PSYCHOMETRIKA—VOL. 70, NO. 4, 685–713
DECEMBER 2005
DOI: 10.1007/S11336-005-1140-9

APPLICATION OF SEQUENTIAL INTERVAL ESTIMATION TO ADAPTIVE MASTERY TESTING

YUAN-CHIN IVAN CHANG

INSTITUTE OF STATISTICAL SCIENCE, ACADEMIA SINICA, TAIPEI, TAIWAN

In this paper, we apply sequential one-sided confidence interval estimation procedures with β-protection to adaptive mastery testing. The procedures of fixed-width and fixed proportional accuracy confidence interval estimation can be viewed as extensions of one-sided confidence interval procedures. It can be shown that the adaptive mastery testing procedure based on a one-sided confidence interval with β-protection is more efficient in terms of test length than a testing procedure based on a two-sided/fixed-width confidence interval. Some simulation studies applying the one-sided confidence interval procedure and its extensions mentioned above to adaptive mastery testing are conducted. For the purpose of comparison, we also have a numerical study of adaptive mastery testing based on Wald's sequential probability ratio test. The comparison of their performances is based on the correct classification probability, averages of test length, as well as the width of the "indifference regions." From these empirical results, we found that applying the one-sided confidence interval procedure to adaptive mastery testing is very promising.

Key words: adaptive mastery testing, confidence interval, β-protection, fixed-width confidence interval, fixed proportional accuracy, SPRT, indifference region.

1. Introduction

Mastery testing technology can be applied to many applications that require decisions about whether a person is above or below a criterion score. To conduct mastery testing, a threshold must be set in advance. Once this prescribed cutoff levelis chosen, the test-takers will be classified into one of two categories—pass or fail, according to their responses to the test items. In this paper, we consider applying sequential interval estimation methods to adaptive mastery testing (AMT) (Kingsbury & Weiss, 1983), a type of variable-length mastery testing. Another kind of variable-length mastery testing commonly discussed in the literature is sequential mastery testing (SMT), which uses sequential Bayesian decision theory or Wald's sequential probability ratio test, and does not concern item selection (Epstein & Knerr, 1977; Glas & Vos, 1998).

The major differences between AMT and SMT are their item selection rules. This can be seen, for example, in the applications of Wald's (1947) Sequential Probability Ratio Test (SPRT) to mastery testing (Kingsbury & Weiss, 1983; Reckase, 1983). Wald's (1947) SPRT was originally built for a statistical hypotheses testing problem where both null and alternative hypotheses are simple. If we apply SPRT to mastery testing with a random item selection rule, then it can be viewed as an SMT. However, if the item selection is tailored to the test-taker's estimated proficiency level, then it is a type of AMT (Ferguson, 1969; Lord, 1971; Kingsbury & Weiss, 1983; Reckase, 1983; Spray & Reckase, 1996; Chang, 2003).

Some researchers have already proposed applying two-sided confidence interval estimation procedures to mastery testing (Kingsbury & Weiss, 1983). In this case, the classification probability is related to the coverage probability of the confidence interval. In addition to the coverage probability of the confidence interval, the width of the confidence interval is also an important criterion, which can serve as a precision measure of the two-sided confidence intervals. If our goal is to estimate test-takers' proficiency levels as in computerized adaptive testing (CAT), then

Requests for reprints should be sent to Yuan-chin Ivan Chang, Institute of Statistical Science, Academia Sinica, Taipei, Taiwan. E-mail: ycchang@sinica.edu.tw.

the width of the confidence interval can also serve as a precision measure here. Therefore, two-sided confidence interval estimation is an appropriate statistical tool for this purpose (Chang & Ying, 2003). However, if our purpose is to classify test-takers into one of two categories (pass or fail/mastery or nonmastery), then intuitively the one-sided confidence interval estimation procedure should be used instead. For more detailed discussion of mastery testing and CAT, we refer readers to Lord (1980) and Wainer (2000).

If the one-sided confidence interval is used, then the length of the interval is no longer an appropriate measure of precision because the length of any one-sided confidence interval is always infinite. To overcome this difficulty, Wijsman (1981, 1982) introduces the idea of β-protection, a measure of precision of one-sided confidence intervals, which measures the probability of a confidence interval covering the "wrong" parameters. Based on this idea, Wijsman (1981, 1982) proposes a sequential procedure for constructing a one-sided confidence interval with the prescribed coverage probability and β-protection, which allows us to specify the coverage probability and the probability of covering "wrong parameters," simultaneously. When applying a one-sided confidence interval to mastery testing, this property allows test administrators to specify the upper bounds of the misclassification probabilities of false pass and false fail, separately. Some nontrivial modifications are needed to adapt it to AMT.

When applying SPRT to mastery testing, a region around the cutoff level must be specified for which it does not matter whether a pass or a fail decision is made. This region is usually called the "indifference region" (Reckase, 1983; Spray, 1993). The closer a test-taker's proficiency level is to the cutoff level, the more items will be required to make a correct decision. Similarly, when interval estimation procedures are applied to mastery testing in practice, an "indifference region" must be enforced in order to prevent the item pool from being exhausted.

Thus, comparison of the performances of classification procedures will be based on test-length, probability of correct classification, as well as width of indifference regions. In this paper, we prove that AMT based on the one-sided confidence interval estimation procedure is more efficient than those based on the fixed-width confidence interval procedure in terms of test length, under the condition that the prescribed classification probabilities and the width of the indifference regions are all the same. In fact, it can be shown that the average test length of the one-sided confidence interval procedure is only one-fourth that of the fixed-width confidence interval procedure. These results are confirmed by simulation studies.

Since the logistic model is one of the most popular statistical models used in the Item Response Theory (IRT) of educational testing, this paper will focus on AMT using the logistic model-based IRT model. Moreover, from either a theoretical or a technical point of view, when the items are randomly selected from an item bank, SMT can be treated as a special case of AMT. Thus, the results in this paper can be applied to SMT as well.

The rest of this paper is organized in the following way. In the next two sections, we first review the sequential interval estimation procedure and then AMT using the logistic model-based IRT model. We then describe the procedures for applying one-sided confidence intervals to AMT and introduce some of the large sample results. Procedures for using two-sided confidence intervals and fixed proportional accuracy confidence intervals are also presented in this section. Simulation studies and discussions are summarized in section 3. Concluding remarks and technical details are given in sections 4 and 5, respectively.

2. Sequential Estimate of Confidence Intervals

The two most popular criteria to measure the quality of confidence intervals are: (1) coverage probability and (2) precision. The precision of a confidence interval is as important as its coverage probability. This is illustrated by the fact that the interval $(-\infty, \infty)$ is a trivial confidence interval.

It will always have 100% coverage probability for any real parameter, but it is of no use in providing information on any parameter of interest.

Moreover, measures of precision may vary with different types of confidence sets. For example, it is not appropriate to use lengths of one-sided confidence intervals as their measure of precision, since they are infinite. Below, we will use an example to introduce the sequential interval estimation procedure and to explain why it is needed.

Before we go into the details of sequential interval analysis, it should be noted that the sample size here is equivalent to the test length of CAT/AMT, when we apply these confidence interval estimation procedures to CAT/AMT. Thus, the procedure with the smallest sample size is equivalent to the procedure with the smallest number of test items.

2.1. *Fixed-Width Confidence Intervals*

Suppose we want to construct a two-sided confidence interval for a parameter of interest. In CAT, for example, this might be the ability level of a test-taker. It is clear that under the same coverage probability, the shorter the length of the confidence interval, the more precise the measure of the ability level of the test-taker. Thus, for a given (fixed) sample size (number of test items) and the prescribed coverage probability, the two-sided confidence interval with the shortest length will be the best. The length of confidence intervals is certainly an appropriate measure of precision in this case.

On the other hand, if coverage probability and precision are fixed first, then the procedure that requires the smallest sample size (number of test items) to construct a confidence interval with the prescribed coverage probability and precision will be the most efficient one.

The following example explains the need for the sequential procedure for constructing a fixed width confidence interval with predefined coverage probability.

Example 1. Suppose that $x_1, x_2, \ldots$ are independent and identically distributed (i.i.d.) random variables with mean μ and variance $\sigma^2 < \infty$. It is known that the sample mean $\bar{x}_n = \sum_1^n x_i/n$ is a strong consistent estimate of μ and is asymptotic normal; i.e., as $n \to \infty$, $\bar{x}_n \to \mu$ a.s. and $\sqrt{n}(\bar{x}_n - \mu) \to_L N(0, \sigma^2)$. Let $CI_n = [\bar{x}_n \pm z_{\alpha/2}/(\sigma\sqrt{n})]$ be a confidence interval for μ. It follows from the asymptotic normality of $\bar{x}_n$, that the coverage probability of CI_n equals $(1-\alpha)$, asymptotically. (The notation $z_t \in \mathcal{R}$, for $t \in (0,\ 1)$, denotes the $(1-t)$ quantile of the standard normal distribution; i.e., $\Phi(z_t) = 1 - t$, where Φ denotes the cumulative distribution function of the standard normal distribution.)

If we require further that the width of CI_n is no greater than $2d$; i.e., the inequality, $\sigma z_{\alpha/2}/\sqrt{n} \leq d$, must hold or, equivalently, the sample size must satisfy $n \geq (\sigma z_{\alpha/2}/d)^2$. This implies that the smallest sample size required such that the confidence interval CI_n has the required coverage probability $1-\alpha$ and length (precision) no greater than $2d$ is

$$\left[\sigma^2 \left(\frac{z_{\alpha/2}}{d}\right)^2\right] + 1\ ,$$

where notation $[r]$ denotes the largest integer less than r, for $r \in R$. If the variance σ^2 is unknown, then there is no fixed sample size procedure which can be used to construct this kind of confidence interval. Thus, a sequential procedure must be used in order to construct a fixed width confidence interval with the prescribed accuracy. By replacing the unknown σ^2 with its estimate, the stopping rule can be defined as the smallest integer such that the above inequality holds. The formal definition of the stopping rule is given later. For details, please see Siegmund (1985).

2.2. *Confidence Set with β-Protection*

As mentioned before, for a one-sided confidence interval, length is no longer an appropriate measure of precision. Instead of measuring the width of a confidence interval, the idea of "β-protection" proposed by Wijsman (1981, 1982) is to measure the probability of a one-sided confidence interval covering the "wrong" parameters.

We will first explain this idea of "β-protection" using the previous example of i.i.d. random variables. The general definition is given below.

