Using Bayesian Decision Theory to Design a Computerized Mastery Test

Charles Lewis and Kathleen Sheehan
Educational Testing Service

A theoretical framework for mastery testing based on item response theory and Bayesian decision theory is described. The idea of sequential testing is developed, with the goal of providing shorter tests for individuals who have clearly mastered (or clearly not mastered) a given subject and longer tests for those individuals for whom the mastery decision is not as clear-cut. In a simulated application of the approach to a professional certification examination, it is shown that average test lengths can be reduced by half without sacrificing classification accuracy. *Index terms: Bayesian decision theory, computerized mastery testing, item response theory, sequential testing, variable-length tests.*

Mastery testing is used in educational and certification contexts to decide, on the basis of test performance, whether an individual has attained a specified level of knowledge, or mastery, of a given subject. A central problem in designing a mastery test is that of maximizing the probability of making a correct mastery decision while simultaneously minimizing test length. A similar problem is frequently encountered in the field of quality control: Acceptance sampling plans must be designed to maximize the probability of correctly classifying the quality of a lot of manufactured material while simultaneously minimizing the number of items inspected. The solution to the acceptance sampling problem that was proposed by Wald (1947), called the sequential probability ratio test (SPRT), exploited the fact that a lot of very poor quality can be expected to reveal its character in a very small sample, whereas lots of medium quality will always require more extensive testing. This is done by testing one randomly selected unit at a time, while allowing for the possibility of a decision on the quality of the lot as a whole after each selection.

In an early application of the sequential testing approach to the mastery testing problem, Ferguson (1969a, 1969b) designed a sequential mastery test that treated examinees' responses to items as a sequence of independent Bernoulli trials. This design requires a pool of calibrated items that can be sampled randomly. The test is conducted by presenting items to examinees one at a time. After each item has been presented, a decision is made either to classify the examinee (as a master or a nonmaster) or to present another item. Ferguson also specified a maximum test length for those individuals for whom the mastery classification is very difficult to make. The decision rule assumes a binomial probability model for item responses and, as in the SPRT, is based on a likelihood ratio statistic.

A major advantage of this approach is that it allows for shorter tests for individuals who have clearly mastered (or clearly not mastered) the subject matter, and longer tests for those individuals for whom the mastery decision is not as clear-cut. The use of the binomial model implies that the probability of a correct response to an item is the same for all items in the pool, or that items are sampled at random.

Alternative sequential mastery testing procedures have been proposed by Reckase (1983) and by Kingsbury and Weiss (1983). Both of these procedures employ non-random adaptive item selection

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 14, No. 4, December 1990, pp. 367-386

0146-6216/90/040367-20$2.25

algorithms and are designed to be used with item pools containing items that vary in difficulty and discrimination. In each procedure, the next item to be presented to an individual is selected based on the amount of information that the item provides concerning the individual's achievement level estimate at that point in the testing process. This adaptive item sampling algorithm is implemented using methods derived from item response theory (IRT). The decision rule proposed by Reckase is a modification of the SPRT, in which the probability of a correct response to an item is allowed to vary from one item to the next. This probability is estimated using an IRT model. The procedure proposed by Kingsbury and Weiss differs from the Reckase procedure in that classification decisions are made using Bayesian confidence intervals.

An alternative IRT-based mastery testing procedure has been proposed by Lord (1980) and implemented by Stocking (1987). In this alternative approach, all examinees receive the same fixed-length test, but the test is designed, constructed, and scored using methods derived from IRT. An optimal test length is determined by specifying a maximum value for the length of the asymptotic confidence interval for estimating ability (θ) from test score in the region of the cutscore. This approach places no restrictions on the variability of items in the pool, but it does require that all examinees take the same fixed-length test.

A new type of mastery testing procedure is introduced in this paper that is constructed using IRT, like Lord (1980), and using Wald's sequential testing approach to provide an adaptive stopping rule, like Ferguson (1969a, 1969b) and Reckase (1983). This approach differs from those presented previously in that (1) the sequential testing process operates on blocks of items, called testlets, rather than individual items, and (2) the decision rule is determined using Bayesian decision theory. This new mastery test is called a computerized mastery test (CMT), because it is designed to be administered and scored using personal computers.

Design for a Computerized Mastery Test

Testlets

A testlet is an ordered collection of items that has been designed to be administered as a unit. One of the advantages of a testlet-based item presentation algorithm is that, regardless of the selection methodology employed, it is never necessary to restrict the item pool to equivalent items.

For example, a test that calls for adaptive testlet selection requires a pool of variably peaked testlets, or testlets that have been designed to be optimally discriminating at a series of fixed points along the θ scale. The fixed-length mastery testing procedure proposed by Lord (1980) could be used to construct such a pool. In contrast, a test that calls for random testlet selection requires a pool of parallel peaked testlets, or testlets that have been designed to be optimally discriminating at the same θ value (i.e., the cutscore). Lord's procedure again could be used to construct such a pool, in which each testlet would contain items that varied in difficulty and discrimination, but all testlets would be constructed to provide equivalent measurement at the cutscore. The feasibility of Lord's procedure for the case of equivalent testlets has been demonstrated by Stocking (1987). Because a testlet is basically a short test, it is not unreasonable to assume that any procedure developed to construct peaked tests could be modified to construct peaked testlets; also, any procedure developed to construct parallel test forms could be modified to construct parallel testlets, or testlets with equivalent measurement properties.

Testlet-based item presentation algorithms also provide a number of secondary advantages. For example, they allow for greater control over problems related to item-ordering and context effects (Wainer & Kiely, 1987). Context effects arise when the appearance of a particular item has an effect

on the difficulty of a subsequent item. Controlling this type of problem can be quite difficult when individual items are randomly selected from a pool. When items are ordered within testlets, however, a careful screening procedure can eliminate most dependency problems before they occur. Although individual testlet screenings cannot control for problems occurring between testlets, the impact of between-testlet item dependencies is lessened by the fact that items in separate testlets are typically spaced further apart in the presentation sequence.

Two additional advantages are: (1) test security is enhanced, because each examinee may potentially be administered a different subset of testlets (depending on the size of the testlet pool); and (2) the record-keeping burden associated with administering tests to repeat examinees is reduced, because the system must only keep track of the subset of five or six testlets administered to each examinee, as opposed to the (possibly) hundreds of individual test items.

Random Versus Adaptive Selection

The CMT design uses random testlet selection. The primary reasons for selecting random rather than adaptive sampling were (1) computational efficiency—unlike adaptive selection, random selection does not require the estimation of examinee θ levels at each stage of testing; (2) simplicity—the CMT was already designed to be adaptive in the sense of an adaptive stopping rule, so the additional complication of an adaptive testlet selection mechanism was not considered desirable; and (3) ease of implementation—unlike adaptive selection, which requires a pool of content-balanced testlets spanning a range of difficulty levels, random testlet selection requires a pool of parallel testlets. The specialized pools required for adaptive testlet selection in many testing programs are difficult to construct, whereas procedures for constructing parallel test forms are often available for use in constructing pools of parallel testlets.

The Definition of Mastery

In Ferguson's (1969a, 1969b) application of the SPRT to the mastery testing problem, the cutscore separating masters from nonmasters was defined in terms of the minimum proportion of correct responses needed to classify an examinee as a master. In Lord's (1980) treatment of mastery testing, an IRT model is used to characterize the relationship between observed test performance and true mastery status, and the cutscore is defined as a point θ_c on the latent achievement scale. Because it may not always be feasible to specify θ_c precisely, Lord also suggested an alternative mastery definition in which two values, θ_n and θ_m, are specified. θ_n is the highest level at which an examinee will be considered a nonmaster, and θ_m is the lowest level at which an examinee will be considered a master. Lord's approach is followed in the present sequential testing procedure.

