LSAC RESEARCH REPORT SERIES

Testlet-Based Adaptive Mastery Testing

Cees A. W. Glas
Hans J. Vos
University of Twente, Enschede, The Netherlands

Law School Admission Council
Computerized Testing Report 99-11
May 2006

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are over 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

In mastery testing, the goal is to decide whether an examinee should be classified as a master or a nonmaster. Well-known examples of mastery testing include testing for pass/fail decision-making, licensure, and certification. The focus of this study was on a combination of sequential mastery testing (SMT) and adaptive mastery testing (AMT). In SMT, testing continues by selecting random items until a master or nonmaster decision can be made with prescribed expected loss. The main advantage of SMT is a shorter expected test length. In AMT, items are not selected at random, but to maximize information at the examinee's current ability estimate. Generally, AMT results in better precision of classifying the examinees as master or nonmaster and a decrease in the cost of testing. A combination of sequential and adaptive mastery testing (ASMT) will be used in this study.

In a previous report by Glas and Vos, a general theoretical framework for ASMT based on Bayesian sequential decision theory and item response theory (IRT) was presented. The performance of ASMT was investigated for the one-parameter logistic (1PL) IRT model both for the selection of individual items and for small intact sets of items (testlets) from the pool. In the present study, the approach was generalized to testlet-based testing under the regular and a hierarchical version of the three-parameter logistic (3-PL) model. Unlike the regular model, which assumes local independence to hold simultaneously within and between testlets, the hierarchical model assumes greater similarity between responses in the same testlets than in different testlets.

The results showed that the expected losses associated with the decisions made by SMT and ASMT conditions were much smaller than for mastery testing based on a linear test. Also, ASMT produced considerably better results than SMT for the two 3-PL models relative to those for the 1-PL model in the previous study. Finally, the 3-PL model with the hierarchical structure that allowed for local dependence in testlets realized a substantial reduction in expected loss relative to the regular 3-PL model. The general conclusion from this study is that mastery testing can be conducted best in the ASMT format under a hierarchical 3-PL model for the item responses.

Abstract

In a previous report by Glas and Vos, a general theoretical framework for adaptive mastery testing was developed that was based on a combination of Bayesian sequential decision theory and item response theory (IRT). It was pointed out how IRT-based sequential mastery testing can be generalized to incorporate adaptive item and testlet selection rules. The performance of the approach was studied in a number of simulations using the Rasch model. In the present report, the approach is generalized in two directions. Firstly, it is shown how the method can be implemented for a more complicated model than the Rasch model; the 3-PL model is used as an example. Secondly, it is shown how specific dependencies emanating from the testlet structure can be explicitly taken into account in the IRT model used. The performance of the method is shown in a number of simulation studies. Further, a number of simulation studies are carried out to address the question of how ignoring the testlet structure affects the precision of the decision procedure.

Introduction

In mastery testing the problem is to decide whether a testee must be classified as a master or a nonmaster. The decision is based on the testee's observed test score. Well-known examples of mastery testing include pass/fail decisions, licensure, and certification. The mastery test can have a fixed-length or variable-length form. In the fixed-length mastery test, the performance on a fixed number of items is used for deciding either mastery or nonmastery. Over the last few decades, the fixed-length mastery problem has been studied extensively by many researchers (e.g., De Gruijter & Hambleton, 1984; van der Linden, 1990). Most of these authors derived analytically or numerically optimal rules by applying (empirical) Bayesian decision theory (e.g., DeGroot, 1970; Lehmann, 1986) to this problem. In the variable-length form, in addition to the actions declaring mastery or nonmastery, the action of continuing to administer items is available (e.g., Kingsbury & Weiss, 1983; Lewis & Sheehan, 1990; Sheehan & Lewis, 1992; Spray & Reckase, 1996). Items could be administered one at a time or in batches of more than one item. These batches are usually denoted testlets. It is plausible that responses to items within a batch may be more strongly related than the responses to items of different batches. One could think of a language comprehension test consisting of a number of texts, where each text is followed by a small number of items.

The main advantage of variable-length mastery tests as compared to fixed-length mastery tests is that they offer the possibility of providing shorter tests for those testees who have clearly attained a certain level of mastery (or clearly nonmastery) and longer tests for those for whom the mastery decision is not as clear-cut (Lewis & Sheehan, 1990). For instance, Lewis and Sheehan (1990) showed in a simulation study that average test lengths could be reduced by half without sacrificing classification accuracy for clear masters or nonmasters.

Two main types of variable-length mastery tests can be distinguished. First, the next item or testlet to be administered can be randomly selected. In this case, the stopping rule (i.e., termination criterion) is adaptive but the item selection procedure is not adaptive. This type of variable-length mastery testing is also known as sequential mastery testing, and, in the sequel, it will be referred to as SMT. In the second main type of variable-length mastery testing, not only the stopping rule but also the selection mechanism is adaptive. The testee's ability level is estimated after each response to an item or testlet, and the difficulty of the item or testlet administered next is tailored to the testee's estimated proficiency level. Kingsbury and Weiss (1983) denote this type of variable-length mastery testing as adaptive mastery testing (AMT).