Example 2. We keep the same notation of the previous example. Assume that μ is the only unknown parameter of the distribution. Let $S_n = [L_n(x_1, x_2, \dots, x_n), \infty)$ be a "left-closed, right-open" one-sided confidence interval of μ with coverage probability

$$P(\mu \in S_n) \approx 1 - \alpha \ , \tag{1}$$

where $\alpha \in (0, 1)$. Suppose $\psi(\mu)$ is a positive real-valued function of μ. (For example, we might simply set $\psi(\mu) = d$, a positive constant.) For a given $\beta \in (0, 1)$, if in addition to (**??**), the one-sided confidence interval S_n satisfies the following extra condition:

$$P(\mu - \psi(\mu) \in S_n) \leq \beta \tag{2}$$

for all μ, then it is called a $1 - \alpha$ confidence interval of μ with β-protection at $\mu - \psi$. Equation (2) simply means that the probability of the confidence interval S_n covering the "wrong" parameter, which is $\mu - \psi$ in the example, is less than β.

Suppose that θ_0 is the unknown parameter to be estimated. Let α and $\beta \in (0, 1)$ be two predefined constants and let $\phi(\theta_0)$ be a real-valued function of θ_0. Then the confidence interval with β-protection is defined as below.

Definition 1. A confidence interval S_n is called an $1 - \alpha$ confidence interval of θ_0 with β-protection at $\phi(\theta_0)$, if:

$$(\mathbf{D1}) P(\theta_0 \in S_n) \geq 1 - \alpha;$$

and

$$(\mathbf{D2}) P(\phi(\theta_0) \in S_n) \leq \beta.$$

For more detailed discussions and extensions of the β-protection confidence interval, we refer the reader to Wijsman (1981, 1982, 1986) and Juhlin (1985).

3. Adaptive Mastery Testing

The logistic model is one of the most popular statistical models used for the IRT-based standardized tests. The item characteristic curve (ICC) of a three-parameter logistic (3-PL) model (Birnbaum, 1968) is defined as

$$P(\theta) = P(Y = 1 \mid \theta; a, b, c) = c + \frac{1 - c}{1 + e^{-Da(\theta - b)}}, \tag{3}$$

where a, b, and c are item parameters and D is a constant. Equation (3) models the probability of a correct answer by a test-taker with latent ability θ to a given item with parameters a, b, and c. Here $Y = 1(0)$ denotes whether the item is answered correctly (or incorrectly), and a, b, and c are the item parameters of discrimination, difficulty, and guessing, respectively. The special case of $c = 0$ (no guessing) is called the two-parameter logistic (2-PL) model. If, in addition to $c = 0$, the discrimination parameters for all the items in the bank are the same, then it is called the Rasch model (Rasch, 1960).

Let the vector (a, b, c) denote an item selected from an item pool $\mathcal{B}$ with item parameters a, b and c. Suppose that we are in the $(n-1)$-th step of an adaptive test, then there have been $n-1$ items, (a_i, b_i, c_i), $i = 1, \ldots, n-1$, administered to the examinee. Let $Y_1, \ldots, Y_{n-1}, i = 1, \ldots, n-1$, be the corresponding responses of the examinee, and let $\mathcal{F}_{n-1}$ denote the σ-field generated by (a_i, b_i, c_i) and Y_i, $i = 1, \ldots, n-1$. Thus, the selection of the n-th item of the adaptive test will be based on knowledge in $\mathcal{F}_{n-1}$. Suppose that θ_0 is the true (but unknown) value of the examinee's latent ability level. Therefore, in the n-th step of the test, the estimate $\tilde{\theta}_n$ of θ_0 can be obtained based on the observation of (a_i, b_i, c_i) and Y_i, $i = 1, \cdots, n$. Hence, in adaptive testing, the property of the estimate $\tilde{\theta}_n$ of θ_0 depends on the estimating method as well as on the item selection scheme.

Suppose that all the items follow the general 3-PL model, of which the 2-PL model and the Rasch model are special cases. Then the likelihood function after n items have been administered is

$$L(\theta) = \prod_{i=1}^{n} P_i^{Y_i}(\theta) Q_i^{Y_i}(\theta). \tag{4}$$

Here P_i is the ICC for the i-th item as defined in (3), and $Q_i = 1 - P_i$. Therefore, the likelihood estimating function can be written as

$$\sum_{i=1}^{n} w_i(\theta) \left[Y_i - c_i - (1 - c_i) G(a_i(\theta - b_i)) \right], \tag{5}$$

where $w_i(\theta) = \partial \log[P_i(\theta)/Q_i(\theta)]/\partial\theta$ and $G(t) = e^t/(1 + e^t)$. Then the maximum likelihood estimate $\hat{\theta}_n$ is a maximizer of (4) or a root of (5).

Assume that the maximization or root finding is over a compact interval containing true parameter θ_0 as an interior point. Let $\hat{\theta}_n$ denote the maximum likelihood estimate of θ_0 and let $I_n(\theta)$ denote the Fisher information

$$I_n(\theta) = \sum_{i=1}^{n} \left[\frac{\partial P_i(\theta)}{\partial \theta} \right]^2 \bigg/ P_i(\theta) Q_i(\theta). \tag{6}$$

Then Chang and Ying (2003, Theorem 3.1) show that the maximum likelihood estimate $\hat{\theta}_n$ is strongly consistent and asymptotically normal under the following regularity conditions:

(C1) The number of items in the bank cannot be exhausted, i.e., the size of the item bank is infinite.

(C2) The parameters of all selected items satisfy the following constraints:

$$\sup_n |b_n| < \infty, \qquad 0 < \inf_n a_n < \sup_n a_n < \infty \quad \text{and} \quad \sup_n c_n < 1□ \text{ a.s.}$$

(C3) There exists a nonrandom sequence v_n such that $I_n(\theta_0)/v_n \to 1$ a.s.

To be more specific, they show that if (C1) and (C2) hold, then as $n \to \infty$, $\hat{\theta}_n \to \theta_0$ a.s. and if, in addition, (C3) holds, then $\sqrt{I_n(\hat{\theta}_n)}(\hat{\theta}_n - \theta_0) \to_{\mathcal{L}} N(0, 1)$.

The results of Chang and Ying (2003) can be extended to cover more general cases. In fact, the weight $w_i(\theta)$ in equation (5) may be replaced by any $\mathcal{F}_{i-1}$ measurable weight and the resulting estimating function is still valid. This can easily be seen as $E[Y_i - P_i(\theta_0) \mid \mathcal{F}_{i-1}] = 0$. This implies that for different item selection rules, the weight $w_i(\theta)$ may vary accordingly, but the estimating equation (5) is still a valid estimating function for estimating θ_0.

Suppose that $\tilde{\theta}_n$ is a unique solution to the estimation equation (7) below for some $\mathcal{F}_{i-1}$-measurable weight w_i, which may depend on the item selection method; that is, $\tilde{\theta}_n$ satisfies

$$\sum_{i=1}^{n} w_i(\tilde{\theta}_n)[Y_i - P_i(\tilde{\theta}_n)] = 0 \ . \tag{7}$$

Let

$$\delta_n^2(\theta) = \sum_{1}^{n} w_i(\theta) \frac{\partial P_i(\theta)}{\partial \theta} \ , \tag{8}$$

recalling that $P_i(\theta) = 1 - Q_i(\theta) = c_i + (1 - c_i)G(a_i(\theta - b_i))$. Similar to the case of maximum likelihood estimate, if $\delta_n(\theta_0)$ satisfies Condition (C3′) below:

(C3′) There exists a nonrandom sequence v_n' such that $\delta_n^2(\theta_0)/v_n' \to 1$ a.s. as $n \to \infty$.

Then it can also be shown that

Theorem 1. *Suppose $\tilde{\theta}_n$ is a unique solution to the estimating equation (5) with $w_i \in \mathcal{F}_{i-1}$ for all i. Then under Conditions (C1) and (C2), $\tilde{\theta}_n$ is strongly consistent; i.e., $\tilde{\theta}_n \to \theta_0$ a.s. as $n \to \infty$. If, in addition to Conditions (C1) and (C2), Condition (C3′) is satisfied, then it is asymptotically normal that $\delta_n(\tilde{\theta}_n)(\tilde{\theta}_n - \theta) \to_{\mathcal{L}} N(0,\ 1)$.*

Based on these asymptotic properties of $\tilde{\theta}_n$, we are able to construct one-sided confidence intervals for θ with β-protection. (The proof is provided in the final section.)

3.1. Application of One-Sided Confidence Intervals to Mastery Testing

Let $\tilde{\theta}_n$ be an estimate of θ_0 as defined above and let $S_n^l = [L_n(\tilde{\theta}_n),\ \infty)$ be a $1-\alpha$ one-sided (left-closed and right-open) confidence interval for θ_0 with β-protection at $\theta_0 - \psi(\theta_0)$, where $\alpha, \beta \in (0,\ 1)$ are two prechosen constants and ψ is a positive real-valued function. Then, according to Definition 1, the confidence set S_n must satisfy the following two inequalities: $P(\theta_0 \in S_n^l) \geq 1 - \alpha$ and $P(\theta_0 - \psi(\theta_0) \in S_n^l) \leq \beta$. It follows from the asymptotic that normality of $\tilde{\theta}_n$, that if

$$\delta_n(\tilde{\theta}_n)(\tilde{\theta}_n - L_n) \geq z_\alpha \ , \tag{9}$$

then

$$P(\theta_0 \geq L_n) = P\left(\delta_n(\tilde{\theta}_n)(\tilde{\theta}_n - \theta_0) \leq \delta_n(\tilde{\theta}_n)(\tilde{\theta}_n - L_n)\right) \geq 1 - \alpha \ . \tag{10}$$

Inequality (9) holds if, and only if,

$$L_n \leq \tilde{\theta}_n - \frac{z_\alpha}{\delta_n(\tilde{\theta}_n)} \ . \tag{11}$$

Thus, inequality (11) is a sufficient condition such that (D1) holds.

On the other hand, in order to satisfy (D2) of Definition 1, we must have

$$\beta \geq P(\theta - \psi \geq L_n) = P\left(\delta_n(\tilde{\theta}_n)(\tilde{\theta}_n - \theta) \leq \delta_n(\tilde{\theta}_n)(\tilde{\theta}_n - L_n - \psi)\right). \tag{12}$$

Again, based on the asymptotic normality of $\tilde{\theta}_n$, and using arguments similar to those above, it can be shown that equation (12) holds, if

$$L_n \geq \tilde{\theta}_n - \psi(\tilde{\theta}_n) + \frac{z_\beta}{\delta_n(\tilde{\theta}_n)} \ . \tag{13}$$

Here z_β is as defined earlier.