The Use of Loss Functions

In making mastery decisions, two types of errors are possible: (1) classifying a nonmaster as a master (a false positive decision); and (2) classifying a master as a nonmaster (a false negative decision). Let α and β denote the probability of occurrence for these two different types of errors, respectively. In the procedures proposed by Ferguson (1969a, 1969b) and Reckase (1983), the decision rule is determined by specifying target values for α and β. In one of the fixed-length mastery tests proposed by Lord, the decision rule is determined by selecting a small value for α (e.g., .05) and then determining the decision rule that minimizes β.

The decision rule used here does not require the specification of values for α and β. Instead, misclassification rates are controlled through a decision theory approach. Early applications of decision theory to mastery testing include Cronbach and Gleser (1965), Hambleton and Novick (1973), Huynh

(1976), Petersen (1976), Swaminathan, Hambleton, and Algina (1975), and van der Linden and Mellenbergh (1977). In this approach, the user's preferences for alternative classification outcomes are established by assigning a real-valued number to each possible combination of mastery decision and true mastery status.

An example of a simple loss function defined for a fixed-length mastery test is given in Table 1. This loss function specifies that a correct mastery decision incurs no loss and an incorrect mastery decision incurs a loss that is equal to a real-valued constant (either A if the incorrect decision is a false positive, or B if the incorrect decision is a false negative). Equal values of A and B indicate that a false positive decision is just as undesirable as a false negative decision. When A is greater than B, the loss function incorporates the belief that false positives are more serious than false negatives.

Table 1
A Simple Loss Function Defined For a Fixed-Length Mastery Test

	True Mastery Status	
Decision	θ_n	θ_m
Pass	A	0
Fail	0	B

To define a loss function for a variable-length mastery test, it is necessary to specify the losses associated with each possible combination of mastery decision and true mastery status at each stage of testing. An example of a loss function for a test in which each examinee may be administered two testlets at most is given in Table 2. In this loss function, the value C represents the cost of administering a single testlet. As indicated in the table, this cost is incurred regardless of whether a correct or incorrect mastery decision is made, and the cost of administering the second testlet is assumed to be equal to the cost of administering the first testlet. Unique loss functions also can be defined to accommodate the demands of particular mastery testing applications. For example, the cost of testing, C, can vary with the number of testlets administered.

Table 2
A Simple Loss Function Defined for a Variable-Length Mastery Test With a Maximum of Two Testlets

	True Mastery Status	
Decision	θ_n	θ_m
Pass at Stage 1	$A + C$	C
Fail at Stage 1	C	$B + C$
Pass at Stage 2	$A + 2C$	$2C$
Fail at Stage 2	$2C$	$B + 2C$

Determining the Decision Rule

The object in adopting a decision theory approach is to determine a decision rule that reflects in some way the preferences for alternative outcomes built into the loss function. However, because the loss function depends on the true mastery status of individuals—and that status is never known in practice—the optimal decision rule to associate with a particular loss function will not be unique. Several methods for dealing with this problem are available. A Bayesian decision theory approach

is followed here, as is discussed in Chernoff and Moses (1959), Lindley (1971), and Wetherill (1975), among others. In this approach, the unique decision rule to associate with a particular loss function is found by minimizing posterior expected loss at each stage of testing.

The Bayesian solution described here is based on the simple loss function defined above, in which A represents the loss associated with a false positive decision, B represents the loss associated with a false negative decision, and both A and B are expressed in the same units as C, the cost of administering a single testlet. For the sake of simplicity, C is assumed to remain constant across stages.

Because this loss function considers only two levels of mastery (θ_n and θ_m), it also includes the implicit assumption that the loss of misclassification is the same, regardless of how incompetent (or competent) a particular examinee might be—that is, the loss of passing an examinee who is at θ_n is assumed to be the same as the loss of passing an examinee who is far below θ_n. Similarly, the loss of failing an examinee who is at θ_m is assumed to be the same as the loss of failing an examinee who is far above θ_m. In addition, this loss function also implies that there is no loss associated with misclassifying an examinee whose true θ lies in the neutral region between θ_n and θ_m. Thus, θ_n and θ_m (the θ levels associated with a maximally competent nonmaster and a minimally competent master), should be selected to be as close together as possible, given measurement constraints.

One of the advantages of selecting such a simplified loss function is that it limits the amount of prior information needed to determine posterior expected loss. In particular, prior beliefs about the true mastery status of examinees can be quantified in terms of two probabilities: P_m, the prior probability that an examinee is at θ_m, and $P_n = 1 - P_m$, the prior probability that an examinee is at θ_n. P_m can either be determined through a subjective assessment of the proportion of true masters in the examinee population, or through an analysis of empirical data providing the observed proportion of masters in the examinee population. Alternatively, $P_m = P_n = .5$ could incorporate a notion of equal prior odds.

To determine posterior expected loss, it is also necessary to have a model that characterizes the relationship between true mastery status and observed test performance. A three-parameter logistic IRT model was used here to provide the conditional probability of observing any particular pattern of ordered item responses, given true mastery status. The parameters of the model allow each item to be characterized in terms of its difficulty, discrimination, and guessing characteristics.

The Computerized Mastery Test

The Operational Format

The operational format of this approach is as follows. A pool of parallel testlets is developed and calibrated. At each stage of testing, an n-item testlet is randomly selected from the pool and administered. After responses to all n items have been observed, a decision is made either to classify the individual, or to administer another testlet when the examinee's cumulative number-correct score indicates an intermediate level of mastery and the number of testlets administered is less than some previously defined maximum. The decision to either classify or to continue testing is made by selecting the option that minimizes posterior expected loss. In determining the loss associated with the option to continue testing, all possible outcomes of future testlet administrations are considered. Thus, the action selected at each stage of testing is optimal with respect to the entire testing procedure.

The Decision Rule for a Fixed-Length Test

The decision rule derived minimizes posterior expected loss for a fixed-length mastery test consisting of a single n-item testlet; it is then generalized to a variable-length test. According to the loss

function defined above for a fixed-length mastery test, there are only two possible decisions—pass or fail—and two mastery states—θ_m and θ_n. The prior expected loss of a pass decision is calculated as:

$$E[\ell(\text{pass}|\Theta)] = \ell(\text{pass}|\theta_n) \cdot P_n + \ell(\text{pass}|\theta_m) \cdot P_m = A \cdot P_n \quad , \tag{1}$$

where P_m is the prior probability assumed for the master state (θ_m), $P_n = 1 - P_m$ is the prior probability assumed for the nonmaster state (θ_n), and $\ell(\cdot|\cdot)$ is the loss function. Similarly, the expected loss of a fail decision is given by

$$E[\ell(\text{fail}|\Theta)] = \ell(\text{fail}|\theta_n) \cdot P_n + \ell(\text{fail}|\theta_m) \cdot P_m = B \cdot P_m \quad . \tag{2}$$

The decision rule that minimizes expected prior loss is thus

$$d(P_m) = \text{pass, if } A \cdot P_n \leq B \cdot P_m$$
$$\text{fail, otherwise,} \tag{3}$$

which is equivalent to

$$d(P_m) = \text{pass, iff } P_m \geq A/(A + B) \quad . \tag{4}$$

Note that this decision rule does not make use of observed item responses.