In a previous report, Glas and Vos (2005) presented a procedure for variable-length mastery testing using a Bayesian sequential decision-theoretic framework. This approach is labeled ASMT (adaptive sequential mastery testing). In this framework, the costs of administering items can explicitly be taken into account. In addition, the item or testlet to be administered next is not randomly selected but is adapted to testee's last ability estimate. Hence, the strong points of both approaches are combined; that is, the selection as well as the stopping rule are adaptive and the cost per observation is explicitly taken into account.

The purpose of the present report is to adapt the ASMT approach to the 3-PL model, and to a generalization of the 3-PL model to an IRT model for testlets. This report is organized as follows. Firstly, a concise review of existing literature and earlier solutions to the variable-length mastery problem is presented. Secondly, the combined approach of sequential and adaptive mastery testing will be described. Then, this general approach will be adapted to the 3-PL model and the 3-PL model for testlets. Following this, a number of simulation studies will be presented that focus on the gain of a sequential procedure over a fixed-length test and the gain of an adaptive sequential test over a classical sequential test. Gain will be defined in terms of the average loss, the average number of items administered, and the percentage of correct decisions. Further, the impact of ignoring the testlet structure on the performance of the procedure will be studied. The report concludes with some new lines of research emanating from applying Bayesian sequential decision theory to AMT.

A Review of Earlier Solutions to the Variable-Length Mastery Problem

In this section, earlier solutions from existing literature to the variable-length mastery problem will be briefly reviewed. First, the application of Wald's (1947) sequential probability ratio test (SPRT) to SMT will be considered. Next, IRT-based AMT strategies will be reviewed. Finally, contributions of Bayesian decision theory to SMT will be presented.

Contributions of SPRT to Variable-Length Mastery Testing

The application of Wald's SPRT, originally developed as a statistical quality control test for light bulbs in a manufacturing setting, to SMT dates back to Ferguson (1969), in which testee's responses to items were assumed to be binomially distributed. Reckase (1983) proposed alternative sequential procedures within an SPRT framework, which, as opposed to Ferguson's approach, did not assume that items have equal difficulty but allowed them to vary in difficulty and discrimination by using an IRT model instead of a binomial distribution (see also Spray & Reckase, 1996).

IRT-Based Item Selection Strategies Applied to Adaptive Mastery Testing

Two IRT-based item selection strategies have been primarily used in implementing AMT. First, Kingsbury and Weiss (1983) proposed to select the item next that maximizes the amount of information at the testee's last ability estimate. In the second approach, the Bayesian item selection strategy, the item to be administered next is the one that minimizes the posterior variance of the testee's last ability estimate. Furthermore, a prior distribution about the testee's ability level must be specified in this approach before the administration of the first item. 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.

Sequential Mastery Testing Based on Bayesian Decision Theory

As mentioned before, most researchers applied (empirical) Bayesian decision theory to the fixed-length mastery problem. Within a Bayesian decision-theoretic framework, the following two basic elements must be specified: a psychometric model for the probability of answering an item correctly given a testee's ability level (i.e., the response function), and a loss structure evaluating the total costs and benefits of all possible decision outcomes. These costs may concern all relevant psychological, social, and economic decisions, which the decision brings along. The Bayesian approach allows the decision-maker to incorporate into the

decision process the costs of misclassifications. Furthermore, a prior distribution must be specified representing prior knowledge of the testee's ability level. Finally, a cut-off point on the latent ability scale separating masters and nonmasters must be specified in advance by the decision maker using methods of standard setting (e.g., Angoff, 1971). In a Bayesian approach, optimal rules are now obtained by minimizing the posterior expected loss associated with each possible decision outcome.

Lewis and Sheehan (1990), Sheehan and Lewis (1992), and Smith and Lewis (1995) have recently applied Bayesian sequential decision theory to the variable-length mastery problem. In addition to the elements needed in a Bayesian decision-theoretic approach, cost per observation must be explicitly specified in this framework. The cost of administering one additional item can be considered as an extension of the loss structure for the fixed-length mastery problem to the variable-length mastery problem. Posterior expected losses associated with the nonmastery and mastery decisions can now be calculated at each stage of sampling. The posterior expected loss associated with continuing sampling is determined by averaging the posterior expected loss associated with each of the possible future decision outcomes relative to the probability of observing those outcomes (i.e., the posterior predictive distributions). Analogous to the fixed-length mastery problem, the optimal sequential rule is found by selecting the action (i.e., mastery, nonmastery, or to continue sampling) that minimizes posterior expected loss at each stage of sampling using techniques of dynamic programming (i.e., backward induction). This technique 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. As indicated by Lewis and Sheehan (1990), the action selected at each stage of sampling is optimal with respect to the entire SMT procedure.

Vos (1999) also applied Bayesian sequential decision theory to SMT. Like Smith and Lewis (1995), he assumed three classification categories (i.e., nonmastery, partial mastery, and mastery). However, as in Ferguson's (1969) SPRT-approach, for the conditional probability of a correct response given the testee's ability level, the binomial distribution instead of an IRT model is considered. This modeling of response behavior corresponds to the assumption that all items have equal difficulty, or that items are sampled at random from a large (real or hypothetical) pool of items. Doing so, and assuming prior knowledge about a testee's ability can be represented by a beta prior $B(\alpha, \beta)$ (i.e., its natural conjugate), it is shown that the number-correct score is sufficient to calculate the posterior expected losses at future stages of the mastery test.