Thus, it follows from (11) and (13), we have the following inequality:

$$\tilde{\theta}_n - \psi + \frac{z_\beta}{\delta_n(\tilde{\theta}_n)} \le \tilde{\theta}_n - \frac{z_\alpha}{\delta_n(\tilde{\theta}_n)}. \tag{14}$$

This implies that

$$\delta_n(\tilde{\theta}_n) \ge \left(\frac{z_\alpha + z_\beta}{\psi}\right). \tag{15}$$

This suggests that we can set

$$L_n = \hat{\theta}_n - z_\alpha/\delta_n(\tilde{\theta}_n)\ ,$$

and if, in addition, inequality (15) is satisfied, then conditions (D1) and (D2) will hold. Hence, the procedure with the smallest n such that (15) holds, which implies (D1) and (D2) are satisfied, will be the most efficient one in terms of sample size (test length).

If the true θ_0 was known, it follows from (15) that the optimal sample size for constructing a $1 - \alpha$ one-sided confidence interval for θ_0 with β-protection at $\theta - \psi$ is

$$\begin{aligned} n_{\theta_0} &= \text{ smallest integer greater than } n_0 \text{ such that } \delta_n^2(\theta_0) \ge \left(\frac{z_\alpha + z_\beta}{\psi(\theta)}\right)^2 \\ &= \inf\left\{ n \ge n_0 : \delta_n^2(\theta) \ge \left(\frac{z_\alpha + z_\beta}{\psi(\theta)}\right)^2 \right\}\ . \end{aligned} \tag{16}$$

In other words, n_{θ_0} is the minimum sample size required such that the confidence set S_n^l satisfies Definition 1.

In practice, the true θ_0 is usually unknown, so there is no fixed sample size procedure that can be used for constructing this type of confidence set. By replacing the unknown θ_0 with its estimate $\tilde{\theta}_n$ in (**??**), we can define a stopping rule

$$\begin{aligned} T_\beta(\psi) &= \text{first integer } n \ge n_0 \text{ such that inequality (15) holds} \\ &= \inf\left\{ n \ge n_0 : \delta_n^2(\tilde{\theta}_n) \ge \left(\frac{z_\alpha + z_\beta}{\psi(\tilde{\theta}_n)}\right)^2 \right\}\ , \end{aligned} \tag{17}$$

where $n_0 > 0$ is a prefixed constant (initial sample size). That is, based on (15), we can define a stopping rule for constructing a one-sided confidence interval of θ_0 with β-protection at ψ, sequentially. (Note that in CAT/AMT, n_0 is the number of test items used as an initial test set or testlet.)

As we have seen in (8), $\tilde{\delta}_n = \delta_n(\tilde{\theta}_n)$ is usually an increasing function of the sample size n. (Its rate of increase depends on the item selection schemes used in the tests.) So, if n is large enough such that inequality (15) holds, both Conditions (D1) and (D2) will be satisfied. Based on this property, it can be shown that $P(T_\beta(\psi) < \infty) = 1$; i.e., the sequential procedure will stop, eventually. Then we have the following theorem.

Theorem 2. *Assume that Conditions (C1), (C2), and (C3′) are satisfied.Suppose that $\tilde{\theta}_n$ satisfies the assumption in Theorem 1. Let $L_n = \tilde{\theta}_n - z_\alpha/\delta_n(\tilde{\theta}_n)$ and let $S_n^l = [L_n, \infty)$ be a left-closed, right-open interval, then:*

(i) $P(T < \infty) = 1; \lim_{\psi\to 0} P(\theta_0 \in S_T) = 1 - \alpha$, and $\lim_{\psi\to 0} P(\theta_0 - \psi(\theta_0) \in S_T) = \beta$;
(ii) $\lim_{\psi\to 0} E\left[\frac{T_\psi}{n_{\theta_0}}\right] = 1$.

Statement (i) means that the sequential procedure will stop with probability 1 and the confidence set will satisfy the prescribed condition of coverage probabilities (D1) and (D2), asymptotically. Statement (ii) means that the ratio of the stopping time to the optimal fixed sample size n_{θ_0} converges to 1 as ψ goes to 0. In other words, the test procedure will eventually stop and it will have correct classification probabilities, as required, when it stops. It is also asymptotically efficient, since the ratio of average test length to optimal test length goes to 1. (Note that statement (ii) above is called "asymptotic efficiency" in Chow and Robbins (1965).)

As mentioned before, if Conditions (C1)–(C3) (or C3′) are satisfied, then any estimate $\tilde{\theta}_n$, which is obtained as a root of the estimating equation (7) with some $\mathcal{F}_{i-1}$-measurable weights w_i's, will have similar asymptotic properties. Since we are not concerned with item selection rules here, for simplicity throughout the remainder of this paper, we will assume that the maximum likelihood estimate is used to estimate θ_0, and item selection is based on the maximum Fisher's information principal, as in Chang and Ying (2003).

3.2. Decision Rule and Indifference Region

3.2.1. Single one-sided confidence interval procedure. Suppose θ_0 is the true (but unknown) proficiency level of a test-taker, and let θ_c denote the predefined cutoff level. Let $S_n^l = [L_n, \infty)$ be the $1 - \alpha$ one-sided confidence interval of θ with β-protection at $\theta - \psi$ for $\psi > 0$. Then, we can apply this β-protected confidence interval of θ_0 to AMT in the following way (we will call it the "single one-sided confidence interval procedure" in this paper): When the sampling stops,

$$\begin{cases} \text{Pass the test-taker} & \text{if } L_T \geq \theta_c, \\ \text{Fail the test-taker} & \text{if } L_T < \theta_c \end{cases} \tag{18}$$

As will be shown later, the required classification accuracy could not be maintained, when the true θ is actually within $[\theta_c, \theta_c + \psi]$. Indeed, the interval $[\theta_c, \theta_c + \psi]$ is an "indifference region" of this procedure.

Definition 2. (Indifference Region, Spray (1993)). $\Theta_0 \subset \Theta$ is called an "α-indifference region" for an $\alpha \in (0, 1)$ if and only if for all $\theta \notin \Theta_0$, the probability of misclassification is less than α, and for $\theta \in \Theta_0$ the probability of misclassification is greater than α.

Thus, if we apply the single one-sided (left-close, right-open) confidence interval procedure with coverage probability $1 - \alpha$ and β-protection probability β to AMT, then the false pass and the false fail probabilities are α and β, respectively.The indifference region is $[\theta_c, \theta_c + \psi]$, where ψ is a positive constant as defined above.

Due to the asymmetric misclassification probabilities of the procedure, we have a modified definition of the indifference region to note the asymmetry.

Definition 3. $\Theta_0 \subset \Theta$ is called an "$\alpha\beta$-indifference region" for $0 < \alpha, \beta < 1$ if and only if for all $\theta \notin \Theta_0$, the probability of misclassification is less than $\max\{\alpha, \beta\}$, and for $\theta \in \Theta_0$ the probability of misclassification is greater than $\min\{\alpha, \beta\}$.

Now, using Theorem 2, we can show

Corollary 1. *Under the assumption of Theorem 2, if decision rule (18) is applied to AMT; then:*

(i) for all $\theta_0 \notin [\theta_c, \theta_c + \psi]$ the misclassification probability for $\theta_0 > \theta_c + \psi$ (false fail) is less than β and the misclassification probability for $\theta_0 < \theta_c$ (false pass) is less than α; and

(ii) for $\theta_0 \in [\theta_c, \theta_c + \psi]$ the misclassification probabilities are greater than $\min\{\alpha, \beta\}$.

If the true latent trait level $\theta_0 \in [\theta_c,\ \theta_c + \psi]$, then Corollary 1(ii) states that the correct classification probabilities might not satisfy the required solution (see also equations (27) and (28)). That is, the interval $[\theta_c,\ \theta_c + \psi]$ is an "**indifference region**" for this procedure. Note that the misclassification probabilities of the false pass and the false fail in this case are not necessarily the same. That's why we call it an $\alpha\beta$-indifference region to denote its asymmetry. This is a nice property and similar to that achieved by SPRT to mastery testing.

Obviously, we can also apply one "right-open, left-closed" one-sided interval (such as $(-\infty,\ R_n])$ to AMT. The results are similar, but the indifference region becomes $[\theta_c - \psi,\ \theta_c]$.

As has been shown, application of the one-sided confidence interval to AMT will give asymmetrical misclassification probabilities (i.e., $\alpha \neq \beta$) and asymmetrical indifference regions (asymmetric around the cutoff level θ_c). These properties provide the test administrator with options to set up the adaptive mastery test according to their interests or testing purposes.

3.2.2. Double one-sided confidence intervals with β-protection. If symmetry of the indifference region is required, then it can be achieved by applying two one-sided confidence intervals together to AMT; that is, applying "left-closed, right-open" and "left-open, right-closed" confidence intervals at the same time to the mastery test. This is called a "*double one-sided confidence intervals procedure*" in this paper. The fixed-width confidence interval procedure can be viewed as a special case.

Suppose that $\alpha, \beta \in (0, 1)$ are two given constants and that ψ' is a positive real-valued function of θ. Define $R_n = \hat{\theta}_n + z_\beta/\delta_n(\hat{\theta}_n)$ and let $S_n^r = (-\infty,\ R_n]$ be a right-closed, left-open interval. Then it follows, from similar arguments as before, that we can define a stopping rule

$$T' = T'_{\psi'} = \inf\left\{ n \geq n_0 : I_n(\hat{\theta}_n) \geq \left(\frac{z_\alpha + z_\beta}{\psi'}\right)^2 \right\}, \tag{19}$$

such that when the sampling stops the interval $S_{T'}^r$ is a $1 - \beta$ one-sided confidence interval for θ_0 with α-protection at $\theta_0 + \psi'$. (Here we switch the roles of α and β.)