The additional information about mastery status derived from examinees' observed item responses is incorporated into the decision rule by taking expectations with respect to the posterior distribution of Θ, rather than the prior distribution of Θ. The posterior distribution of Θ is obtained by conditioning on observed number-correct score, X, as

$$P(\Theta = \theta_m|X = s) = \frac{P(X = s|\Theta = \theta_m) \cdot P_m}{P(X = s|\Theta = \theta_m) \cdot P_m + P(X = s|\Theta = \theta_n) \cdot P_n} \equiv P_{m|x} \quad , \tag{5}$$

where, as before, P_m and P_n represent the prior probabilities for θ_m and θ_n, respectively, and the probability of observing any particular number-correct score X (on a single n-item testlet) is determined from the assumed IRT model, as

$$P(X = s|\Theta = \theta_m) = \sum \prod_{j=1}^{n} P_j(\theta_m)^{x_j} [1 - P_j(\theta_m)]^{1-x_j} \quad , \tag{6}$$

where the summation is taken over all response patterns such that the total score is s (for $s = 0, \ldots, n$), $x_j = 1$ or 0 (depending on whether the response pattern considered is defined with a correct or incorrect response to the jth item), and $P_j(\theta_m)$ is the conditional probability of a correct response to the jth item by an examinee with proficiency level θ_m (as given by the IRT model).

In a departure from some IRT-based Bayesian procedures, posterior probabilities are calculated here conditional on the observed number-correct score, rather than the entire vector of observed item responses. This simplification was adopted because of the branching nature of the sequential testing process. This simplification does not imply, though, that the number-correct score is sufficient for estimating Θ. It merely implies that the probabilities associated with response vectors having the same number-correct score can be meaningfully grouped together and treated as a unit, and it will result in a degradation of measurement accuracy only when the amount of information lost by combining probabilities within number-correct score groups is significant.

The posterior expected losses can now be calculated as

$$E[\ell(\text{pass}|\Theta)|X] = \ell(\text{pass}|\theta_n) \cdot P_{n|x} + \ell(\text{pass}|\theta_m) \cdot P_{m|x} = C + A \cdot P_{n|x} \tag{7}$$

and

$$E[\ell(\text{fail}|\Theta)|X] = \ell(\text{fail}|\theta_n) \cdot P_{n|x} + \ell(\text{fail}|\theta_m) \cdot P_{m|x} = C + B \cdot P_{m|x} \quad , \tag{8}$$

where the posterior probability of mastery $P_{m|x}$ is given in Equation 5 and $P_{n|x} = 1 - P_{m|x}$.

The decision rule that minimizes posterior expected loss is thus

$$d(P_{m|x}) = \begin{array}{l}\text{pass, if } A \cdot P_{n|x} \leq B \cdot P_{m|x} \\ \text{fail, otherwise,}\end{array} \tag{9}$$

which is equivalent to

$$d(P_{m|x}) = \text{pass, iff } P_{m|x} \geq A/(A + B) \quad . \tag{10}$$

This decision rule does not require a large number of on-line calculations. The pass/fail cutoff point $A/(A + B)$ can be calculated prior to test administration. Also, even though the measure of examinee performance $P_{m|x}$ is a function of data that will not be available until after the test has been administered, it will not be difficult to determine ahead of time which values of X will result in values of $P_{m|x}$ that are above the cutoff point and which will result in values that are below the cutoff point. Thus, on-line classification decisions can be made on the basis of observed number-correct scores.

The Assumption of Parallel Testlets

The testing procedure outlined above calls for administering a single n-item testlet to each examinee. In the more general case, testlets are randomly sampled from a pool of parallel testlets, and the number of testlets administered to each examinee is determined from his or her vector of observed item responses. In this context, "parallel testlets" refers to testlets that are both content-balanced and equivalent with respect to the likelihood of each possible number-correct score at the two points θ_m and θ_n.

Let X_i be the number-correct score observed for the ith testlet administered. For all pairs of testlets t and t', the parallel assumption implies that

$$P(X_i = s|\Theta = \theta_m, \text{ testlet } = t) = P(X_i = s|\Theta = \theta_m, \text{ testlet } = t') \tag{11}$$

and

$$P(X_i = s|\Theta = \theta_n, \text{ testlet } = t) = P(X_i = s|\Theta = \theta_n, \text{ testlet } = t') \tag{12}$$

for all possible scores ($s = 0, \ldots, n$). (A graphical procedure for evaluating the validity of the parallel testlet assumption is presented below.)

An in-depth treatment of the consequences of lack of parallelism in a testlet pool can be found in Sheehan and Lewis (1989). They concluded that between-testlet variation affects the efficiency of the testing procedure, but not the specification of an optimal decision rule. Thus, whether or not testlets are parallel, the best estimate of the likelihood of a particular number-correct score on a randomly selected testlet is the average of the probabilities calculated for that score on all testlets in the pool. As long as conditioning does not occur on the particular testlet administered, the likelihood of an observed number-correct score is the same, regardless of whether or not the testlets in the pool are truly parallel.

Although the decision rule proposed above is unaffected by lack of parallelism in an achieved testlet pool, there are still three reasons why the test designer should strive to make the testlets as parallel as possible. First, the efficiency of the test is inversely related to the level of between-testlet

variation, because the decision rule is based on the average likelihood of a number-correct score. As between-testlet variation increases, the precision of the average likelihood as an estimate of the likelihood for a specific testlet decreases, and efficiency thus decreases. Second, parallel testlet pools promote fairness, which is related to degree of parallelism, because each examinee may potentially be administered a different subset of testlets but the decision rule (as currently formulated) is the same for all testlets. Third, parallel testlets minimize the effect in this application of performing sampling without replacement.

The Decision Rule for a Variable-Length Test

A sequential testing design is now considered in which the decision to classify or continue testing is reevaluated after each testlet has been administered. Assume that a maximum test length of k testlets has been specified. As in the case of a single testlet, the decision rule at each stage of testing will require the posterior probability of mastery at that stage. To simplify the notation, let

$$P_{m|t} = P(\Theta = \theta_m | X_1, X_2, \ldots, X_t) \quad , \tag{13}$$

where X_t is the score observed for the ith testlet ($i = 1, \ldots, k$). This probability can be calculated iteratively, with

$$P_{m|t} = \frac{P(X_t|\Theta = \theta_m) \cdot P_{m|t-1}}{P(X_t|\Theta = \theta_m) \cdot P_{m|t-1} + P(X_t|\Theta = \theta_n) \cdot P_{n|t-1}} \quad , \tag{14}$$

where $P(X_t|\Theta = \theta_m)$ and $P(X_t|\Theta = \theta_n)$ refer to pool-wide average probabilities. Note that when $i = 1$, $P_{m|t-1}$ is the prior probability of mastery, P_m.

The expected losses associated with the decisions to pass or fail at stage i are expressed as functions of $P_{m|t}$:

$$\mathrm{E}[\ell(\mathrm{pass}|\Theta)|X_1, \ldots, X_t] = iC + A \cdot (1 - P_{m|t}) \tag{15}$$

and

$$\mathrm{E}[\ell(\mathrm{fail}|\Theta)|X_1, \ldots, X_t] = iC + B \cdot P_{m|t} \quad . \tag{16}$$

To determine the expected loss associated with the decision to administer another testlet at stage i (for $i < k$), it is necessary to consider all possible outcomes at stage $i + 1$. For example, if stage $i + 1$ were to result in a "pass immediately" decision, the loss of deciding to continue testing at stage i would be equal to the loss of deciding to pass the examinee at stage $i + 1$. However, because the set of all possible outcomes at stage $i + 1$ includes the option of administering another testlet, all possible outcomes at stage $i + 2$ must also be considered.