Bayesian Sequential Decision Theory Applied to Adaptive Mastery Testing

In this section, the new approach of applying Bayesian sequential decision theory to AMT will be introduced. Before doing so, some necessary notations will be introduced and the variable-length mastery problem in general will be formalized. Then, the loss function assumed will be discussed. Next, it will be shown how IRT models can be incorporated to support ASMT.

Formalization of the Variable-Length Mastery Problem

In the following, it will be assumed that the variable-length mastery problem consists of S ($S \geq 1$) stages labeled $s = 1, ..., S$ and at each stage one of the testlets can be given. At each stage of sampling, one or more items labeled i are administered and the observed item response for a randomly sampled student will be denoted by a discrete random variable U_{si}, with realization u_{si}. It is assumed that the random variables U_{si} are independent and identically distributed and that U_{si} takes the value 0 and 1 for a correct and incorrect response, respectively. Let $\mathbf{u}_s$ be the response to the s-th testlet. For $s = 1, ..., S$, the decisions will be based on a statistic $\mathbf{w}_s$, which is a function of the response patterns $\mathbf{u}_s$, that is, $\mathbf{w}_s = f(\mathbf{u}_1, ..., \mathbf{u}_s)$. In many cases, $\mathbf{w}_s$ will be the response pattern $\mathbf{u}_1, ..., \mathbf{u}_s$, itself. However, below it will become clear that some computations are only feasible if the information of the complete response pattern is aggregated. At each stage of sampling s ($s = 1, ..., S - 1$), a decision rule $d(\mathbf{w}_s)$ can be defined as

$$\mathrm{d}(\mathbf{w}_s) = \begin{cases} m \text{ the respondent is judged a master, sampling stops;} \\ n \text{ the respondent is judged a nonmaster, sampling stops;} \\ c \text{ sampling is continued.} \end{cases} \tag{1}$$

At the final stage of sampling, only the two mastery classification decisions m and n are available. Mastery will be defined in terms of the latent ability continuum of the IRT model. Therefore, let θ and θ_c denote the testee's ability level and some prespecified cut-off point on the latent continuum, respectively. Examinees with ability θ below this cut-off point are considered nonmasters, while testees with ability θ above this cut-off point are considered masters.

Linear Loss

As noted before, a loss function evaluates the total costs and benefits for each possible combination of action and testee's ability θ. Unlike Lewis and Sheehan (1990), Sheehan and Lewis (1992), and Smith and Lewis (1995), threshold loss will not be adopted here. The reason is that this loss function, although frequently used in the literature, may be less realistic in some applications. An obvious disadvantage of threshold loss is that it assumes the same constant loss for testees with ability θ to the left or to the right of θ_c, no matter how large their distance from θ_c. For instance, it seems more realistic to assume that for nonmasters the loss is an increasing function of θ. Moreover, the threshold loss function is discontinuous; at the cut-off point θ_c, this function "jumps" from one constant value to another. This sudden change seems unrealistic in many decision-making situations. 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 the testee's ability level θ (see also Huynh, 1980; van der Linden & Vos, 1996; Vos, 1997a, 1997b, 1999).

For the variable-length mastery problem, the linear loss function for the master and nonmaster decision can be restated at each stage of sampling as

$$L(m,\theta) = max\{sC,\ sC + A(\theta - \theta_c)\} \tag{2}$$

with $A < 0$ and

$$L(n,\theta) = max\{sC,\ sC + B(\theta - \theta_c)\}, \tag{3}$$

with $B > 0$; C is the cost of delivering one testlet; sC is the cost of delivering s testlets. For the sake of simplicity, following Lewis and Sheehan (1990), these costs are assumed to be equal for each decision outcome as well as for each sampling occasion. The above defined function consists of a constant term and a term proportional to the difference between the testee's ability level θ and the specified cut-off point θ_c. The condition $A < 0$ and $B > 0$ is equivalent to the statement that for action m, loss is a strictly decreasing function of the latent ability θ, whereas loss for action n is assumed to be strictly increasing in θ. The definitions (2) and (3) guarantee that the losses are at least sC. As a departure from van der Linden and Mellenbergh (1977), for instance, this specific formulation of linear loss is chosen because in many problems it has been convenient to work with non-negative loss functions (see, for instance, DeGroot, 1970, p.125).

The loss parameters A, B, and C have to be either theoretically or empirically assessed. For assessing loss functions empirically, most texts on decision theory propose lottery methods (e.g., Luce & Raiffa, 1957, Chapter 2). In general, these methods use the notions of desirability of outcomes to scale the consequences of each pair of actions and testee's ability level.

