$$

and nonmastery is declared otherwise. It now can easily be verified from (1)-(2), and using Table 1, that mastery is declared when s_k is such that

$$(l_{10}+ke)\text{Prob}(t \leq t_c s_k)+(ke)\text{Prob}(t>t_c \mid s_k)<(ke)\text{Prob}(t \leq t_c \mid s_k)+(l_{01}+ke)\text{Prob}(t>t_c \mid s_k) \qquad (3)$$

and that nonmastery is declared otherwise. Rearranging terms, it can easily be verified from (3) that mastery is declared when s_k is such that

$$\text{Prob}(t \leq t_c \mid s_k) \; < \; 1/(1+R), \qquad (4)$$

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

Assuming a beta prior for t, 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 $B_t(a, b)$ with parameters α and β (α, β > 0) is chosen as 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 $B_t(a + s_k, k - s_k + b)$. Hence, assuming a beta prior for t, it follows from (4) that mastery is declared when s_k is such that

$$B_{t_c}(a + s_k, k - s_k + b) < 1/(1+R) \tag{5}$$

and that nonmastery is declared otherwise.

The beta prior might be specified as either an empirical (i.e., empirical Bayes approach) or subjective prior (i.e., subjective Bayes approach). In the first approach, empirical data from other students of the group to which the individual student belongs (i.e., 'comparable group') are used for estimating the parameters α and β. In the second approach, prior knowledge about t is specified by subjective assessment. A subjective beta prior will be assumed in this paper. More specifically, the uniform distribution on the standard interval [0,1] is taken as a noninformative prior; that is, the beta distribution $B_t(a, b)$ with $\alpha = \beta = 1$. In other words, prior true level of functioning can take on all values between 0 and 1 with equal probability. This particular prior is used for illustrative purposes in the present paper.

It then follows immediately from (5) that mastery is declared when s_k is such that

$$B_{t_c}(1 + s_k, k - s_k + 1) < 1/(1+R) \tag{6}$$

and that nonmastery is declared otherwise. The beta distribution has been extensively tabulated (e.g., Pearson, 1930). Normal approximations are also available (Johnson & Kotz, 1970, sect. 2.4.6).

Derivation of Bayesian Sequential Rules

Let $d(x_1,\ldots,x_k)$ denote the Bayesian sequential rule at stage k of testing. At each stage of testing k, $d(x_1,...,x_k)$ can then be found by using the following backward induction computational scheme: First, the Bayesian sequential rule at the final stage of testing n is computed. Since the continue testing option is not available at that stage of testing, it follows immediately that the Bayesian sequential rule (i.e., $d(x_1,...,x_n)$) coincides with the Bayesian rule for the fixed-length mastery problem, that is, declare mastery if the inequality in (6) holds for $s_k = s_n$ and $k = n$; otherwise, declare nonmastery.

Subsequently, the Bayesian sequential rule at the next to last stage of testing (n-1) is computed by comparing the minimum of the two classification decisions, that is, min{E[L(m, t) | s_{n-1}], E[L(n, t) | s_{n-1}]}, with the posterior expected loss associated with the continue testing option. As noted before, the posterior expected loss associated with administering one more item at stage (n-1) of testing, given an observed response pattern (x_1,..,x_{n-1}), is computed by averaging the posterior expected losses associated with each of the possible future decision outcomes at the final stage of testing n relative to the probability of observing those outcomes (i.e., backward induction).

Let Prob(x_n | s_{n-1}) denote the probability of observing response x_n (x_n = 0 or 1) on the final stage of testing n, given observed number-correct score s_{n-1} on the (n-1) previous stages of testing, then, the posterior expected loss associated with administering one more item after (n-1) items have been administered, E[L(c, t) | s_{n-1}], is computed as follows:

$$E[L(c,t)\,|\,s_{n-1}] = \sum_{x_n=0}^{x_n=1} \min\{E[L(m,t)\,|\,s_n],\, E[L(n,t)\,|\,s_n]\} * \mathrm{Prob}(x_n\,|\,s_{n-1}) \qquad (7)$$

Note that (7) averages the posterior expected losses associated with each of the possible future decision outcomes relative to the probability of observing those outcomes. Generally, Prob(x_k|s_{k-1}) is called the posterior predictive probability of observing response x_k (x_k = 0 or 1) at stage k of testing, conditional on having obtained an observed number of correct responses s_{k-1} on the (k-1) previous stages of testing. It will be indicated later on how this conditional probability can be computed.