The uncertainty associated with future testlet administrations can be accounted for by averaging the expected loss associated with each of the various outcomes in proportion to the probability of observing those outcomes. In particular, the probability of observing each possible score X_{t+1} at stage $i + 1$, given the scores for the first i stages, can be calculated as a function of the posterior probability of mastery at stage i:

$$P(X_{t+1} = s|X_1, \ldots, X_t) = P(X_{t+1} = s|\Theta = \theta_n) \cdot P_{n|t} + P(X_{t+1} = s|\Theta = \theta_m) \cdot P_{m|t} \equiv P_{s|t} \quad , \tag{17}$$

where $P(X_{t+1} = s|\Theta = \theta_m)$ and $P(X_{t+1} = s|\Theta = \theta_n)$ again represent average pool-wide probabilities. This is called the predictive probability at stage i.

To determine the expected loss of the continue-testing option, it is useful to introduce risk functions at each stage of testing. Beginning with stage k, define

$$r_k(P_{m|k}) = \min\{\mathrm{E}[\ell(\text{pass}|\Theta)|X_1, \ldots, X_k], \mathrm{E}[\ell(\text{fail}|\Theta)|X_1, \ldots, X_k]\}$$
$$= \min[kC + A \cdot (1 - P_{m|k}), kC + B \cdot P_{m|k}] \quad . \tag{18}$$

The expected loss of deciding to administer another testlet at stage $k - 1$ can now be written in terms of the risk at stage k:

$$\mathrm{E}[\ell(\text{continue testing}|\Theta)|X_1, \ldots, X_{k-1}] = \sum_s P_{s|k-1}\, r_k(P_{m|k}) \quad , \tag{19}$$

where $P_{m|k}$ is evaluated for each value that X_k may take on. The risk function at stage $k - 1$ may now be defined as

$$r_{k-1}(P_{m|k-1}) = \min\{\mathrm{E}[\ell(\text{pass}|\Theta)|X_1, \ldots, X_{k-1}], \mathrm{E}[\ell(\text{fail}|\Theta)|X_1, \ldots, X_{k-1}],$$
$$\mathrm{E}[\ell(\text{continue testing}|\Theta)|X_1, \ldots, X_{k-1}]\}$$
$$= \min[(k-1)C + A \cdot (1 - P_{m|k-1}), (k-1)C + B \cdot P_{m|k-1}, \sum_s P_{s|k-1}\, r_k(P_{m|k})] \quad . \tag{20}$$

In general, the risk at stage i is defined in terms of the risk at stage $i + 1$:

$$r_i(P_{m|i}) = \min[iC + A(1 - P_{m|i}), iC + B \cdot P_{m|i}, \sum_s P_{s|i}\, r_{i+1}(P_{m|i+1})] \quad . \tag{21}$$

The decision rule that minimizes posterior expected loss at stage i for $i = 1, \ldots, k - 1$ can now be defined as

$$d_i(P_{m|i}) = \text{pass, if } r_i(P_{m|i}) = iC + A \cdot (1 - P_{m|i})$$
$$\text{fail, if } r_i(P_{m|i}) = iC + B \cdot P_{m|i}$$
$$\text{continue testing, otherwise;} \tag{22}$$

and for stage $i = k$,

$$d_k(P_{m|k}) = \text{pass, if } kC + A \cdot (1 - P_{m|k}) \leq kC + B \cdot P_{m|k}$$
$$\text{fail, otherwise;} \tag{23}$$

or equivalently,

$$d_k(P_{m|k}) = \text{pass, if } P_{m|k} \geq A/(A + B)$$
$$\text{fail, otherwise.} \tag{24}$$

This decision rule, which can be evaluated with respect to any pool of parallel testlets, provides the optimal decision to make at each stage of testing in a multistage test.

Rationale for Conditioning With Respect to Observed Number-Correct Score

Conditioning should be performed with respect to an individual's number-correct score, rather than their complete vector of observed item responses. Consider a test with a maximum of six stages and a testlet pool containing parallel 10-item testlets. To determine the expected loss associated with the continue-testing option at stage 1, all possible outcomes must be considered at successive stages of the test. When outcomes are grouped by number-correct score, the total number to consider is $11^5 = 161{,}051$ (11 corresponds to the set of all possible number-correct scores on a single testlet, and 5 corresponds to the possible additional testlets that could be administered). For each of these

outcomes, the probability that the outcome will occur (i.e., the predictive probability) must be calculated, as well as the loss that would be incurred if it did occur (i.e., the risk). When outcomes are not grouped by number-correct score, the total number possible is $(2^{10})^5 \approx 10^{15}$. Thus, the reduction in computation achieved by grouping outcomes by number-correct score is significant.

Determining Probability Metric Threshold Values

In the decision rule given above, the number of required on-line calculations increases with the number of stages and the size of the testlet. It is conceivable that in some testing situations the number of required on-line calculations will be more than the test administration computers can handle without noticeable delays between stages. One alternative testing procedure is to perform some of the on-line calculations off-line before the test is administered. It is expected that this procedure will be applicable to tests of any size, and that the testing computers will not require more than a minimum level of sophistication.

To determine the threshold values, first the risk is computed at stage k for a set of values of $P_{m|k}$ (e.g., 0, .001, . . . , 1). For each value of $P_{m|k}$, the decision is selected that minimizes expected posterior loss by applying Equation 24. Next, the risk is computed at stage $k - 1$ for the same set of values of $P_{m|k-1}$. The task of computing the risk at stage $k - 1$ is simplified considerably by the fact that the risk at stage k has already been determined. At each stage $i < k$, the largest values of $P_{m|i}$ will result in the decision "pass immediately," and the smallest values will result in the decision "fail immediately." Thus the threshold values are defined as

λ_{1i} = largest value such that application of Equation 22 results in the decision to fail immediately whenever $P_{m|i} < \lambda_{1i}$,

λ_{2i} = smallest value such that application of Equation 22 results in the decision to pass immediately whenever $P_{m|i} \geq \lambda_{2i}$.

(Note that application of Equation 22 will result in the decision to continue testing whenever $\lambda_{1i} \leq P_{m|i} \leq \lambda_{2i}$.) It is easily seen that $\lambda_{1k} = \lambda_{2k} = A/(A + B)$.

The procedure outlined above can be used to determine the threshold values λ_{1i} and λ_{2i}, for $i = 1, \ldots, k$, prior to an actual test administration. Given that the threshold values have already been determined, the testing procedure at each stage reduces to three steps: (1) administer a randomly selected testlet and observe the number-correct score X_i; (2) update the examinee's estimate of $P_{m|i}$ using Equation 14; and (3) make a decision by comparing the updated estimate of $P_{m|i}$ to a stored threshold value. Use of this alternative procedure vastly reduces the number of required on-line calculations because it is no longer necessary to calculate the risk at future stages of the test.

Translating Probability Metric Threshold Values to the Number-Correct Score Metric

In deriving the decision rule for the fixed-length test, the number of on-line calculations was reduced by translating the threshold value $\lambda_{1i} = \lambda_{2i} = A/(A + B)$ from the probability metric to the number-correct score metric. A similar translation can be performed in the case of a variable-length test—that is, by computing the posterior probabilities corresponding to all possible combinations of $X_1, \ldots, X_k$, a set of approximate threshold values (Y_{1i}, Y_{2i}, $i = 1, \ldots, k$) can be determined, such that

$\Sigma X_j < Y_{1i}$ for most combinations in which $P_{m|i} < \lambda_{1i}$, and

$\Sigma X_j \geq Y_{2i}$ for most combinations in which $P_{m|i} \geq \lambda_{2i}$ for $i = 1, \ldots, k$,

where the summation is over X_j for $j = 1, \ldots, i$. The translation from the posterior probability metric to the number-correct score metric is a many-to-one transformation (i.e., many values of the

posterior probability statistic are mapped into a single number-correct score). Thus, in many applications, on-line classification decisions can be made on the basis of cumulative number-correct scores.