At stage s, the decision whether the respondent is a master or a nonmaster, or whether another testlet will be administered, is based on the expected losses of the three possible decisions given the observation $\mathbf{w}_s$. The expected losses of the first two decisions are computed as

$$E(L(m,\theta) \mid \mathbf{w}_s) = sC + A\int_{-\infty}^{\theta_c} (\theta - \theta_c)\, p(\theta \mid \mathbf{w}_s)\, d\theta \tag{4}$$

and

$$E(L(n,\theta) \mid \mathbf{w}_s) = sC + B\int_{\theta_c}^{\infty} (\theta - \theta_c)\, p(\theta \mid \mathbf{w}_s)\, d\theta, \tag{5}$$

where $p(\theta \mid \mathbf{w}_s)$ is the posterior density of θ given $\mathbf{w}_s$. The expected loss of the third possible decision is computed as the expected risk of continuing testing. If the expected risk of continuing testing is smaller than the expected loss of a master or a nonmaster decision, testing will be continued. The expected risk of continuing testing is defined as follows. Let $\{\mathbf{w}_{s+1}\}$ be the range of $\mathbf{w}_{s+1}$. Then, for $s = 1, \ldots, S-1$, the expected risk of continuing testing is defined as

$$E\left[R(\mathbf{w}_{s+1} \mid \mathbf{w}_s)\right] = \sum_{\{\mathbf{w}_{s+1}\}} p(\mathbf{w}_{s+1} \mid \mathbf{w}_s) R(\mathbf{w}_{s+1} \mid \mathbf{w}_s), \tag{6}$$

where the so-called posterior predictive distribution $p(\mathbf{w}_{s+1} \mid \mathbf{w}_s)$ is given by

$$p(\mathbf{w}_{s+1} \mid \mathbf{w}_s) = \int p(\mathbf{w}_{s+1} \mid \theta) p(\theta \mid \mathbf{w}_s) d\theta, \tag{7}$$

and risk is defined as

$$R(\mathbf{w}_{s+1} \mid \mathbf{w}_s) = min\{E(L(m, \theta) \mid \mathbf{w}_{s+1}),$$

$$E(L(n, \theta) \mid \mathbf{w}_{s+1}), E[R(\mathbf{w}_{s+2} \mid \mathbf{w}_{s+1})]. \tag{8}$$

The risk associated with the last testlet is defined as

$$R(\mathbf{w}_S \mid \mathbf{w}_{S-1}) = min\{E(L(m, \theta \mid \mathbf{w}_S)), E(L(n, \theta \mid \mathbf{w}_S))\}. \tag{9}$$

So, given an observation $\mathbf{w}_s$, the expected distribution of $\mathbf{w}_{s+1}$, $\mathbf{w}_{s+2}$, ..., $\mathbf{w}_S$ is generated and an inference about future decisions is made. Based on these inferences, the expected risk of continuation (6) is computed and compared with the expected losses of a mastery or nonmastery decision. If the risk of continuation is lower than these two expected losses, testing is continued. If this is not the case, the classification decision with the lowest expected loss is made.

Notice that definition 8 implies a recursive definition of the expected risk of continuation. In practice, the computation of the expected risk of continuing testing can be done by backward induction as follows. First, the risk of the last testlet (9) is computed for all possible values of $\mathbf{w}_S$. Then, the posterior predictive distribution $p(\mathbf{w}_S \mid \mathbf{w}_{S-1})$ is computed using (7), followed by the expected risk $E(R(\mathbf{w}_S \mid \mathbf{w}_{S-1})$ defined in (6). This, in turn, can be used for computing the risk $R(\mathbf{w}_{S-1} \mid \mathbf{w}_{S-2})$ for all $\mathbf{w}_{S-1}$ using (8), and this iterative process continues until s is reached and the decision can be made whether to administer testlet $s + 1$, or to decide on mastery or nonmastery.

The 3-PL Testlet Model

In many test situations, the testee is presented with a number of clusters of items rather than a set of independent items. Above, the example of a language comprehension test consisting of a number of texts with items nested under texts was given. In IRT models used for analyzing data from tests made up of testlets, the dependence structure must be explicitly taken into account. Recently, Bradlow, Wainer, and Wang (1999) have presented a multilevel IRT model based on the 3-PL model that meets this requirement. A generalization to the 3-PL model was proposed by Wainer, Bradlow and Du (1999). These authors used an MCMC (Markov Chain Monte Carlo) procedure to estimate the parameters in the model. Glas, Wainer and Bradlow (1999) have proposed an MML (marginal maximum likelihood) estimation procedure in which they show how the model can be used in a CAT environment. In the present report, this model will be used in the framework of ASMT.

As above, $\mathbf{u}_s$ is the response pattern on testlet s, and $\mathbf{u}$ is the response pattern on the complete test, that is, $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_s, ..., \mathbf{u}_S)$. It is assumed that the person parameters θ have a standard normal distribution. Further, it is assumed that the response pattern of testlet s depends on an ability parameter ξ_s, and conditionally on θ, the parameters ξ_s are independent and normally distributed with mean zero and variance equal to τ^2, that is, $\xi_s \mid \theta \sim N(\theta, \tau^2)$. Then, the variables $\xi_1 - \theta$, $\xi_2 - \theta$, ..., $\xi_s - \theta$, ..., $\xi_S - \theta$, and θ have a joint multivariate normal distribution given by

$$\begin{bmatrix} \xi_1 - \theta \\ \xi_2 - \theta \\ \dots \\ \xi_S - \theta \\ \theta \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 & 0 & \dots & 0 & 0 \\ 0 & \tau^2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \tau^2 & 0 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} \right).$$

Without conditioning on the ability parameter θ, the testlet-related ability parameters ξ_s are dependent. To see this, consider the transformation

$$\begin{bmatrix} 1 & 0 & \dots & 0 & 1 \\ 0 & 1 & \dots & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & \dots & 0 & 1 \end{bmatrix} \begin{bmatrix} \xi_1 - \theta \\ \xi_2 - \theta \\ \dots \\ \xi_S - \theta \\ \theta \end{bmatrix} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \dots \\ \xi_S \\ \theta \end{bmatrix}.$$