Note that the stopping time T' is symmetric for z_α and z_β; that is, the stopping rules for constructing the left-close, right-open confidence interval (T) and the right-close, left-open confidence interval (T') are the same. Therefore, if we let $L_n = \hat{\theta}_n - z_\alpha/\delta_n(\hat{\theta}_n)$ and let $S_n^l = [L_n,\ \infty)$, then $S_{T'}^l$ is also a $1 - \alpha$ confidence interval with β-protection at $\theta_0 - \psi'$. Hence, by using the same stopping rule T', we can construct two (different direction) one-sided confidence intervals, S_n^l and S_n^r, at the same time. Therefore, when the sampling stops, we will have

$$\begin{aligned} &P\big(\theta \in S_{T'}^r\big) \geq 1 - \beta, \qquad P\big(\theta + \psi' \in S_{T'}^r\big) \leq \alpha, \qquad P\big(\theta \in S_{T'}^l\big) \geq 1 - \alpha, \\ &\quad \text{and } P\big(\theta - \psi' \in S_{T'}^l\big) \leq \beta. \end{aligned} \tag{20}$$

Now, we can apply the double one-sided confidence intervals procedure to mastery testing in the following way:

$$\begin{cases} \text{Pass the test-taker} & \text{if } L_{T'} \geq \theta_c, \\ \text{Fail the test-taker} & \text{if } R_{T'} < \theta_c, \\ \text{No decision} & \text{if } L_{T'} < \theta_c \leq R_{T'}. \end{cases} \tag{21}$$

Remark 1. If $L_{T'} < \theta_c \leq R_{T'}$, then it is very likely that the true θ_0 actually belongs within the indifference region. Therefore, no decision should be made in this case, since the misclassification probabilities may be greater than required.

Again, if we assume θ_0 is known, then the smallest sample size for constructing confidence intervals satisfying the four inequalities in (20) is

$$n' = n'_{\theta_0} = \inf\left\{n \geq n_0 : I_n(\theta) \geq \left(\frac{z_\alpha + z_\beta}{\psi'(\theta)}\right)^2\right\}.$$

By Theorems 1 and 2, we know that $P(T' < \infty) = 1$, $\lim_{\psi' \to 0} E[T'/n'] = 1$, and we have the following corollary:

Corollary 2. *Under the assumptions of Theorems 1 and 2, if the double one-sided confidence intervals procedure is applied to AMT with confidence sets S_n^r, S_n^l (as defined above) and decision rule (21), then for all $\theta \notin [\theta_c - \psi',\ \theta_c + \psi']$, the misclassification probabilities of the false pass and the false fail are less than α and β, respectively; the classification probabilities of correct pass and correct fail are greater than $1 - \beta$ and $1 - \alpha$, respectively.*

It can be shown that the indifference region for the double one-sided confidence intervals procedure is $[\theta_c - \psi',\ \theta_c + \psi']$, which is symmetric around the predecided threshold θ_c. Now the width of the indifference region is $2\psi'$. Hence, if we require these two procedures (the single and double one-sided confidence intervals procedures) to have the same width of indifference regions, then we must set $\psi' = \psi/2$. According to the definitions of stopping times (see equations (17) and (19)), this will imply that $T'_{\psi'} = 4T_\psi$ almost surely.

Note that the fixed-width confidence interval can be constructed using this double one-sided confidence intervals procedure. If we set $\alpha = \beta$, then all classification probabilities are the same. Based on previous discussions, we can conclude that given the same classification probabilities and the same width of indifference region, the sample size (test items) required for the double one-sided confidence intervals procedure (i.e., the fixed-width confidence interval procedure) is four times the sample size of the single one-sided confidence interval procedure. Therefore, if we are not concerned with the asymmetry of the indifference region, and the false pass and false fail probabilities are equal, the procedure using a single one-sided confidence interval is more efficient than the one using a fixed-width confidence interval.

Remark 2. In the double one-sided procedure, if we choose $\psi(\theta)$ to be a constant and $\beta = \alpha$, then the interval $[L_T,\ R_T]$ is a $1 - 2\alpha$ fixed-width (with length $2\psi'$) confidence interval for θ_0 (Ghosh & Sen, 1991; Ghosh, Mukhopadhyay, & Sen, 1997). That is, the fixed-width confidence interval can be treated as a special case of the double one-sided confidence procedure.

Similarly, the fixed proportional accuracy confidence interval (see Woodroofe, 1982; Siegmund, 1985; Chang & Martinsek, 1992) can be constructed using this double one-sided confidence intervals procedure by using a different choice for ψ. Below, we will discuss some possible choices for ψ. Since the true ability level θ is unknown in practice, $\psi = \psi(\theta)$ will usually be unknown. In this case, we can replace the unknown θ in the stopping rule with its estimate. Note that if $\psi(\theta)$ is not a constant, then the "length" of the indifference region is no longer fixed, and the expected test length will vary for different latent ability levels.

Choice of ψ. There are several possible choices for function ψ. For example, we can:

(1) simply let ψ be a positive constant; i.e., $\psi = d$ for $d > 0$ as before; or
(2) let $\psi(\theta) = k|\theta - \theta_c|$ for some constant $k \in (0, 1)$, where θ denotes the true ability level of interest.

If we choose $\psi(\theta) = k|\theta - \theta_c|$, then the indifference region varies according to the value of the true latent ability θ. In actual fact, the interval $[\theta_c \pm \psi]$ will degenerate to the singleton

$\{\theta_c\}$ as $|\theta - \theta_c|$ goes to 0. By the definition of the stopping rule (see (17) or (19)), the smaller the ψ, the larger the test length (T); or, equivalently, the larger $|\theta_0 - \theta_c|$ is, the fewer the test items that will be needed. This property allows us to use fewer items to classify test-takers with ability levels that are far from the cutoff level. Therefore, choosing a varying ψ will be more efficient than using the procedure with a constant ψ in average for the same ability level θ.

Remark 3. In theory, when the indifference region is degenerated to $\{\theta_c\}$, we should be able to classify all test-takers correctly, except for those with ability level θ_c. However, when $\psi = k|\theta_0 - \theta_c|$ goes to 0, the required number of test items goes to infinity. Although, in theory, we do not require an indifference region in this case, in practice, we may still need to set up an indifference region for the procedure with nonconstant ψ. Otherwise, truncation rules will be needed.

4. Simulation

The ability levels $\theta = \pm 2, \pm 1, \pm 0.6, \pm 0.3$, and 0 are included in the numerical studies. In all the simulation studies in this paper, the items are generated based on the 3-PL model with $a \in (0.5,\ 3)$, $b \in (-3.5,\ 3.5)$, and $c \in (0,\ 0.1)$, where the constant $D = 1$. Items are selected based on the maximum Fisher information principle of selecting items with parameter b matching the most recent estimate of the examinee's unknown proficiency level θ, while a and c are randomly (uniformly) generated from the above ranges (see Chang & Ying, 1999).

All three methods mentioned in the previous sections are studied, including single one-sided and double one-sided procedures with constant ψ, as well as the single one-sided procedure with varying ψ. We conduct 1000 runs for each combination of different choice of θ's and ψ's with an initial sample size $n_0 = 15$. In all the studies, we set $\alpha = \beta = 0.05$; so $z_\alpha = z_\beta$ and $\Phi(z_\alpha) = \Phi(z_\beta) = 0.05$. The cutoff level θ_c is equal to 0 for all the simulation studies.

For comparison purposes, we also have a small scale numerical study for SPRT using the same adaptive item selection method. The SPRT is set up for testing $H_0 : \theta = 0$ versus $H_1 : \theta = 0.5$; so the indifference region of the SPRT is [0, 0.5]. For more complete numerical studies applying SPRT to the (adaptive) mastery test, please see Reckase (1983), Spray and Reckase (1996), and Kingsbury and Weiss (1983).

Remark 4. Note that as mentioned in Reckase (1983), the original SPRT was created for the statistical testing problem of (simple versus simple) hypotheses under an assumption that the observations are independent. For a theoretical discussion of SPRT under more general situations, we refer the reader to Tartakovsky (1998) and Chang (2003).

Before we get into numerical details, we can gain a general idea of the properties of estimation and the distribution of test length from Figures 1 and 2. Figure 1 shows the Q–Q plot and histogram of the estimate of true ability level $\theta = 0.6$ with $n_0 = 15$ and $\psi = 0.5$. This again confirms the asymptotical normality of the sequential estimate of the true θ. These graphs are standard, so we omit the plots of other ability levels. Figure 2 contains histograms of test lengths for $\theta = \pm 0.6$ and ± 1 with the same n_0 and ψ as above. For procedures with constant ψ, the test length averages should be very close, since they use the same stopping rule.

However, the width of the indifference region of the double one-sided procedures is twice that of the other two types of procedures. The effect of the indifference region can be seen in Tables 1 to 3 by comparing the classification probabilities among different ability levels. For example, the double one-sided procedure cannot have the required classification probabilities for $\theta = \pm 0.3$, while the left-closed, right-open procedure can have the required classification probability at $\theta = -0.3$.

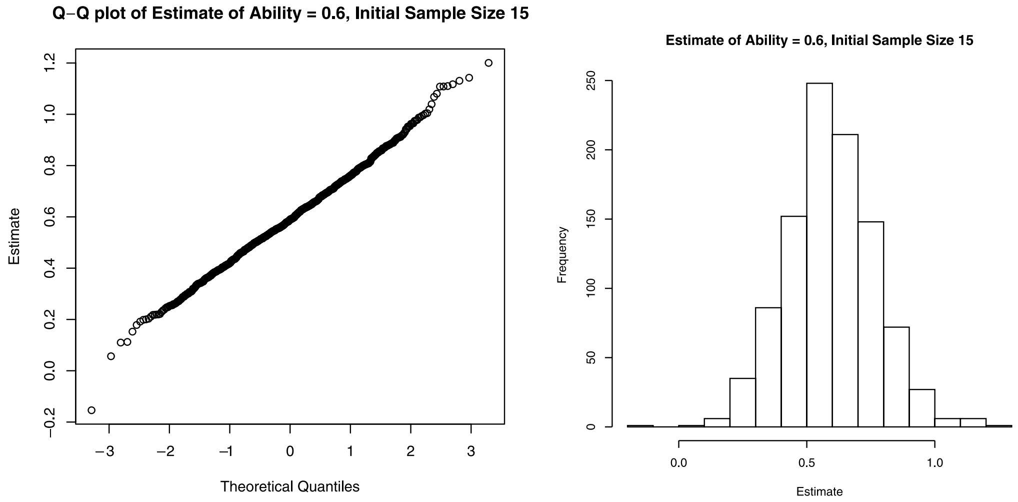

FIGURE 1.
Distribution of test length and estimate of ability = 0.6 with initial sample Size 15.