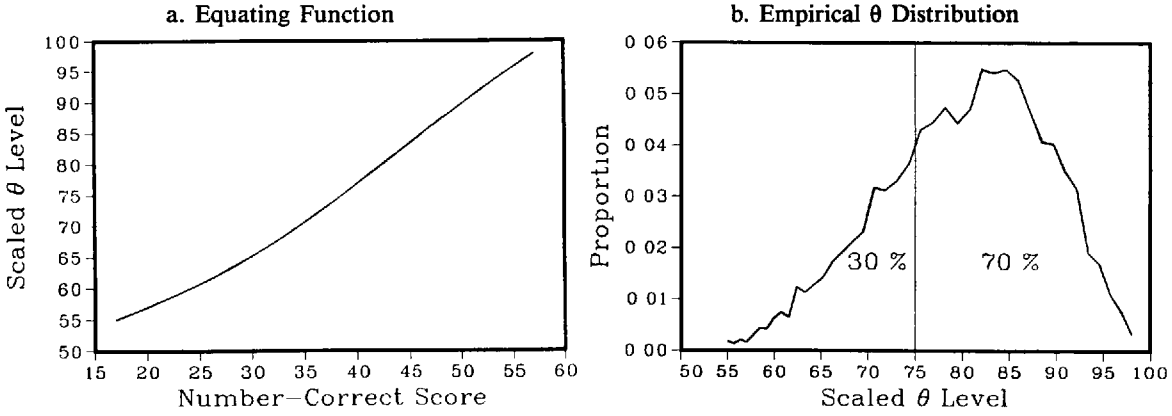
An Application

Data

The procedure described above was used to design a variable-length mastery test for use in a professional certification program. The data available for this effort consisted of responses collected in several past administrations of a fixed-length 60-item mastery test that was administered in the standard paper-and-pencil format. The results of this test were reported to examinees on a scale that ranged from 0 to 100 with a cutscore of 75. The cutscore had been established using the procedure described in Ebel (1972).

Some results of a typical paper-and-pencil administration of this exam are presented in Figure 1. Figure 1a provides an equating function that can be used to translate scores from the number-correct metric to the reporting scale metric. The reported cutscore of 75 corresponds to a number-correct score of 39. Figure 1b provides the distribution of estimated abilities, which shows that approximately 70% of the examinee population received a passing score.

Figure 1
An Equating Function and an Empirical θ Distribution Estimated From a Paper-and-Pencil Administration of the Fixed-Length 60-Item Form (N = 4,280)



Repeated paper-and-pencil administrations of this exam previously resulted in an item pool containing 110 items. The three-parameter logistic IRT model was fit to these data, and estimated model parameters were available (Kingston, 1987). The estimated parameters indicated a wide variation in the difficulty and discrimination levels of the items in the pool. The items also belonged to two non-overlapping content categories.

Testlet Construction

Additional constraints imposed on this test development effort included the following: (1) the test length was limited to between 20 and 60 items, and (2) the test content was constrained to include the two content categories in a 60/40 ratio. A testlet length of 10 items was selected, and each examinee would be required to respond to at least two but no more than six testlets.

Because the item pool contained only 110 items, Lord's procedure (1980) was not used to construct

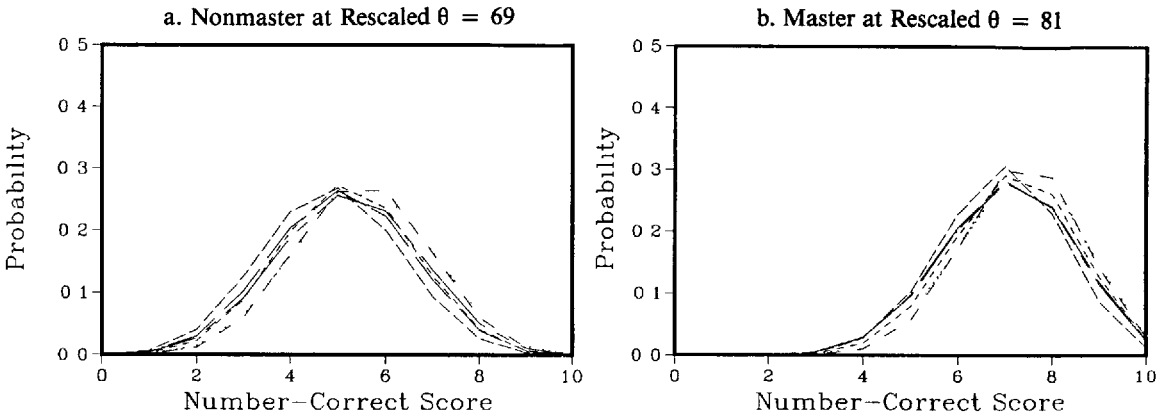
testlets. Instead, an alternative procedure was developed that retained some elements of Lord's procedure. First, 11 testlets were constructed by cross-classifying the item pool according to content category and estimated item difficulty, and then sequentially assigning items to testlets. Each resulting testlet contained six items from the first content category and four items from the second content category. Second, the six testlets that appeared most "parallel" (in terms of median difficulty and discrimination) were then selected from the 11 that were available. The estimated item parameters of these six selected testlets are summarized in Table 3.

Table 3
Minimum, Median, and Maximum of Item Parameter Estimates for Six Testlets Ordered by Median Difficulty

Number	Minimum	Median	Maximum
Item Difficulty ($\hat{b}$)			
1	-1.789	-.965	1.016
2	-2.074	-.501	.945
3	-1.992	-.490	4.701
4	-2.678	-.409	.640
5	-1.701	-.111	.547
6	-2.364	-.108	2.132
Item Discrimination ($\hat{a}$)			
1	.633	.862	1.469
2	.200	.768	1.333
3	.168	.760	1.627
4	.196	.621	1.014
5	.321	.666	1.477
6	.446	.728	1.389
Lower Asymptote ($\hat{c}$)			
1	.000	.090	.341
2	.000	.179	.500
3	.000	.166	.326
4	.000	.166	.335
5	.000	.187	.348
6	.000	.205	.285

As a final step, two checks were performed. First, each testlet was evaluated for unwanted item dependencies, and several offending items were replaced with alternative items that had been matched for content category and difficulty. Second, the validity of the testlet interchangeability assumption was evaluated by comparing the theoretical distribution of number-correct scores estimated for each testlet at the previously-selected points θ_n and θ_m. These distributions are plotted in Figure 2. The points θ_n and θ_m correspond to ability levels of 69 (Figure 2a) and 81 (Figure 2b) on the test reporting scale, respectively. The closeness of the curves indicates that for examinees near the cutscore, the probability of observing a particular number-correct score is virtually the same regardless of the particular testlet administered. Based on this comparison, it was decided that the measurement accuracy of the test would not be seriously degraded by treating the six selected testlets as if they were truly interchangeable. (This procedure is not presented as an optimal method for constructing parallel testlets; rather, it a method that can be used when the test designer must contend with a small item pool. More appropriate methods would include a facility for matching test information curves, and would provide a pool of at least 10 testlets.)

Figure 2
Theoretically-Derived Number-Correct Score Distributions
for the Six Testlets (Different Line Types Represent Different Testlets)



Choice of Decision Rule

To determine the decision rule that minimizes posterior expected loss for this particular pool of six testlets, four additional parameters had to be specified: the prior probability of mastery P_m, and the loss function parameters A, B, and C. Although the proportion of true masters in the population was expected to be near .7, the value $P_m = .5$ was selected in order to incorporate a notion of equal prior odds.