Given an observed response pattern (x_1,…,x_{n-1}), the Bayesian sequential rule at stage (n-1) of testing (i.e., d(x_1,...,x_{n-1})) is now given by (8):

$$d(x_1,...,x_{n-1}) = \begin{cases} \text{declare mastery} & \text{if } E[L(m,t)\,|\,s_{n-1}] \text{ is a minimum} \\ \text{declare nonmastery} & \text{if } E[L(n,t)\,|\,s_{n-1}] \text{ is a minimum} \\ \text{continue testing} & \text{if } E[L(c,t)\,|\,s_{n-1}] \text{ is a minimum.} \end{cases}$$

To compute the posterior expected loss associated with the continue testing option at stage (n-2), the so-called risk at stage (n-1) of testing is needed. The risk at stage (n-1) of testing, Risk($x_1,\ldots,x_{n-1}$), is defined as the minimum of the posterior expected losses associated with all available decisions, that is, declare mastery, declare nonmastery, or continue testing. In other words:

$$\text{Risk}(x_1,...,x_{n-1}) = \min\{E[L(m, t) \mid s_{n-1}], E[L(n, t) \mid s_{n-1}], E[L(c, t) \mid s_{n-1}]\} \quad (9)$$

The posterior expected loss associated with administering one more item after (n-2) items have been administered with s_{n-2} of them being answered correctly, $E[L(c, t) \mid s_{n-2}]$, can then be computed by using the following recurrent relation (i.e., computing the expected risk):

$$E[L(c, t) \mid s_{n-2}] = \sum_{x_{n-1}=0}^{x_{n-1}=1} \text{Risk}(x_1,...,x_{n-1})*\text{Prob}(x_{n-1} \mid s_{n-2}) \quad (10)$$

Given an observed response pattern ($x_1,\ldots,x_{n-2}$), the Bayesian sequential rule at stage (n-2) of testing (i.e., $d(x_1,...,x_{n-2})$) can now be computed analogous to the computation of $d(x_1,\ldots,x_{n-1})$ under (8). Following the same computational backward scheme as in determining the Bayesian sequential rules at stages (n-1) and (n-2), the Bayesian sequential rules at stages (n-3),...,1 are computed.

Computation of Posterior Predictive Probabilities

For computing the posterior expected loss associated with administering one more item after (k-1) items have been administered with s_{k-1} of them being answered correctly (i.e., $E[L(c, t) \mid s_{k-1}]$), as can be seen from (7) and (10), the posterior predictive probability Prob($x_k \mid s_{k-1}$) is needed. To compute this probability, a prior probability distribution must be specified. As noticed before, here the uniform prior as a special case of the beta prior $B_t(a, b)$ with $\alpha = \beta = 1$, is taken as a prior probability distribution.

If s_{k-1} items have been answered correctly after (k-1) items have been administered, and assuming a uniform prior, in combination with the binomial distribution for the psychometric model, it is known (e.g., DeGroot, 1970) that the probability on a correct response to the k-th item (i.e., Prob($x_k = 1 \mid s_{k-1}$)) is equal to $(1+s_{k-1})/(k+1)$. Since the probabilities on a correct and incorrect response must sum to 1, it follows immediately that the probability on an incorrect response to the k-th item (i.e., $\text{Prob}(x_k = 0 \mid s_{k-1})$) is equal to $[1-(1+s_{k-1})/(k+1)] = (k-s_{k-1})/(k+1)$.

Simulation of Different Mastery Testing Strategies

In a Monte Carlo simulation the Bayesian sequential strategy will be compared with other existing approaches to mastery testing.

Description of the Testing Strategies Used for Comparison

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

The second comparison will be made with Wald's SPRT procedure. The limits of the indifference region in which testing 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 larger than $\alpha/(1-\beta)$, nonmastery if this likelihood ratio was smaller than $(1-\alpha)/\beta$, and otherwise testing 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 procedure, using a mastery proportion-correct value of 0.6.