We now summarize the numerical results for each procedure below:

Study 1: Constant ψ. Tables 1 to 3 summarize the simulation results for $\psi = 0.3, 0.5$, and 0.7, Respectively. The column headings for these three tables, from left to right, are: (1) ability level (θ), (2) coverage frequency of the fixed-width confidence interval with width 2ψ (Fixed C.F.); (3) coverage frequency of the left-closed, right-open, one-sided confidence interval (Left C.F.), and its corresponding Frequency of; (4) Pass; and (5) Fail; (6) coverage frequency of the right-closed, left-open, one-sided confidence interval (Right C.F.); and its corresponding frequency of; (7) Pass; and (8) Fail; frequency of; (9) Pass; and (10) Fail of the double

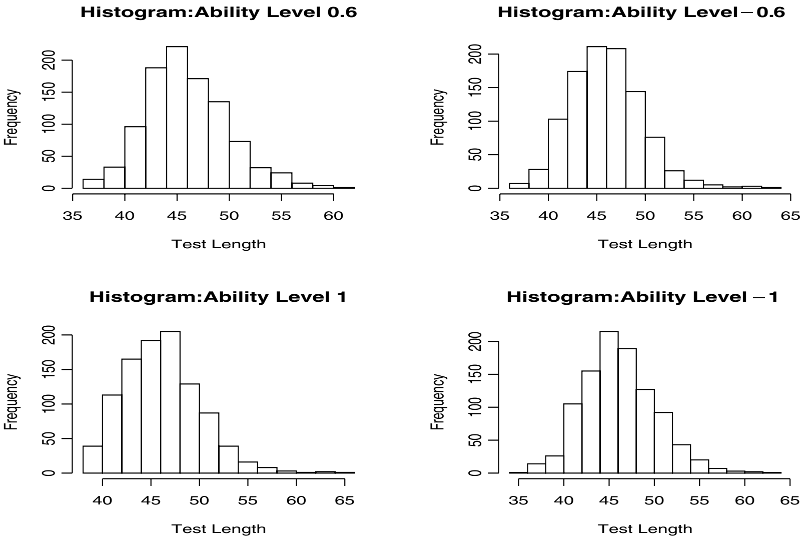

FIGURE 2.
Test length of procedures with initial sample size 15.

TABLE 1.
Initial sample size $n_0 = 15$ and $\psi = 0.3$

Ability	Fixed	Left			Right			Double		Test Length
θ	C.F.	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD
2	0.987	0.957	1	0	0.979	1	0	1	0	82.42/5.47
1	0.980	0.959	1	0	0.962	1	0	1	0	81.79/5.43
0.6	0.978	0.958	0.999	0.001	0.964	1	0	0.999	0	81.70/5.36
0.3	0.977	0.964	0.709	0.291	0.973	1	0	0.709	0	81.81/5.34
0	0.989	0.963	0.037	0.963	0.983	0.983	0.017	0.037	0.017	82.06/5.33
−0.3	0.984	0.964	0.002	0.998	0.963	0.292	0.708	0.002	0.708	81.78/5.39
−0.6	0.984	0.960	0.002	0.998	0.969	0.003	0.997	0.002	0.997	81.89/5.41
−1	0.989	0.969	0.007	0.993	0.976	0.007	0.993	0.007	0.993	81.57/5.50
−2	0.962	0.943	0.022	0.978	0.973	0.022	0.978	0.022	0.978	82.08/5.32

one-sided confidence intervals procedure and (11) average test length and its standard deviation (Test Length). The columns titled "C.F." denote the coverage frequencies of the corresponding confidence intervals. The tables show that all the coverage probabilities are close to the target probability of 95%.

The effect of the indifference region can be clearly seen from Tables 1 to 3. For example, in Table 1 with $\psi = 0.3$, the one-sided confidence interval procedures (the left-close, right-open and the right-close, left-open) fail at either $\theta = 0.3$ or $\theta = -0.3$ (but not both), while the double one-sided procedure fails at both $\theta = 0.3$ and -0.3.

By comparing the results with different ψ's, we found that when ψ was getting larger, only θ's that are far away from the cutoff level can be classified correctly with the required classification probabilities. On the other hand, the choice of ψ does not affect the coverage frequencies of any interval. The required coverage frequency can still be maintained for all θ's in the studies.

Note that the average test length and its standard deviation are very close among different ability levels, as expected for the constant ψ. The sum of the frequency of pass and fail in the double one-sided procedure is not necessarily equal to one. This is the case when the estimate of θ, $\hat{\theta}_T$, falls into the region $[L_T, R_T]$, indicating that the true θ might belong within the indifference region and no decision will be made in this case.

Since the results of the case with initial sample size $n_0 = 10$ are very similar to $n_0 = 15$, we provide only the summarization of a simple case, where $\psi = 0.3$, in Table 4. Comparing Table 4

TABLE 2.
Initial sample size $n_0 = 15$ and $\psi = 0.5$

Ability	Fixed	Left			Right			Double		Test Length
θ	C.F.	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD
2	0.993	0.967	1	0	0.971	1	0	1	0	46.58/4.02
1	0.994	0.967	1	0	0.972	1	0	1	0	46.60/3.97
0.6	0.991	0.971	0.941	0.059	0.968	1	0	0.941	0	46.40/3.95
0.3	0.994	0.968	0.432	0.568	0.972	1	0	0.432	0	46.57/3.92
0	0.991	0.963	0.037	0.963	0.973	0.973	0.027	0.037	0.027	46.52/4.07
−0.3	0.992	0.966	0.002	0.998	0.958	0.587	0.413	0.002	0.413	46.54/3.91
−0.6	0.991	0.969	0.003	0.997	0.971	0.074	0.926	0.003	0.936	46.71/3.96
−1	0.988	0.961	0.007	0.993	0.974	0.007	0.993	0.007	0.993	46.61/4.00
−2	0.943	0.915	0.047	0.953	0.980	0.047	0.953	0.047	0.953	46.97/4.24

TABLE 3.
Initial sample size $n_0 = 15$ and $\psi = 0.7$

Ability	Fixed	Left			Right			Double		Test Length
θ	C.F.	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD
2	0.998	0.954	1	0	0.973	1	0	1	0	35.18/3.45
1	0.994	0.953	0.997	0.003	0.969	1	0	0.997	0	35.10/3.25
0.8	0.997	0.952	0.962	0.038	0.966	1	0	0.962	0	35.04/3.35
0.6	0.998	0.960	0.829	0.171	0.964	1	0	0.829	0	34.88/3.29
0	0.996	0.953	0.047	0.953	0.964	0.964	0.036	0.047	0.036	35.00/3.26
−0.6	0.992	0.952	0.007	0.993	0.965	0.188	0.812	0.007	0.812	35.19/3.37
−0.8	0.989	0.951	0.009	0.991	0.963	0.053	0.947	0.009	0.947	35.07/3.35
−1	0.993	0.961	0.005	0.995	0.974	0.010	0.990	0.005	0.990	35.00/3.22
−2	0.995	0.917	0.041	0.959	0.970	0.041	0.959	0.041	0.959	35.09/3.22

with Table 1, we can see that performance in the case of $n_0 = 10$ is very similar to that in the case of $n_0 = 15$. However, it can be seen in Table 4 that the average test length is a little bit smaller than in the case of $n_0 = 15$, but variance (standard deviation) becomes slightly larger. A reasonable explanation of this phenomenon is that the estimate of θ obtained from a procedure with a smaller number of initial test items might not be as stable as the one obtained from a procedure starting with more initial test items.

Study 2: Varying ψ There are many possible choices for ψ. In this simulation study, we set $n_0 = 15$ and let $\psi(\theta) = |\theta - \theta_c|$; i.e., let the "$\beta$-protected parameter" vary according to the distance between true θ and the cutoff level θ_c. In theory, we should have no indifference region for the case of varying ψ; that is, this kind of procedure should be able to classify all values of θ in the range, except for the singleton, $\theta = \theta_c$.

By the definition of the stopping rule, it is clear that the sample size will be larger when the true θ is close to θ_c. In other words, the closer θ is to the threshold, the more test items will be needed in order to make a correct classification. If the value of $|\theta - \theta_c|$ goes to zero, then the expected stopping rule will go to infinity; i.e., the item pool may be exhausted by those θ's very close to the cutoff level θ_c. Therefore, an indifference region should be enforced here to prevent the item pool from being exhausted. Otherwise, some truncation rule will be needed. (Note that, in practical applications, the true proficiency level θ is unknown, as is $\psi(\theta)$. Thus, an estimate of ψ, say $\hat{\psi}_n = \psi(\hat{\theta}_n)$, will be used instead.)

In Table 5, we summarize the results of $\psi(\theta) = |\theta - \theta_c|$ for $\theta = -1, -0.8, -0.6, \ldots, 1$. The classification probabilities of all different proficiency levels in this study achieve the predescribed requirements. That is, there is no indifference region when the varying $\psi = |\theta_0 - \theta_c|$ is used. We

TABLE 4.
Initial sample size $n_0 = 10$ and $\psi = 0.3$

Ability	Fixed	Left			Right			Double		Test Length
θ	C.F.	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD
2	0.982	0.954	1	0	0.964	1	0	1	0	78.83/5.69
1	0.979	0.959	1	0	0.970	1	0	1	0	78.65/5.9
0	0.976	0.967	0.033	0.967	0.972	0.972	0.028	0.033	0.028	78.62/5.35
−1	0.986	0.963	0	1	0.965	0	1	0	1	78.44/5.23
−2	0.978	0.967	0	1	0.954	0	1	0	1	78.65/5.88

TABLE 5.
Initial sample size $n_0 = 15$ and $\psi = |\theta - \theta_c|$

Ability	Left			Right			Double		Test Length
θ	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD
1	0.952	1	0	0.972	1	0	1	0	28.44/5.08
0.8	0.950	1	0	0.983	1	0	1	0	32.29/6.89
0.6	0.948	1	0	0.982	1	0	1	0	40.85/11.57
0.4	0.952	1	0	0.991	1	0	1	0	57.75/22.77
0.2	0.927	1	0	0.989	1	0	1	0	135.00/87.01
−0.2	0.983	0.003	0.997	0.966	0.003	0.997	0.003	0.997	138.74/86.41
−0.4	0.980	0.002	0.998	0.950	0.003	0.998	0.002	0.998	59.85/24.58
−0.6	0.979	0.002	0.998	0.953	0.002	0.998	0.002	0.998	40.51/11.69
−0.8	0.978	0.001	0.999	0.957	0.002	0.998	0.002	0.998	33.20/7.40
−1	0.973	0.006	0.994	0.965	0.007	0.993	0.007	0.993	28.63/5.43

also found that under the same prescribed misclassification probabilities (α and β), the average test lengths for $\theta = 1, 0.8, 0.6, -0.6, -0.8, -1$ are shorter than those in the case where ψ is a constant.