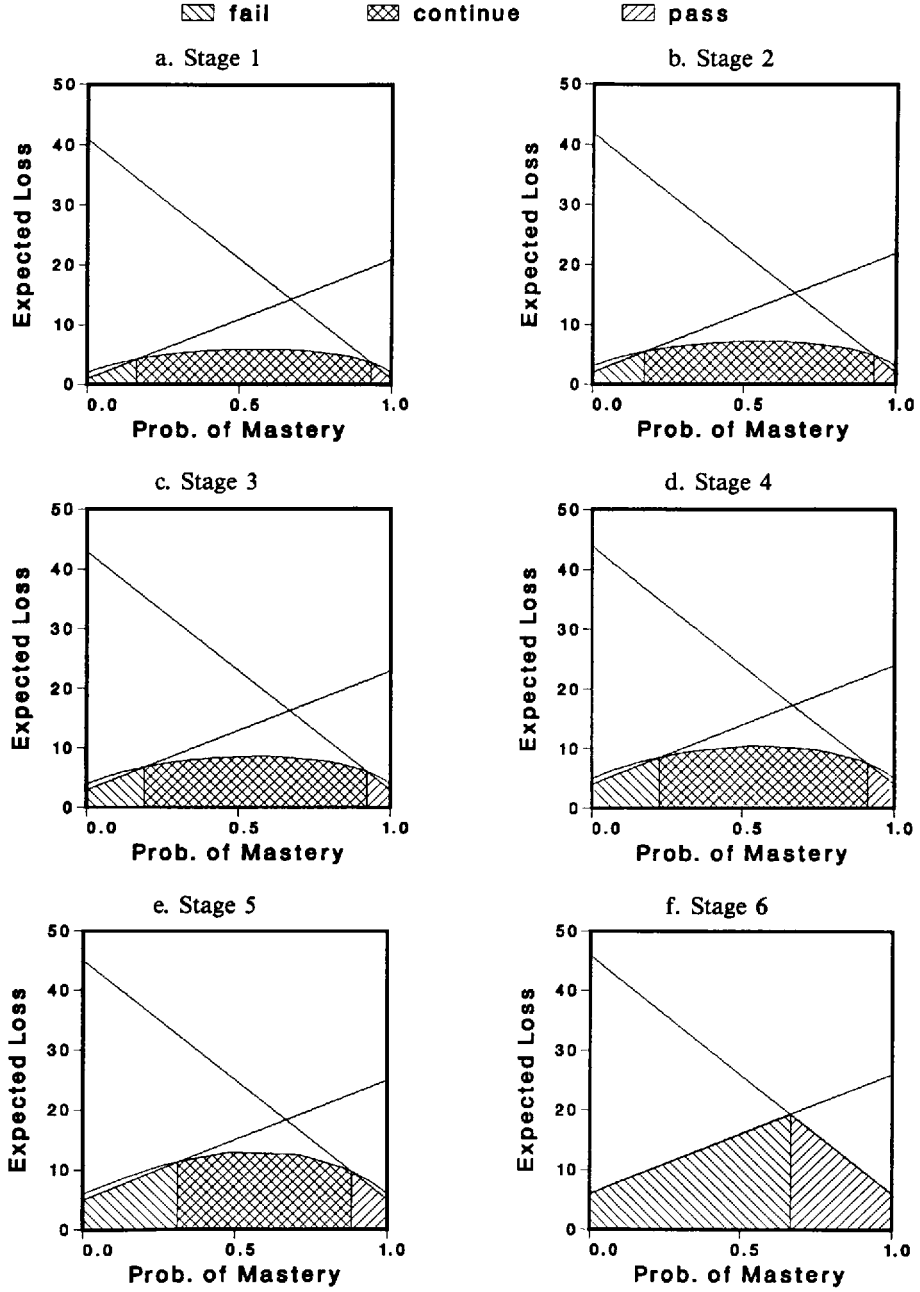
The loss function parameters, A, B, and C were selected as follows: (1) to set the scale of measurement, the cost of administering a testlet (C) was set equal to 1; (2) to incorporate the belief that a false positive decision was twice as serious as a false negative decision, A was set to $2B$; and (3) a simulation study was performed to evaluate the operating characteristics of the alternative decision rules that resulted when B was allowed to vary between 2 and 100. The operating characteristics investigated in the simulation included average test length, expected passing rate, expected proportion of false positive decisions, and expected proportion of false negative decisions.

Based on the simulation results, it was determined that the value $B = 20$ provided a decision rule with desirable operating characteristics. Thus, the loss of passing a nonmaster was taken to be 40, and the loss of failing a master was taken to be 20, on a scale in which one unit corresponded to the cost of administering a single testlet. This is referred to as a 40/20 loss function.

Figure 3 gives a stage-by-stage view of the expected losses associated with the 40/20 loss function, as applied to the six selected testlets. In each plot, the posterior probability of mastery $P_{m|t}$ is plotted along the X axis, and the posterior expected loss curves calculated for each possible decision are plotted along the Y axis. The plot for stage 6 (Figure 3f) is the final stage of the test, so only two decisions are possible: pass or fail. The expected loss curve for the pass decision decreases linearly as the posterior probability of mastery increases. The expected loss curve for the fail decision increases linearly as the posterior probability of mastery increases. The point at which the two curves intersect is the threshold value $\lambda_{16} = \lambda_{26} = A/(A + B) = 2/3$. Expected loss is minimized below this point by making a fail decision, and above this point by making a pass decision. The region in which a fail decision is optimal is indicated by negatively sloped diagonal lines, whereas positively sloped diagonal lines indicate the region in which a pass decision is optimal.

Figures 3a through 3e include a third curve representing the expected loss of the continue-testing

Figure 3
Posterior Expected Loss for the 40/20 Decision Rule



option. The area where this curve provides the minimum expected loss is shaded using a cross-hatched pattern. For stage 5 (Figure 3e), expected loss is minimized by making a fail decision for low values of $P_{m|s}$, by making a pass decision for high values, and by deciding to administer a sixth testlet for

intermediate values. Figures 3a through 3f show that the range of values of $P_{m|t}$ that will result in the decision to continue testing decreases with the number of testlets administered. Thus, an examinee's chances of being classified increase with each testlet administered.

The threshold values obtained for the 40/20 loss function are listed in Table 4. For purposes of comparison, the table also lists the threshold values that would have resulted if equal losses ($A = B = 20$) had been specified. Values are reported both in the probability metric and the cumulative number-correct score metric. The probability metric threshold values correspond to the intersection points of the curves in Figure 3. The cumulative number-correct score values were obtained by computing the posterior probabilities corresponding to all possible combinations of number-correct scores. Although either one of these sets of values could be used in an actual on-line testing situation, the cumulative number-correct score threshold values are the most likely choice because they are much simpler to understand and easier to use.

Table 4
Probability Metric and Number-Correct Score Threshold Values for 20/20 and 40/20 Loss Functions

		Loss = 20/20		Loss = 40/20	
Stage	Items	Lower Value	Upper Value	Lower Value	Upper Value
Probability Metric Threshold Values					
1	10	.1525	.8525	.1675	.9275
2	20	.1525	.8475	.1775	.9275
3	30	.1625	.8375	.1975	.9225
4	40	.1875	.8225	.2325	.9125
5	50	.2325	.7725	.3175	.8825
6	60	.5000	.5000	.6666	.6666
Number-Correct Score Threshold Values					
1	10	-	-	-	-
2	20	11	15	11	16
3	30	17	21	17	22
4	40	23	27	23	28
5	50	30	33	30	34
6	60	37	38	38	39

The probability metric threshold values are the cutoffs that are to be applied to each examinee's updated posterior probability of mastery $P_{m|t}$ at the completion of each additional testlet. Under the 20/20 rule, examinees with posterior probabilities below .1525 at stage 1 will be failed immediately, examinees with posterior probabilities of .8525 or greater will be passed immediately, and examinees with intermediate values will be required to respond to an additional testlet. The corresponding values for the 40/20 rule are shifted slightly upward. For example, in order to be passed after responding to just one testlet under the 40/20 rule, examinees must have a posterior probability of mastery that meets or exceeds the higher cutoff value of .9275. This more stringent requirement reflects the asymmetric nature of the 40/20 loss function.