Then it follows that the joint distribution of the ability variables is a multivariate normal distribution given by

$$\begin{bmatrix} \xi_1 \\ \xi_2 \\ \dots \\ \xi_S \\ \theta \end{bmatrix} \sim N\left[\begin{bmatrix} 0 \\ 0 \\ \dots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1+\tau^2 & 1 & \dots & 1 & 1 \\ 1 & 1+\tau^2 & \dots & 1 & 1 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1+\tau^2 & 1 \\ 1 & 1 & \dots & 1 & 1 \end{bmatrix} \right].$$

So summing up, the model has an hierarchical structure: For every testlet, the respondent independently draws an ability parameter ξ_s from a distribution with mean θ and within-person ability variance τ^2. However, the joint distribution of the ability parameters ξ_s, $s = 1, ..., S$, is dependent. With a reparametrization $\gamma_s = \xi_s - \theta$, this model is equivalent to the Bradlow, Wainer, and Wang (1999) and Wainer, Bradlow, and Du (1999) model.

Glas, Wainer, and Bradlow (1999) argue that in a CAT framework, one is interested in estimating an examinee's ability parameter θ and that the within-person ability variance τ^2 can be considered error variance. A reasonable estimate of ability is the so-called EAP estimate, which is the a-posteriori expectation of θ. In the present case, the EAP estimate is given by

$$E(\theta \mid \mathbf{u}) = \frac{\int \theta \left[\prod_s \int p(\mathbf{u}_s \mid \xi_s) h(\xi_s \mid \theta, \tau) d\xi_s \right] g(\theta) d\theta}{\int \left[\prod_s \int p(\mathbf{u}_{s\mid} \xi_s) h(\xi_s \mid \theta, \tau) d\xi_s \right] g(\theta) d\theta} \tag{10}$$

where $p(\mathbf{u}_s \mid \xi_s)$ is the probability of observing response pattern $\mathbf{u}_s$ as a function of ξ_s, as specified by the IRT model; $h(\xi_s \mid \theta, \tau)$ is the density of ξ_s and $g(\theta)$ is the density of θ.

The 3-PL Testlet Model and Sequential Mastery Testing

Unlike the Rasch model, the 3-PL model has no minimal sufficient statistic for ability. Therefore, one approach of applying the general framework for sequential testing to the 3-PL model would be to substitute the complete response pattern $(\mathbf{u}_1, ..., \mathbf{u}_s)$ for $\mathbf{w}_s$. For the testlets where the responses are already known, say the testlets 1, ..., s^*, this presents no problem. But for evaluation of $E(R(\mathbf{w}_s + 1, \mathbf{w}_s) \mid \mathbf{w}_s)$, $s \geq s^*$, however, this entails a summation over the set of all possible response patterns on the future testlets, and exact computation of this expected risk generally presents a major problem. One of the approaches to this problem is approximating (6) using Monte Carlo simulation techniques, that is, simulating a large number of draws from $p(\mathbf{w}_s + 1 \mid \mathbf{w}_s)$ to compute the mean of $R(\mathbf{w}_s + 1, \mathbf{w}_s)$ over these draws. However, this approach proves quite time consuming and is beyond the scope of the present report. The approach adopted here is assuming that the unweighted sum score contains much of the relevant information provided by the testlets $s + 1, ..., S$ with respect to θ. This is motivated by the fact that the expected number-right score $\sum_i P_i(\theta)$ is monotonously increasing in θ (see, for instance, Lord, 1980, pp.46–49). Therefore, the following procedure is used.

Suppose that s^* testlets have been administered. Then, $\mathbf{w}_s$ will be the observed response pattern on the s^* testlets and the, as yet, unobserved sum score on the testlets $d = s^* + 1, ..., s$. So let t_d be the sum score on testlet d, that is, $t_d = \sum_i u_{di}$, and let $r_{s^*,s}$ be the sum over the scores of the testlets $d = s^* + 1, ..., s$, that is, $r_{s^*,s} = \sum_{d=s^*+1}^{s} t_d$. Then $\mathbf{w}_s = (\mathbf{u}_1, ..., \mathbf{u}_{s^*}, r_{s^*,s})$. As described in the previous section, it will be assumed that for

every testlet, the respondent independently draws an ability parameter ξ_s from a normal distribution with mean θ and within-person ability variance τ^2. Then, the probability of response pattern $\mathbf{u}_s$ can be written as

$$p(\mathbf{u}_s \mid \xi_s) = \prod_i P_{is}(\xi_s)^{u_{is}} (1 - P_{is}(\xi_s))^{1-u_{is}}.$$

The probability of a sum score t_s is computed by summing over the set of all possible response patterns $\mathbf{u}_s$, resulting in a sum score t_s, denoted by $\{\mathbf{u}_s \mid t_s\}$, that is,

$$p(t_s \mid \xi_s) = \sum_{\{\mathbf{u}_s \mid t_s\}} p(\mathbf{u}_s \mid \xi_s).$$