Heuristically, if the true θ is farther away from the cutoff level, then it should be easier to classify correctly than those θ's in the near neighborhood of θ_c. That's why the procedures with varying ψ are able to classify test-takers with proficiency levels far from the threshold more efficiently than the procedures with constant ψ's. As expected, the smaller the value of $|\theta_0 - \theta_c|$, the larger the average test length. This result agrees with the definition of the stopping rule.

For comparison purposes, we include a simulation study of AMT using SPRT; i.e., the items are selected adaptively to the estimate of the test-taker's ability, as previous simulation studies. Here, the set up of SPRT is for testing the null hypothesis $H_0 : \theta = 0$ versus the alternative hypothesis $H_1 : \theta = 0.5$; so the indifference region is [0, 0.5]. Again, the initial test length n_0 is 15.

Table 6 summarizes the simulation results of SPRT with the same adaptive item selection method as mentioned, and the column headings, from left to right, are: (1) ability; (2) average test length; (3) standard deviation (SD) of test length; and (4) classification probabilities (frequency of pass and fail).

Average test lengths increase as the true θ moves closer to the boundary of the indifference region. However, they are smaller than those of the estimation-oriented procedures proposed in this paper. For $\theta \notin [0, 0.5]$, the classification probabilities are better than required. When $\theta \in [0, 0.5]$ (i.e., $\theta = 0.2, 0.4$ in Table 6), as expected SPRT can no longer provide satisfactory results. Note that the variation of test length for these two θ's is greater than for the procedure with varying ψ. (Actually, the average test length in the case of $\theta \in [0, 0.5]$ may in theory go to infinity.) By looking at Tables 6 and 5, the classification probabilities and average test lengths of the procedure with varying ψ are very comparable to those of SPRT. This suggests that the procedure with varying ψ may be a good alternative for AMT.

Effect of truncation. The average test lengths in the simulations above seem reasonable, but it can be seen in Figure 2 that some of the test lengths could be unreasonably large for real testing situations. To be more useful in practical testing, it may be necessary to set an upper limit for the test length; i.e., to stop a test if the test length reaches a prefixed upper bound (a constant). This kind of truncation rule is useful in a practical sense, so we conduct simulation studies to find out how the results are affected by truncation. For simplicity, here we only consider the case of $\psi = 0.5$.

TABLE 6.
Sequential Probability Ratio Test (SPRT); Initial sample size $n_0 = 15$; Indifference Region $= [\theta_c, \theta_c + 0.5]$

Ability	Test Length		Reject H_0	Reject H_1
θ	Average	SD	Pass	Fail
1	24.06	4.96	1	0
0.8	26.97	6.78	0.999	0.001
0.6	33.51	11.92	0.992	0.008
0.4	47.53	26.58	0.885	0.115
0.2	80.67	106.55	0.331	0.689
−0.2	28.01	9.08	0.010	0.990
−0.4	23.56	5.76	0.002	0.999
−0.6	20.78	4.10	0	1
−0.8	19.30	3.52	0	1
−1	18.16	2.81	0	1

We begin by considering the property of the estimate of θ. Figure 3 shows Q–Q plots of the quantiles of the estimates of the truncated procedures (when the true $\theta = 1$). The "x-axis" is quantiles of the standard normal distribution. It shows that for both $n_0 = 10, 15$ and upper limits equal to 40 and 50, the distribution of estimates are still very close to the standard normal distribution.

Figure 4 shows histograms of test lengths when upper limits are enforced. The plots in the first row of Figure 4 are the cases of $n_0 = 10$, and the second row are the cases of $n_0 = 15$. From left to right, the upper bounds are 40, 50, and 60. Figure 4 shows that when the upper limit increases from 40 to 60, the effects of truncation fade away. When the upper limit is 60, the effect of truncation is very limited. This can also be seen in the variances of test lengths in

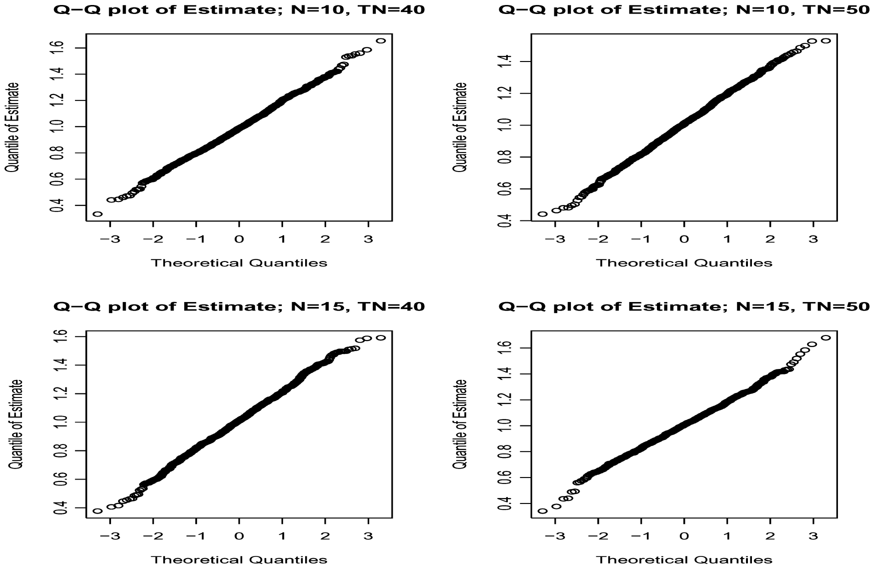

FIGURE 3.
Q–Q plots of the estimates of truncated procedures $\psi = 0.5$.

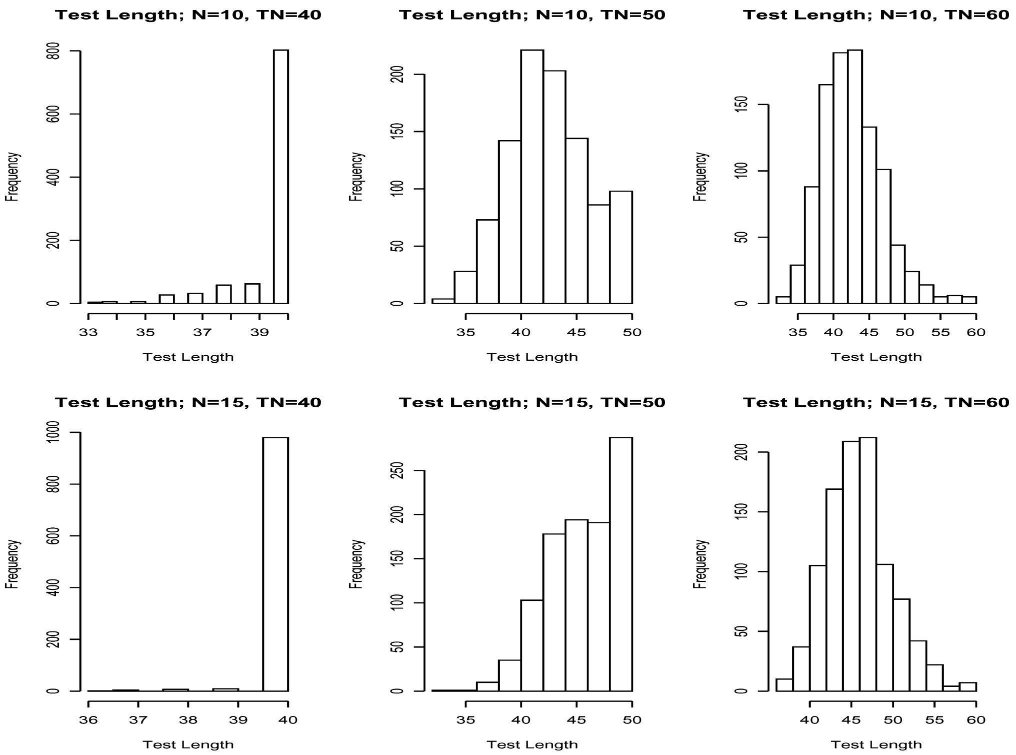

FIGURE 4.
Histograms of test lengths of truncated procedures $\psi = 0.5, n_0 = 10, 15$.

Tables 7 to 11. By comparing Table 9 with Table 2, we find that when the upper limit is 60, the variances of test lengths are very close to those of procedures with no upper limits. Moreover, their classification probabilities are also very close.

In Tables 7 to 11, we summarize the results of the procedures with upper limits enforced. The columns are arranged in an order similar to those as before, except for the last two columns. The last two columns of these tables record the number of tests where the test length reaches the upper bounds and the minimum test length used within 1000 simulations for each different θ. In these tables, we find that the classification probabilities are slightly smaller than before. In general, however, satisfactory results are produced; especially, when the true latent ability levels are far away from the indifference region.

From these tables, we also find that the minimum numbers of test items are all greater than the initial test length (i.e., 10 and 15). This result suggests that the estimate of θ_0 based on the initial test length with $n_0 = 10$ or 15 may not be a good estimate, but it might be "good enough" to provide the information for the adaptive item selection scheme.

Varying ψ with truncation. Tables 12 to 14 summarize the results of using varying $\psi = |\theta - \theta_c|$ with initial test length $n_0 = 15$ and upper limits 40, 50, and 60, respectively. All columns are in similar order, except that the column for coverage frequency of the fixed-width confidence interval is omitted. Note that in the last column of these tables, 15* means that there are some cases that use only the initial test items; that is, no extra items are required to make a decision. The number in parentheses denotes how many times this kind of situation happens during 1000 runs. There are only a few cases that do not require any extra items.