In order to make a fair comparison of the Bayesian sequential strategy with the two 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 100 reflecting the fact that costs for administering another random item were assumed to be rather small relative to the costs associated with incorrect decisions.

Using the backward induction computational scheme discussed earlier, for given maximum test length n, a computer program called BAYES was developed to determine the appropriate action (i.e., nonmastery,

mastery, or continue testing) for the Bayesian sequential strategy at each stage of testing k for different number-correct score s_k. A copy of the program BAYES is available from the author upon request.

Type of Test and Maximum Test Lengths

The simulation study was conducted using a set of items that were perfect replications of each other (i.e. , uniform pool). More specifically, it was assumed that each item could be described by a one-parameter logistic model (Rasch, 1960) with discrimination parameter a of 1 and equal difficulty parameter b of 0. This item pool reflected the choice of the binomial distribution for modeling response behavior in the Bayesian sequential procedure. Furthermore, the simulation runs were executed using a 100-item pool. Conventional fixed-length mastery tests (CTs) of three different lengths of 10, 25, and 50 items were randomly drawn from this 100-item pool. 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 three CTs served as subpools from which the SPRT and Bayesian sequential procedures drew items during the simulations.

Item Response Generation

Item responses for 1000 simulated students, drawn from a N(0,1) distribution, were generated for each item in the 100-item pool. For known ability of the simulated student and given item difficulty, first the probability of a correct answer was calculated using the one-parameter logistic model. Next, this probability was compared with a random number drawn from the 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 the 100-item pool 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.

Results of the Monte Carlo Simulation

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

Average Test Lengths

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

As can be seen from Table 2, the BAYES strategy resulted in considerably test length reductions at each level of maximum test length (MTL). Table 2 also shows that the BAYES procedure resulted in a greater reduction of average test lengths than the SPRT strategy at all MTL levels. Finally, like under the SPRT strategy, it can be inferred from Table 2 that the reduction in average test length increased under the BAYES strategy with increasing MTL. More specifically, the average test length was reduced by 41.6%, 62.2%, and 77.7% for the 10-item MTL, 25-item MTL, and 50-item MTL, respectively.

Table 2. Average Number of Items to be Administered

	Maximum Test Length		
Strategy	10	25	50
CT	10	25	50
SPRT	8.64	14.47	17.42
BAYES	5.84	9.46	11.15

Classification Accuracy

Table 3 shows phi correlations between true classification status (i.e., true master or true nonmaster) and estimated classification status (i.e., declaring master or nonmaster) for the three testing procedures at each MTL level. These phi correlations (i.e., correspondence coefficients) can be considered as an indicator of the quality/validity of the mastery/nonmastery decisions, and are denoted by Weiss and Kingsbury (1984) as classification validity indicators. As can be seen from Table 3, the phi correlations increased under the BAYES strategy with increasing MTL level.

Table 3. Phi Correlations between True Classification Status and Estimated Classification Status

	Maximum Test Length		
Strategy	10	25	50
CT	0.652	0.787	0.718
SPRT	0.584	0.723	0.718
BAYES	0.610	0.620	0.692

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 (see also Weiss and Kingsbury, 1984). 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.

As is immediately clear from Tables 2 and 3, the BAYES strategy yielded shorter average test lengths than the other two strategies at each MTL level, whereas the phi correlations were generally somewhat lower at each MTL level. To examine which strategy is most efficient, therefore, we compute for the other two strategies the average test length at each MTL level for achieving the same phi correlation as under the BAYES strategy. In other words, we match the average test length on the classification accuracy.

Doing so, the BAYES strategy resulted in a phi correlation of 0.692 for an average test length of 11.15 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 13.17 items on the average to achieve this same phi correlation of 0.692, whereas the CT procedure would need to administer 14.4 items on the average. Similarly, as shown by Tables 2 and 3, the BAYES strategy resulted in a phi correlation of 0.620 for 9.46 items to be administered on the average at the 25-item MTL level. The SPRT and CT procedures would need to administer 10.15 and 9.54 items on the average to achieve this same phi correlation of 0.620, respectively. Finally, as indicated by Tables 2 and 3, the BAYES procedure resulted in a phi correlation of 0.610 for 5.84 items to be administered on the average at the 10-item MTL level. The SPRT and CT procedures would need to administer 9.73 and 8.61 items on the average to achieve this same phi correlation of 0.610, respectively. Hence, for the specific parameter values chosen in this simulation study, it can be concluded that the BAYES procedure was the most efficient of the three testing procedures at each MTL level.