Table 4 does not list number-correct score threshold values for stage 1. This is because a minimum test length of two testlets had previously been established, and this restriction was incorporated into all the decision rules considered by changing stage 1 probability metric threshold values to 0 and 1, respectively. Because it is impossible for an examinee to respond in such a manner as to achieve a posterior probability of mastery less than 0 or greater than 1, all examinees are required to respond

to at least two testlets, regardless of their score on the first testlet.

Table 4 shows that by stage 2, under the 20/20 rule, examinees with cumulative number-correct scores of 11 or less will be failed, and those with cumulative number-correct scores of 15 or higher will be passed; those with scores of 16 or higher will be passed under the 40/20 rule. In order to determine the cutoff scores for stage 3, not all of the $11^3 = 1{,}331$ possible score combinations need be considered, so the number of required calculations remains manageable. Also, this method of determining a decision rule is particularly attractive because these calculations can be performed in advance of an actual on-line testing session.

Using Simulation Techniques to Select a Loss Function

Simulation Design

A number of alternative loss functions were evaluated before selecting the 40/20 loss function. Each alternative function was evaluated with respect to a single set of simulated data. The simulated dataset included item responses generated according to the three-parameter logistic IRT model for each item in the six selected testlets (a total of 60 items). Data were simulated for 100 examinees at each of 41 levels ranging from 55 to 98 on the reported score metric (total $N = 4{,}100$). The score levels used in the simulation were selected to be representative of the range of abilities observed in paper-and-pencil administrations of the items included in the six selected testlets (Stocking, 1987).

Although many different loss functions were evaluated, the three reported here include the 20/20, the 40/20, and the 100/20 loss functions. For purposes of comparison, these loss functions were evaluated with respect to a variable- and a fixed-length test. The variable-length test was defined such that each examinee was required to respond to at least two testlets and no more than six testlets; therefore, test lengths ranged from 20 to 60 items. The fixed-length test was defined to include the same six testlets used to construct the variable-length test, for a total length of 60 items.

For each test, Table 5 provides the average test length, the expected pass rate, and the expected error rates. These statistics were obtained by weighting the simulation results to reflect the expected proportion of examinees at each of the 41 score levels considered, based on the distribution given in Figure 1. The error rates given as a percent of the total population provide the number of incorrect decisions (false positives or false negatives) expressed as a percent of the total decisions made. The error rates given as a percent of a subpopulation provide the percent of nonmasters misclassified as masters (false positives), and the percent of masters misclassified as nonmasters (false negatives). The approximate percentages of masters and nonmasters in the examinee population were 70% and 30%, respectively.

The results presented for the three variable-length tests show that, as expected, more severe losses led to longer tests. For example, when the loss of incorrectly passing a nonmaster is considered to be 20 times the cost of administering another testlet, on the average examinees will be required to respond to about 25 items. When this loss is considered to be 100 times the cost of administering another testlet, an average test length of about 30 items can be expected.

The error rates listed in Table 5 were obtained by comparing the true mastery status of each simulee to the mastery classifications that resulted from applying the various decision rules. The error rates listed for the fixed-length tests are those expected for a standard paper-and-pencil test developed from the same item pool as the variable-length tests and scored with a similar loss function. For each loss function considered, the variable-length testing procedure achieved similar decision accuracy as the fixed-length testing procedure, but used fewer items. For example, under the 100/20 decision rule, the average test length was decreased by half, but there was only a slight change in the expected error

Table 5
Average Test Length, Pass Rate, and Error Rates as a Percent of Total Population and Subpopulation for Variable- and Fixed-Length Mastery Tests and Three Decision Rules

Decision Rule	Average Test Length	Pass Rate	Total Population		Subpopulation	
			False Positive	False Negative	False Positive	False Negative
Variable-Length Tests						
20/20	25.2	65.5	3.3	7.5	11.0	10.8
40/20	27.7	64.0	2.6	8.2	8.6	11.8
100/20	30.2	60.5	1.7	10.8	5.6	15.5
Fixed-Length Tests						
20/20	60	69.0	3.7	4.3	12.3	6.2
40/20	60	65.8	2.7	6.6	8.8	9.4
100/20	60	62.1	1.8	9.3	6.0	13.4

rates. Thus, on a per-item basis, the decision accuracy of the variable-length test exceeded that of the fixed-length test for each loss function considered.

The adaptive nature of the 40/20 CMT is illustrated in Figure 4, which provides the bivariate distribution of true θ and test length. The plot shows that examinees with true θ levels located near the cutscore of 75 will be administered tests of 50 or 60 items, whereas those with θ levels at either of the two extremes will be administered a test of 20 or 30 items, at most.

An alternative view of the operating characteristics of these three loss function specifications is provided in Figure 5. The percent of simulees classified as masters is plotted as a function of θ. Results for the three variable-length tests are given in Figure 5a, and results for the three fixed-length tests are given in Figure 5b. All curves were obtained by smoothing the unweighted simulation results.

Figure 4
Bivariate Distribution of θ and Test Length for a Variable-Length CMT With a 40/20 Decision Rule

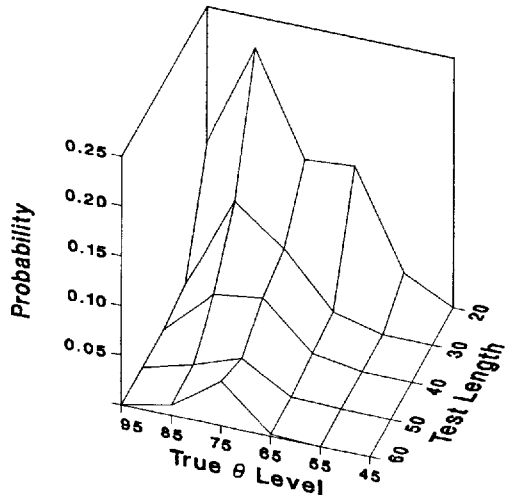
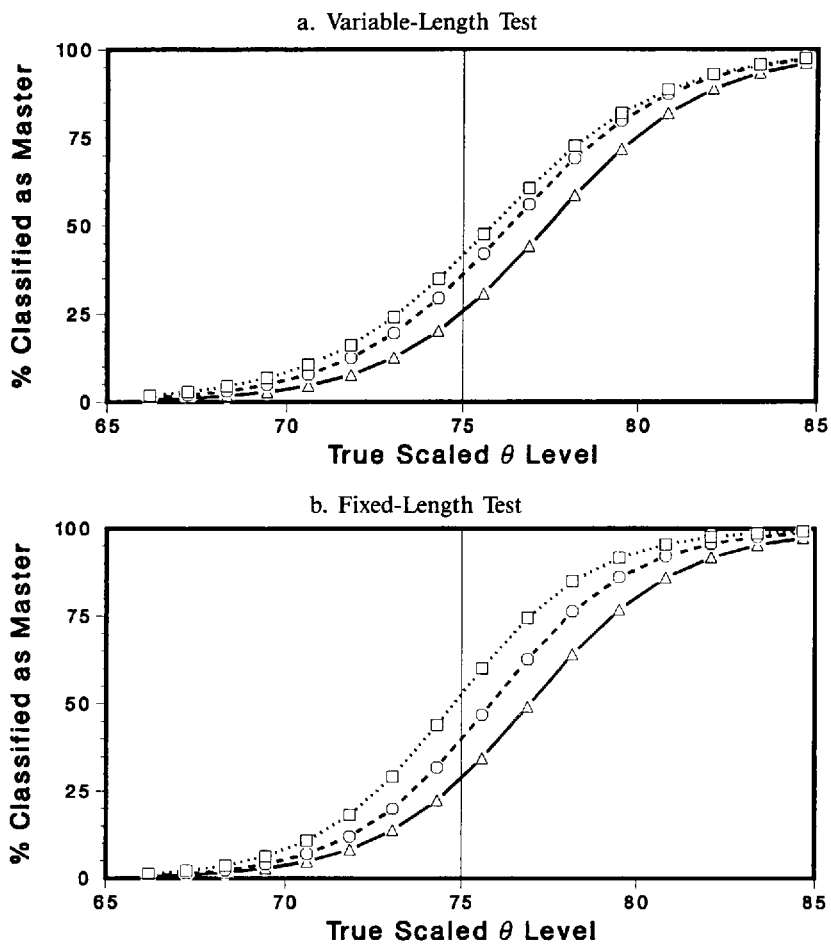