The recursion formulas needed for the computation of these probabilities can, for instance, be found in Kolen and Brennan (1995, pp. 181–183). The probability of t_s conditional on θ is given by

$$p(t_s \mid \theta; \tau) = \int p(t_s \mid \xi_s) h(\xi_s \mid \theta; \tau) d\xi_s.$$

Finally, let $\{t_{s^*+1}, ..., t_s \mid r_{s^*,s}\}$ be the set of all testlet scores $t_{s^*+1}, ..., t_s$ compatible with a total score $r_{s^*,s}$. Then,

$$p(r_{s^*,s} \mid \theta; \tau) = \sum_{\{t_{s^*+1}, ..., t_s \mid r_{s^*,s}\}} p(t_s \mid \theta; \tau)$$

and the posterior distribution of θ given $\mathbf{w}_s$ is given by $p(\theta \mid \mathbf{w}_s; \tau) \propto p(\mathbf{u}_1, ..., \mathbf{u}_{s^*} \mid \theta; \tau) p(r_{s^*,s} \mid \theta; \tau) g(\theta)$. Inserting this definition into (4) to (9) defines the sequential mastery testing procedure for the 3-PL testlet model.

Performance of Sequential and Adaptive Sequential Mastery Testing

The first research questions addressed in this section will be whether sequential testing improves upon a fixed test and whether adaptive sequential testing improves upon sequential testing. The second research question is to what extent ignoring the testlet structure and performing the ASMT procedure as if the variance component τ^2 did not exist lead to deterioration of precision.

The design of the studies will be explained using Table 1. This table pertains to simulation studies where $\tau = 0.0$, that is, to studies where the variance due to the presence of testlets is zero so that the standard 3-PL model holds. The study concerns a situation where, at most, 40 items are administered. The results in the first panel of the table pertain to a cut-off point θ_c equal to 0.10. The second panel pertains to a cut-off point equal to 1.00. Consider the first panel. The 10 rows represent 10 simulation studies of 2,000 replications each. In the model, the prior distribution of ability is standard normal. Therefore, for every replication, a true θ was drawn from a standard normal distribution. In the first simulation study, every one of the 2,000 simulees was presented a fixed test of 40 items. For every simulee, the item difficulties were drawn from a standard normal distribution, and the item discrimination parameters were drawn from a log-normal distribution with parameters $\mu = 0.0$ and $\sigma^2 = 0.25$. The guessing parameter was equal to 0.25 for all items.

TABLE 1

Relation between selection method and loss—K = 40, θ ~ N(0, 1), τ = 0.0, 2,000 replications per study

Number of Testlets	Items Per Testlet	Selection Method	Proportion Testlets Given	Proportion Correct Decisions	Mean Loss
Cut-off Point $\theta_c = 0.10$					
1	40	fixed test	1.00	.87	.4440
4	10	sequential	.38	.83	.2297
4	10	cutting point	.40	.89	.2007
4	10	eap estimate	.40	.89	.2026
10	4	sequential	.33	.84	.2098
10	4	cutting point	.30	.91	.1434
10	4	eap estimate	.28	.92	.1353
40	1	sequential	.40	.85	.2316
40	1	cutting point	.15	.88	.1192
40	1	eap estimate	.20	.93	.1084
Cut-off Point $\theta_c = 1.00$					
1	40	fixed test	1.00	.93	.4278
4	10	sequential	.32	.91	.1699
4	10	cutting point	.36	.93	.1730
4	10	eap estimate	.36	.92	.1748
10	4	sequential	.29	.91	.1526
10	4	cutting point	.26	.93	.1324
10	4	eap estimate	.25	.92	.1322
40	1	sequential	.39	.91	.1922
40	1	cutting point	.11	.90	.0990
40	1	eap estimate	.13	.96	.0645

The remaining 9 rows of the first panel relate to a 2-factor design, the first factor being the test administration design, the second being the selection method. The test administration design is displayed in the first 2 columns of the table. The 3 designs used were 4 testlets of 10 items, 10 testlets of 4 items and 40 testlets of 1 item each. Three selection methods were studied. In the studies labeled "sequential," the SMT procedure was used. The studies labeled "cutting point" and "eap estimate" entail ASMT procedures. The label "eap estimate" refers to a selection procedure based on maximum information at the EAP estimate of ability as defined by (10). In the studies labeled "cutting point," testlets were selected with maximum information at the cutting point θ_c. For all simulations, the parameters of the loss functions (2) and (3) were equal to $A = -1.00$, $B = 1.00$, and $C = 0.01k_t$, where k_t stands for the number of items in a testlet. The motivation for this choice of C is keeping the total cost of administering 40 items constant.