TABLE 7.
Truncated at 40; Initial sample size $n_0 = 15$, $\psi = 0.5$

Ability	Fixed	Left			Right			Double		Test Length	Hit	Min
θ	C.F.	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD	Boundary	Items
2	0.978	0.959	1	0	0.966	1	0	1	0	39.98/0.16	987	37
1	0.984	0.956	0.999	0.001	0.963	1	0	0.999	0	39.96/0.30	979	36
0.8	0.987	0.962	0.979	0.021	0.961	1	0	0.979	0	39.96/0.30	978	35
0.6	0.984	0.960	0.887	0.113	0.959	1	0	0.887	0	39.96/0.31	978	36
0.4	0.981	0.957	0.592	0.408	0.976	1	0	0.592	0	39.94/0.34	964	36
−0.4	0.976	0.949	0.004	0.996	0.971	0.462	0.538	0.004	0.538	39.95/0.33	971	36
−0.6	0.972	0.953	0.004	0.996	0.971	0.190	0.810	0.004	0.810	35.86/2.82	144	27
−0.8	0.981	0.959	0.001	0.999	0.973	0.026	0.974	0.001	0.974	39.96/0.30	973	36
−1	0.981	0.949	0.006	0.994	0.962	0.007	0.993	0.006	0.993	39.96/0.36	981	33
−2	0.932	0.909	0.045	0.955	0.966	0.045	0.955	0.045	0.955	39.96/0.26	977	36

TABLE 8.
Truncated at 50; Initial sample size $n_0 = 15$, $\psi = 0.5$

Ability	Fixed	Left			Right			Double		Test Length	Hit	Min
θ	C.F.	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD	Boundary	Items
2	0.994	0.961	1	0	0.973	1	0	1	0	46.27/3.16	233	36
1	0.989	0.965	1	0	0.969	1	0	1	0	46.05/3.19	209	33
0.8	0.994	0.960	0.995	0.005	0.969	1	0	0.995	0	45.98/3.13	188	36
0.6	0.992	0.974	0.932	0.068	0.968	1	0	0.932	0	45.96/3.19	195	34
0.4	0.995	0.973	0.618	0.382	0.963	1	0	0.618	0	46.01/3.12	207	34
−0.4	0.984	0.946	0.005	0.995	0.968	0.364	0.636	0.005	0.636	46.13/3.03	199	35
−0.6	0.969	0.960	0.007	0.993	0.969	0.185	0.815	0.007	0.815	36.01/3.42	4	27
−0.8	0.987	0.947	0.010	0.990	0.970	0.012	0.988	0.010	0.988	46.15/3.06	206	37
−1	0.980	0.947	0.006	0.994	0.975	0.007	0.993	0.006	0.993	45.19/3.14	187	36
−2	0.954	0.919	0.039	0.961	0.981	0.039	0.961	0.039	0.961	46.18/3.11	223	36

TABLE 9.
Truncated at 60; Initial sample size $n_0 = 15$, $\psi = 0.5$

Ability	Fixed	Left			Right			Double		Test Length	Hit	Min
θ	C.F.	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD	Boundary	Items
2	0.996	0.967	1	0	0.974	1	0	1	0	46.50/3.94	7	36
1	0.991	0.964	1	0	0.965	1	0	1	0	46.47/3.93	5	36
0.8	0.988	0.959	0.994	0.006	0.968	1	0	0.994	0	46.40/3.83	2	37
0.6	0.992	0.961	0.938	0.062	0.970	1	0	0.938	0	46.20/3.81	5	34
0.4	0.994	0.966	0.665	0.335	0.971	1	0	0.665	0	46.24/4.04	2	35
−0.4	0.989	0.958	0.003	0.997	0.969	0.356	0.644	0.003	0.644	46.39/4.13	7	36
−0.6	0.976	0.959	0.004	0.996	0.980	0.166	0.834	0.004	0.834	36.07/3.32	0	36
−0.8	0.982	0.960	0.08	0.992	0.961	0.016	0.984	0.008	0.984	46.50/3.78	4	36
−1	0.989	0.959	0.007	0.993	0.974	0.008	0.992	0.007	0.992	46.51/3.92	3	36
−2	0.969	0.939	0.028	0.972	0.971	0.028	0.972	0.028	0.972	46.87/3.88	6	37

TABLE 10.
Truncated at 40; Initial sample size $n_0 = 10$, $\psi = 0.5$

Ability	Fixed	Left			Right			Double		Test Length	Hit	Min
θ	C.F.	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD	Boundary	Items
2	0.992	0.968	1	0	0.973	1	0	1	0	39.64/1.00	841	33
1	0.984	0.959	1	0	0.976	1	0	1	0	39.52/1.03	810	34
0.8	0.980	0.958	0.986	0.014	0.976	1	0	0.986	0	39.62/1.00	820	33
0.6	0.984	0.953	0.921	0.079	0.978	1	0	0.921	0	39.60/1.02	823	33
0.4	0.980	0.947	0.645	0.355	0.965	1	0	0.645	0	39.63/0.98	825	34
−0.4	0.981	0.952	0.009	0.991	0.961	0.385	0.615	0.009	0.615	39.57/1.06	813	33
−0.6	0.958	0.950	0.017	0.983	0.971	0.186	0.814	0.017	0.814	32.60/3.20	31	24
−0.8	0.963	0.939	0.028	0.972	0.978	0.033	0.967	0.028	0.967	39.51/1.23	803	30
−1	0.963	0.942	0.024	0.976	0.967	0.025	0.975	0.024	0.975	39.61/1.02	825	33
−2	0.991	0.958	0	1	0.973	0	1	0	1	39.57/1.07	832	30

TABLE 11.
Truncated at 50; Initial sample size $n_0 = 10$, $\psi = 0.5$

Ability	Fixed	Left			Right			Double		Test Length	Hit	Min
θ	C.F.	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD	Boundary	Items
2	0.996	0.977	1	0	0.964	1	0	1	0	43.19/3.78	79	34
1	0.993	0.979	1	0	0.970	1	0	1	0	42.84/3.90	79	32
0.8	0.987	0.957	0.994	0.006	0.960	1	0	0.994	0	43.09/3.72	74	32
0.6	0.984	0.947	0.942	0.058	0.974	1	0	0.942	0	43.04/3.81	70	32
0.4	0.988	0.965	0.649	0.351	0.966	1	0	0.649	0	42.91/3.79	69	32
−0.4	0.977	0.950	0.017	0.983	0.969	0.369	0.631	0.017	0.631	42.95/3.73	61	32
−0.6	0.960	0.938	0.021	0.979	0.962	0.189	0.811	0.021	0.811	32.83/3.62	2	24
−0.8	0.970	0.941	0.023	0.977	0.977	0.030	0.970	0.023	0.970	42.85/3.69	61	33
−1	0.952	0.926	0.038	0.962	0.968	03038	0.962	0.038	0.962	42.99/3.84	79	32
−2	0.990	0.957	0	1	0.976	0	1	0	1	43.00/3.64	162	30

TABLE 12.
Varying $\psi = |\theta - \theta_c|$; Initial test length = 15, upper limit = 40

Ability	Left			Right			Double		Test Length	Hit	Min
θ	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD	Boundary	Items
1	0.946	0.998	0.002	0.972	1	0	0.998	0	28.37/4.81	32	17
0.8	0.945	0.984	0.016	0.968	1	0	0.984	0	31.47/5.58	149	19
0.6	0.949	0.878	0.122	0.970	1	0	0.878	0	35.7/5.15	457	18
0.4	0.944	0.571	0.429	0.973	1	0	0.571	0	38.66/3.41	801	17
0.2	0.947	0.208	0.792	0.961	0.995	0.005	0.208	0.005	39.70/1.64	956	25
−0.2	0.960	0.004	0.996	0.961	0.807	0.193	0.004	0.193	39.71/1.49	964	20
−0.4	0.968	0.002	0.998	0.949	0.453	0.547	0.002	0.547	38.70/3.28	806	20
−0.6	0.968	0.002	0.998	0.955	0.140	0.860	0.002	0.860	35.71/5.28	451	15* (1)
−0.8	0.954	0.002	0.998	0.961	0.021	0.979	0.002	0.979	32.21/5.54	182	19
−2.0	0.967	0.007	0.993	0.969	0.010	0.990	0.007	0.990	28.19/4.91	38	15* (1)

TABLE 13.
Varying $\psi = |\theta - \theta_c|$; Initial test length = 15, upper limit = 50

Ability	Left			Right			Double		Test Length	Hit	Item
θ	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD	Boundary	Range
1	0.936	1	0	0.982	1	0	1	0	28.21/4.87	0	17–46
0.8	0.993	0.998	0.002	0.981	1	0	0.998	0	32.03/6.69	20	17–50
0.6	0.943	0.958	0.042	0.973	1	0	0.958	0	37.97/8.21	162	16–50
0.4	0.942	0.692	0.308	0.959	1	0	0.692	0	45.01/7.16	561	20–50
0.2	0.931	0.260	0.740	0.966	0.998	0.002	0.260	0.002	48.88/3.95	895	21–50
−0.2	0.973	0.004	0.996	0.943	0.731	0.269	0.004	0.269	49.11/3.48	905	21–50
−0.4	0.95	0.002	0.998	0.961	0.336	0.664	0.002	0.664	45.55/6.88	597	15*–50 (1)
−0.6	0.960	0.003	0.997	0.962	0.056	0.944	0.003	0.944	39.00/8.32	201	19–50
−0.8	0.965	0.007	0.993	0.958	0.014	0.986	0.007	0.986	32.56/6.98	36	19–50
−1	0.965	0.008	0.992	0.964	0.009	0.991	0.008	0.991	28.72/5.53	9	18–50

TABLE 14.
Varying $\psi = |\theta - \theta_c|$; Initial test length = 15, upper limit = 60

Ability	Left			Right			Double		Test Length	Hit	Item
θ	C.F.	Pass	Fail	C.F.	Pass	Fail	Pass	Fail	Ave./SD	Boundary	Range
1	0.943	1	0	0.974	1	0	1	0	28.11/4.92	0	17–51
0.8	0.944	1	0	0.981	1	0	1	0	32.44/6.94	3	17–60
0.6	0.942	0.980	0.020	0.971	1	0	0.980	0	39.56/10.25	70	20–60
0.4	0.950	0.794	0.206	0.972	1	0	0.794	0	49.86/10.86	389	15*–60 (1)
0.2	0.934	0.315	0.685	0.966	1	0	0.315	0	57.65/6.91	834	15*–60 (1)
−0.2	0.954	0.004	0.996	0.963	0.697	0.303	0.004	0.303	58.23/5.49	856	20–60
−0.4	0.956	0.002	0.998	0.956	0.228	0.772	0.002	0.772	50.80/10.48	431	20–60
−0.6	0.949	0.010	0.990	0.970	0.035	0.965	0.010	0.965	40.56/10.29	84	15*–60 (2)
−0.8	0.972	0.005	0.995	0.969	0.007	0.993	0.005	0.993	33.42/7.35	4	15*–60 (1)
−1	0.959	0.013	0.987	0.963	0.013	0.987	0.013	0.987	28.80/5.77	1	15*–60 (1)

Most of the time, test lengths will reach the upper limits, except for θ with values far from θ_c. When we set the upper limit to 60, the performances of the cases where $\theta = -1, -0.8, -0.6, 0.6, 0.8, 1$ are very close to those of the procedures without upper limits. As a result of truncation, variances of test length are smaller than before.