Figure 5
Percent Classified as Master Under Three Alternative Decision Rules



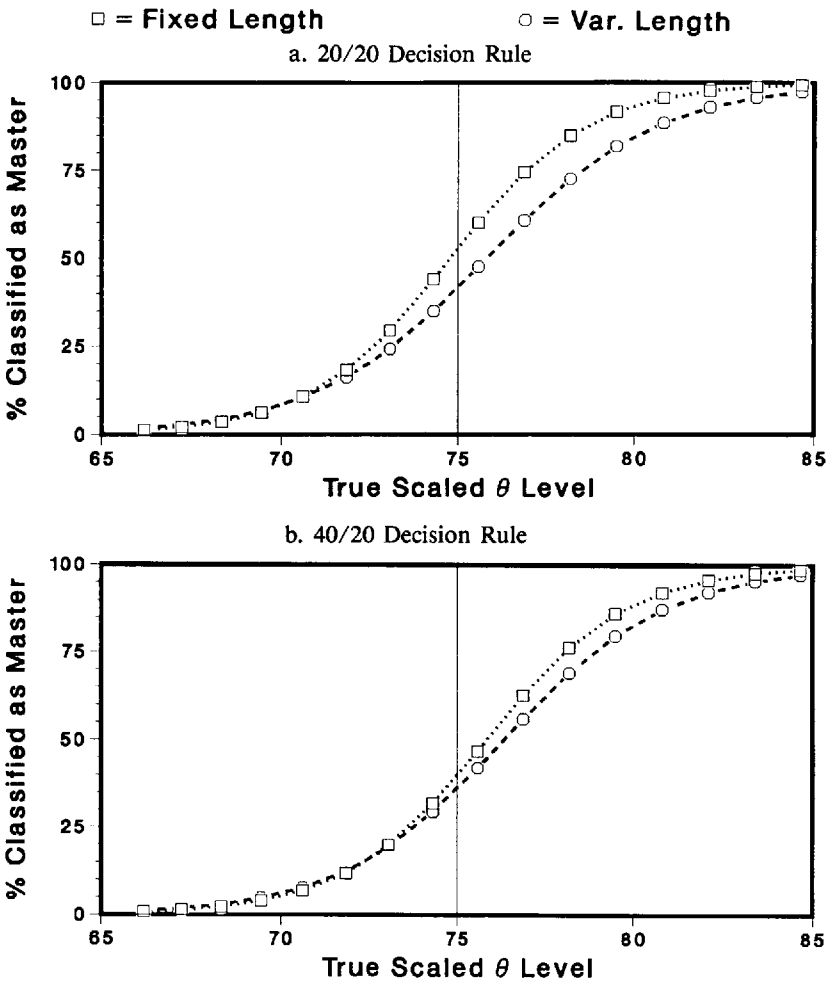
The plots show that the more extreme loss function specifications tend to result in fewer classification errors for both modes of administration.

The trade-off between classification accuracy and test length is illustrated in Figure 6. In Figure 6a, the results of applying the 20/20 decision rule under a fixed-length testing format (i.e., all examinees respond to 60 items) are compared to the results of applying the same decision rule in a variable-length testing format (i.e., test lengths from 20 to 60 items). Figure 6b provides a similar comparison for the 40/20 decision rule. The plots show that the shorter average test lengths associated with the variable-length testing format for these decision rules are accompanied by observable decreases in classification accuracy.

Discussion

The sequential mastery testing procedure described here provides a theoretical framework for balancing the competing goals of classification accuracy and test efficiency. Implementation of this ap-

Figure 6
Percent Classified as Master as a Function of True θ Level
for Fixed-Length and Variable-Length Tests



proach depends on the availability of (1) a computerized test delivery system, (2) a pool of pretested items, and (3) a model relating observed test performance to true mastery status. Although the procedure was developed using an IRT model, alternative models may also prove useful.

This approach to sequential mastery testing incorporates three simplifications: (1) all examinees are assumed to be at one of two ability levels, θ_n or θ_m; (2) conditioning is performed with respect to observed number-correct scores, rather than the entire vector of observed item responses; and (3) posterior distributions are estimated using pool-wide average likelihood functions, rather than testlet-specific likelihood functions. These simplifications, however, were not incorporated into the simulated data that were used to determine the operating characteristics of alternative decision rules. Instead, the simulated data assumed the range of θs given in Figure 1, and responses were generated according to the three-parameter logistic IRT model, which allows for between-testlet variation as well as variation in the likelihoods of response vectors having the same number-correct score.

Thus, the reasonable error rates obtained in the simulation can be interpreted as evidence that the simplifying assumptions have not seriously degraded the measurement accuracy of the test.

References

Chernoff, H., & Moses, L. E. (1959). *Elementary decision theory.* New York: Wiley.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana IL: University of Illinois Press.

Ebel, R. L. (1972). *Essentials of educational measurement.* Englewood Cliffs NJ: Prentice Hall.

Ferguson, R. L. (1969a). *Computer-assisted criterion-referenced measurement* (Working Paper No. 41). Pittsburgh PA: University of Pittsburgh Learning and Research Development Center. (ERIC Documentation Reproduction No. ED 037 089)

Ferguson, R. L. (1969b). *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.

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

Huynh, H. (1976). Statistical consideration of mastery scores. *Psychometrika, 41,* 65–78.

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.

Kingston, N. M. (1987). *Feasibility of using IRT-based methods for divisions D, E, and I of the Architect Registration Examination.* Unpublished manuscript, Educational Testing Service, Princeton NJ.

Lindley, D. V. (1971). *Making decisions.* London and New York: Wiley-Interscience.

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

Petersen, N. S. (1976). An expected utility model for "optimal" selection. *Journal of Educational Statistics, 1,* 333–358.

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. 237–255). New York: Academic Press.

Sheehan, K., & Lewis, C. (1989) *Computerized mastery testing with nonequivalent testlets.* Unpublished manuscript, Educational Testing Service, Princeton NJ.

Stocking, M. L. (1987). *NCARB: Mastery test design project.* Unpublished manuscript. Educational Testing Service, Princeton NJ.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1975). A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement, 12,* 87–98.

van der Linden, W. J., & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. *Applied Psychological Measurement, 7,* 593–599.

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

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

Wetherill, G. B. (1975). *Sequential methods in statistics.* London: Chapman & Hall.

Author's Address

Send requests for reprints or further information to Charles Lewis, Educational Testing Service, Princeton NJ 08541, U.S.A.