For the SMT condition, the item difficulties of the first testlet were all equal to zero, and all other item parameters were drawn as described above. Also, in the ASMT conditions, the first testlet had all item difficulties equal to zero. The reason for starting both the SMT and ASMT procedures with testlets with similar item difficulties was to create comparable conditions in the initial phase of the procedures. The subsequent testlets were chosen from a bank of 50 testlets that was generated as follows. First, $50k_t$ item difficulties were drawn from the standard normal distribution and $50k_t$ item discrimination parameters were drawn from the lognormal distribution with parameters $\mu = 0.0$ and $\sigma^2 = 0.25$. Then, these $50k_t$ items were ordered with respect to the point on the latent continuum where they reached their maximum information. The first k_t items comprised the first testlet in the bank, the second k_t items comprised the second testlet, etc. In this way, 50 testlets were created that were relatively homogeneous in difficulty and attained their maximum information at distinct points of the latent ability scale. In the "'eap estimate" condition, at stage s, $s = 1, \ldots, S - 1$, an EAP estimate of ability as defined by (10) was computed and the expected risk of a "continue sampling" decision was computed using the item parameters of the $S - s$ testlets with highest information at this estimate. If a "continue sampling" decision was made, the next testlet administered was the most informative testlet of the $S - s$ testlets initially selected. For the procedure in the "cutting point" condition, testlets were selected from the testlet bank described above that were most informative at the cutting point θ_c.

The results of the simulation studies for $\tau = 0.0$ are given in the last three columns of Table 1 in terms of the proportion of testlets given, the proportion of correct decisions, and the mean loss over 2,000 replications. With respect to the latter measure, for every replication, loss was computed using (2) or (3) evaluated at the true value of θ, with s as the number of testlets actually given. From Table 1, the following conclusions can be drawn.

- Mean loss in the condition where the cut-off point was equal to 1.00 was lower than in the condition where the cut-off point was 0.10. This is explained by the fact that the mean distance between the cut-off point and ability is larger when the cut-off point is 1.00, so in that case, overall uncertainty with respect to the mastery status is lower.

- Mean loss in the SMT and ASMT conditions was much lower than in the fixed test condition. So, the conclusion is that a sequential strategy has effect here.

- Mean loss in the ASMT condition was lower than in the SMT condition in all cases except when the cut-off point was 1.00 and the number of testlets was 4. So, the conclusion is that adaptive testlet selection also has a positive effect with regards to mean loss.

- Within the two ASMT conditions, there was no clear effect of the "cutting point" condition versus the "eap estimate" condition.

- Mean loss decreased as the number of testlets increased.

- It can be seen that the decrease of mean loss was mainly due to a dramatic reduction in the proportion of testlets given. The number of correct classifications remained stable, so the decrease was not brought at the expense of an increased proportion of incorrect decisions.

In the three columns under the heading "testlet model" of Table 2 and Table 3, analogous results are given for $\tau = 0.5$ and $\tau = 1.0$, respectively. Comparing these columns with the last three columns of Table 1, it can be seen that mean loss increases in most conditions. This is explained by the fact that τ^2 acts as a source of error-variance that increases with the posterior variance of θ. Further inspection of these results shows that they are consistent with the results of Table 1, so the conclusions summarized above are corroborated. Finally, under the heading "3-PL model," the last three columns of Table 2 and Table 3 give the results of simulation studies where the standard 3-PL model was used instead of the 3-PL testlet model. That is, though the data were simulated using $\tau = 0.5$ and $\tau = 1.0$, respectively, the computations for the decision procedure were carried out using $\tau = 0.0$. It can be seen that the overall pattern of mean loss noted in Table 1 remains the same, but under most conditions, mean loss increases somewhat. So, it can be concluded that ignoring the testlet structure results in a decrease of precision, but this decrease is limited.

TABLE 2
Relation between selection method and loss—K = 40, θ ~ N(0, 1), τ = 0.5, 2,000 replications per study

			Testlet Model			3-PL Model		
Number of Testlets	Items per Testlet	Selection Method	Proportion Testlets Given	Proportion Correct Decisions	Mean Loss	Proportion Testlets Given	Proportion Correct Decisions	Mean Loss
Cut-off Point θ_c = 0.10								
1	40	fixed test	1.00	.81	.5006	1.00	.82	.4957
4	10	sequential	.40	.82	.2610	.37	.80	.2646
4	10	cutting point	.43	.85	.2408	.39	.84	.2369
4	10	eap estimate	.42	.85	.2444	.40	.83	.2472
10	4	sequential	.28	.84	.2043	.33	.82	.2322
10	4	cutting point	.26	.86	.1736	.29	.87	.1737
10	4	eap estimate	.28	.86	.1640	.28	.88	.1632
40	1	sequential	.32	.84	.2105	.40	.82	.2559
40	1	cutting point	.17	.87	.1331	.15	.86	.1324
40	1	eap estimate	.18	.88	.1203	.21	.89	.1325
Cut-off Point θ_c = 1.00								
1	40	fixed test	1.00	.90	.4488	1.00	.89	.4625
4	10	sequential	.33	.90	.1865	.31	.89	.1880
4	10	cutting point	.34	.89	.1859	.37	.91	.1855
4	10	eap estimate	.34	.89	.1855	.36	.92	.1812
10	4	sequential	.18	.88	.1437	.28	.91	.1535
10	4	cutting point	.22	.92	.1218	.25	.92	.1302
10	4	eap estimate	.15	.90	.1156	.23	.91	.1329
40	1	sequential	.18	.89	.1343	.38	.91	.1918
40	1	cutting point	.13	.93	.0765	.11	.90	.1058
40	1	eap estimate	.13	.94	.0759	.14	.95	.0710

TABLE 3
Relation between selection method and loss—K = 40, θ ~ N(0, 1), τ = 1.0, 2,000 replications per study