DISCUSSION

Optimal rules for the sequential mastery problem (nonmastery, mastery, or continue testing) were derived using the framework of Bayesian sequential decision theory. The binomial probability distribution was assumed for modeling response behavior, whereas threshold loss was adopted for the loss function involved. Prior knowledge about student's true level of knowledge, needed for computing the posterior expected losses as well as the posterior predictive distributions, was assumed to be represented by the uniform prior.

In a Monte Carlo simulation, the Bayesian sequential procedure was compared with a conventional fixed-length mastery test and the SPRT procedure. Maximum test length varied from 10 to 50 items. For the specific parameter values chosen in this simulation study, it turned out that the Bayesian sequential strategy was most efficient (i.e., combination of highest classification accuracy and shortest average number of items to be administered) for test pools reflecting the one-parameter logistic model (i.e., Rasch model) at each level of maximum test length.

It is important to notice, however, that the Bayesian sequential 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 Bayesian sequential strategy might be appropriate in psychodiagnostic. 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 Bayesian sequential approach might be considered as an alternative to other testing strategies, such as SPRT, AMT, or fixed-length mastery 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 Bayesian 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 Bayesian sequential procedure, 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 testing, the action declaring mastery, declaring nonmastery, or continue testing is then chosen which minimizes the posterior 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, the 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. Vos (1999) applied Bayesian sequential decision theory to SMT for both threshold and linear loss in case the three classification decisions of declaring nonmastery, partial mastery, and mastery are open to the decision-maker (see also Smith & Lewis, 1995).

RESUMEN

Un procedimiento bayesiano en el contexto de los tests secuenciales de clasificación. El propósito de este artículo consiste en obtener reglas óptimas en los tests secuenciales de clasificación. En un test secuencial de clasificación, la decisión a tomar es clasificar a una persona como maestro, no maestro, o continuar el test y administrar el siguiente ítem. Se aplica la teoría de la decisión secuencial bayesiana; es decir, las reglas óptimas resultan de minimizar las pérdidas esperados posteriores asociadas con todas las posibles reglas de decisión en cada momento de test. La principal ventaja de este acercamiento es que los costes de aplicar el test pueden ser tenidos en cuenta de manera explícita. Se asume el modelo binomial para determinar la probabilidad de respuesta correcta dado un cierto nivel de funcionamiento de la persona. Se adopta una pérdida umbral como función de perdida. El artículo termina con un estudio de simulación en el que la estrategia secuencial bayesiana se compara con otros procedimientos disponibles en la literatura para problemas de decisión similares.

Palabras clave: Tests secuenciales de clasificación, reglas secuenciales bayesianas, distribución binomial, pérdida umbral, estrategia más eficiente.

REFERENCES

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

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.

DeGroot, M.H. (1970). *Optimal statistical decisions*. New York: McGraw-Hill.

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

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.

Johnson, N.L., & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions*. Boston: Houghton Mifflin.

Kingsbury, G.G., & 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* (pp. 257-283). New York: Academic Press.

Lehmann, E.L. (1959). *Testing statistical hypotheses* (3rd ed.). New York: Macmillan.

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

Luce, R.D., & Raiffa, H. (1957). *Games and decisions*. New York: John Wiley and Sons.

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

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.

Pearson, K. (1930). *Tables for statisticians and biometricians*. London: Cambridge University Press.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

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* (pp. 237-257). New York: Academic Press.

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets.
Applied Psychological Measurement, *16*, 65-76.

Smith, R.L., & 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.

Spray, J.A. (1993). *Multiple-category classification using a sequential probability ratio test*

(Research Rep. No. 93-7). Iowa City, IA: American College Testing.

Spray, J.A., & Reckase, M.D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, *21*, 405-414.

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

van der Linden, W.J. (1990). Applications of decision theory to test-based decision making. In R.K. Hambleton & J.N. Zaal (Eds.), *New developments in testing: Theory and applications*, 129-155. Boston: Kluwer.

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

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

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

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

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

Vos, H.J., & 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. Kluwer-Nijhoff, Boston, MA.

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

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

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