			Testlet Model			3-PL Model		
Number of Testlets	Items per Testlet	Selection Method	Proportion Testlets Given	Proportion Correct Decisions	Mean Loss	Proportion Testlets Given	Proportion Correct Decisions	Mean Loss
Cut-off Point θ_c = 0.10								
1	40	fixed test	1.00	.74	.5879	1.00	.73	.5976
4	10	sequential	.35	.74	.3238	.52	.79	.3383
4	10	cutting point	.38	.77	.3053	.51	.78	.3417
4	10	eap estimate	.38	.76	.3100	.50	.78	.3299
10	4	sequential	.32	.78	.2743	.35	.79	.2627
10	4	cutting point	.28	.80	.2253	.34	.83	.2225
10	4	eap estimate	.27	.80	.2262	.34	.82	.2224
40	1	sequential	.43	.80	.2954	.36	.82	.2551
40	1	cutting point	.16	.80	.1895	.21	.81	.1972
40	1	eap estimate	.25	.83	.1946	.22	.83	.1835
Cut-off Point θ_c = 1.00								
1	40	fixed test	1.00	.88	.4795	1.00	.82	.5564
4	10	sequential	.36	.88	.2106	.31	.85	.2209
4	10	cutting point	.36	.88	.2089	.36	.88	.2202
4	10	eap estimate	.35	.87	.2233	.37	.88	.2199
10	4	sequential	.19	.89	.1442	.28	.88	.1773
10	4	cutting point	.17	.88	.1339	.25	.90	.1477
10	4	eap estimate	.17	.89	.1381	.23	.89	.1506
40	1	sequential	.20	.90	.1332	.37	.90	.1955
40	1	cutting point	.16	.90	.1080	.12	.87	.1386
40	1	eap estimate	.18	.92	.1100	.15	.92	.1120

Discussion

In a previous report, Glas and Vos (2005) introduced a general theoretical framework for SMT based on a combination of Bayesian sequential decision theory and item response theory. In the present report, this general framework was further refined, and it was shown how this framework can be adapted to the 3-PL model and an IRT model for testlets. In the previous report (Glas & Vos, 2005), simulation studies using the Rasch model showed that SMT does indeed lead to a considerable decrease of loss, but reduction of loss due to adaptive testlet selection was less pronounced. In the present study, using the 3-PL model and the testlet response model, the former conclusion was corroborated. However, the latter conclusion does not generalize to the present IRT models: Adaptive item and testlet selection does result in substantial reduction of loss. An explanation might be that in the 3-PL and testlet response model, there is more variation in the amount of information that testlets supply, so picking the most informative testlet rather than a random testlet has a greater effect than in the Rasch model. However, this conclusion is quite tentative, and more research needs to be done to provide a definitive explanation for the phenomenon. The question of to what extent ignoring the testlet structure and performing the ASMT procedure using the regular 3-PL model lead to greater loss could not be answered unambiguously: Ignoring the testlet structure does generally lead to inflation of loss, but the effects are not very large. Also, more research needs to be done here to properly survey the extent of this phenomenon.

The general approach sketched here can be applied to several other IRT models, for instance, to multidimensional IRT models (see, for instance, McDonald, 1997, or Reckase, 1997). The loss structure involved must allow for both conjunctive and compensatory testing strategies in this situation. In decision theory, much work has already been done under the name of "multiple-objective decision making" (Keeney & Raiffa, 1976). It still needs to be examined how the results reported there could be applied to the problems of ASMT in cases of multidimensional IRT models.

Another point of further study is the adoption of minimax sequential decision theory instead of Bayesian sequential decision theory. Optimal rules are found in this approach by minimizing the maximum expected losses associated with all possible decision rules. In fact, as pointed out by van der Linden (1981), 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, and, therefore, can be considered as a bit conservative and pessimistic (Coombs, Dawes, & Tversky, 1970). Analogous to Bayesian sequential decision theory, the cost of test administration is also explicitly taken into account in this approach.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thornike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council of Education.

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168 .

Chang, H., & Stout, W. F. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58*, 37–52.

Coombs, C. H., Dawes, R. M., Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

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 cut-off 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, PA.

Glas, C. A. W., & Vos, H. J. (2005). *Adaptive sequential mastery testing using the Rasch model and Bayesian sequential decision theory*. Computerized Testing Report No. 99-02). Newtown, PA: Law School Admission Council.

Glas, C. A. W., Wainer, H., & Bradlow (1999). MML and EAP estimates for the testlet response model. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computer Adaptive Testing: Theory and practice*. Boston: Kluwer-Nijhoff.

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

Keeney, D., & Raiffa, H. (1976). *Decisions with multiple objectives: preferences and value trade-offs*. New York: Wiley.

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.

Kolen, M. J., & Brennan, R. L. (1995). *Test Equating*. New York: Springer.

Lehman, E. L. (1986). *Testing statistical hypothesis* (2nd ed.). New York: Wiley.

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

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

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

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden and R. K. Hambleton (Eds.). *Handbook of modern item-response theory* (pp.257–269). New York: Springer.

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.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item-response theory* (pp.271–286). New York: Springer.

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., & 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*(3), 271–92.

Wainer, H., Bradlow, E. T., & Du, Z. (1999). Testlet response theory: An analogue for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.). *Computer adaptive testing: Theory and practice* (pp. 245–269). Boston: Kluwer-Nijhoff.

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