5. Conclusion

In this paper, we apply the method of sequential one-sided confidence interval estimation with β-protection to AMT. From the theoretical arguments, it is clear that applying the fixed-width confidence interval procedure to AMT will require four times the test items (test length) than applying the single one-sided confidence interval procedure, when the widths of the indifference regions and the classification probabilities are equal.

The numerical results also show that applying one-sided confidence interval procedures to AMT is very promising. The asymmetric properties (i.e., the indifference region, and the classification probabilities) allow the test administrator to choose different false-pass and false-fail probabilities, or indifference regions, according to the practical testing situation. The double one-sided confidence intervals procedure or the fixed width confidence procedure will have a wider indifference region. Therefore, this procedure is recommended only when symmetry of the indifference region and classification probabilities are required.

There is another advantage of using the one-sided confidence interval procedure in AMT. That is, we can apply the item selection rules designed for CAT directly to AMT, since the proposed procedures in this paper are based on "estimation-oriented" construction (Wijsman, 1982).

Although we are not concerned with the item selection rule here, the assumption of the estimating equation of θ in this paper is very general, which allows us to apply other item selection rules to the proposed procedures. It is possible to make these procedures even more efficient with a carefully designed item selection method. For example, a carefully designed initial test set (i.e., the initial testlet with n_0 items), created according to the ability distribution of the test-takers, will provide a good starting point for the adaptive item selection rule and might make the procedure more efficient. This possibility will be studied in the future.

In theory, it is possible to choose ψ to be a monotonic function of $|\theta - \theta_c|$. In this case, the "indifference region" might be degenerated to a singleton θ_c. Letting ψ be a function of the distance of the true proficiency level to the threshold will require fewer items for classifying the test-takers with proficiency levels far from the threshold. This idea seems a very promising procedure to be applied to AMT. On the other hand, test length increases as the absolute value of the difference of $\theta - \theta_c$ decreases. Thus, in practical testing situations, an indifference region or a truncation rule still needs to be set. Here we only have a simulation study for a simple ψ, so the role varying ψ is not very clear at the moment. Further study is needed to clarify this point.

Appendix

In adaptive testing, items chosen for a test-taker is based on the estimate of his/her proficiency level. So, they are no longer independent. Hence, in order to apply a β-protection confidence interval estimation procedure to AMT, we have to extend Wijsman's (1981) results to the case of dependent observations.

Proofs of the strong consistency and asymptotic normality of $\tilde{\theta}_n$ are similar to that of Chang and Ying's (2003) Theorem 3.1, and only the highlights of the proof of Theorem 1 will be given here.

Proof of Theorem 1. Let θ_0 denote the true proficiency level, and let $\epsilon_i = Y_i - P_i(\theta_0)$. Suppose $\tilde{\theta}_n$ is the unique solution to the estimating function (5); i.e., it satisfies

$$\sum_1^n w_i(\tilde{\theta}_n)[Y_i - P_i(\tilde{\theta}_n)] = 0, \tag{22}$$

where $w_i \in \mathcal{F}_{i-1}, i = 1, 2, \ldots$. Under Conditions (C1) and (C2) and by Theorem 2 of Chow (1965), and similar arguments of Chang and Ying (2003), thus implies the strong consistency of $\tilde{\theta}_n$, i.e., $\tilde{\theta}_n \to \theta_0$ almost surely a.s. as $n \to \infty$.

Now, let's turn to the asymptotic normality of $\tilde{\theta}_n$. By the mean-value theorem, it can be shown that

$$\sum_1^n w_i(\theta_0)[Y_i - P_i(\theta_0)] = \delta_n^2(\theta^*)(\tilde{\theta}_n - \theta_0), \tag{23}$$

where $\delta_n^2(\theta^*) = \sum_1^n w_i(\theta^*)P_i'(\theta^*)$, $P' = \partial P/\partial\theta$ is the first derivative of P, and θ^* lies between θ_0 and $\tilde{\theta}_n$. (Note that if $w_i = \partial \log(P_i(\theta)/Q_i(\theta))/\partial\theta$ for all i, then $\delta_n^2(\theta^*) = I_n(\theta^*)$.)

Let $\tilde{\delta}_n = \delta_n(\tilde{\theta}_n)$. It follows from equation (23) that

$$\tilde{\delta}_n(\tilde{\theta}_n - \theta_0) = \frac{\tilde{\delta}_n v_n'^{1/2}}{\delta_n^2(\tilde{\theta}^*)}\delta_n^2(\tilde{\theta}^*)(\tilde{\theta}_n - \theta_0). \tag{24}$$

By Condition (C3'), $\delta_n^2(\theta_0)/v_n' \to 1$ almost surely as $n \to \infty$. Then, by a martingale central limit theorem (see Pollard (1984)), it can be shown that $\tilde{\delta}_n(\tilde{\theta}_n - \theta_0) \to_{\mathcal{L}} N(0,\ 1)$. This completes the proof of the asymptotic normality of $\tilde{\theta}_n$. □

Proof of Theorem 2. To prove that $\lim_{\psi\to 0} P(\theta_0 \in S_{T_\psi}) = 1 - \alpha$, it is sufficient to prove that

$$\tilde{\delta}_{T_\psi}(\tilde{\theta}_{T_\psi} - \theta_0) \to_{\mathcal{L}} N(0,\ 1). \tag{25}$$

Now, by Theorem 3.1 and Lemma 1 of Chow and Robbins (1965), we know that $P(T_\psi < \infty) = 1$ and $T_\psi/n_{\theta_0} \to 1$ almost surely as $\psi \to 0$. Then, by Anscombe's (1952) Theorem, equation (25) holds, if $\{\tilde{\delta}_n(\tilde{\theta}_n - \theta_0) : n \geq n_0\}$ is u.c.i.p. For further discussion about uniformly continuous in probability (u.c.i.p.), please see Woodroofe (1982).

It follows from the strong consistency of $\tilde{\theta}_n$ that

$$\tilde{\delta}_n(\tilde{\theta}_n - \theta_0) = [1 + o(1)]\left[v_n'^{-1/2}\sum_i^n w_i(\theta_0)\epsilon_i\right] \text{a.s.} \tag{26}$$

Let $S_n = \sum_1^n w_i\epsilon_i$ and $S_n^* = v_n'^{1/2}\sum_1^n w_i\epsilon_i$. Then, by using the similar arguments of Woodroofe (1982), and applying the Hàjek–Rènyi inequality (see Chow and Teicher, 1988, p. 247), we conclude that $\{\tilde{\delta}_n(\tilde{\theta}_n - \theta_0) : n \geq 1\}$ is u.c.i.p. This implies that $\tilde{\delta}_{T_\psi}(\tilde{\theta}_{T_\psi} - \theta_0) \to_{\mathcal{L}} N(0,\ 1)$. Hence, Theorem 2 (ii) holds.

To show that $\lim_{\psi\to 0} E[T_\psi/n_\theta] = 1$, it is sufficient to show that $\{\psi^2 T_\psi : \psi \in (0,\ 1)\}$ is uniformly integrable. This actually follows from the last-time arguments of Theorem 3.3 of Chang and Ying (2003) (see also Chang, (1999, 2001), and so it is omitted here. □

Proof of Corollary 1. Let θ be the true proficiency level. Suppose $\theta \geq \theta_c$, then, by the decision rule above, the probability of false fail becomes

$$P(\text{Fail}\,|\,\theta \geq \theta_c) = P(L_T < \theta_c\,|\,\theta \geq \theta_c) \leq P(L_T < \theta - \psi) \leq \beta, \tag{27}$$

provided that $\theta - \psi > \theta_c$. By (D2), it follows that the probability of correct pass is

$$P(\text{Pass}\,|\,\theta \geq \theta_c) = P(L_T \geq \theta_c\,|\,\theta \geq \theta_c) = 1 - P(L_T < \theta_c\,|\,\theta \geq \theta_c) \geq 1 - \beta, \tag{28}$$

provided that $\theta - \psi > \theta_c$.

If $\theta < \theta_c$, then by (D1) and similar arguments above, the probability of correct fail and false pass are $P(\text{Fail} \mid \theta < \theta_c) = P(L_T < \theta_c \mid \theta < \theta_c) = 1 - P(L_T \geq \theta_c \mid \theta < \theta_c) \geq 1 - \alpha$; and $P(\text{Pass} \mid \theta < \theta_c) = P(L_T \geq \theta_c \mid \theta < \theta_c) \leq P(\theta < L_T) \leq \alpha$, respectively. Equations (27) and (28) hold under the condition that $\theta - \psi > \theta_c$, which implies that $[\theta_c, \theta_c + \psi]$ is an indifference region of the procedure.

The proof of Corollary 2 follows from arguments similar to those found in Theorem 2 and Corollary 1. Only a sketch of the proof of the misclassification probability is given here.

Proof of Corollary 2. Again, let θ be the true proficiency level. Then the misclassification probabilities are

$$\begin{aligned} P(\text{Fail the test-taker} \mid \theta \geq \theta_c) &= P\{R_{T'} < \theta_c \mid \theta > \theta_c\} \\ &\leq PR_{T'} < \theta \mid \theta > \theta_c\} < \beta(\text{False Fail}), \end{aligned} \tag{29}$$

and

$$\begin{aligned} P(\text{Pass the test-taker} \mid \theta < \theta_c) &= P\{L_{T'} > \theta_c \mid \theta < \theta_c\} \\ &\leq P\{L_{T'} > \theta \mid \theta < \theta_c\} < \alpha(\text{False Pass}). \end{aligned} \tag{30}$$

For $\theta \notin [\theta_c - \psi', \theta_c + \psi']$, the probabilities of making the right decision are

$$\begin{aligned} &P(\text{Pass the test-taker } \mid \theta > \theta_c + \psi') = P(L_{T'} \geq \theta_c \mid \theta \geq \theta_c) \\ &\quad = 1 - P(L_{T'} < \theta_c \mid \theta \geq \theta_c) \geq 1 - P(L_{T'} > \theta - \psi' \mid \theta \geq \theta_c) \geq 1 - \beta, \end{aligned} \tag{31}$$

and

$$\begin{aligned} &P(\text{Fail the test-taker } \mid \theta < \theta_c - \psi') = P(R_{T'} < \theta_c \mid \theta < \theta_c) \\ &\quad = 1 - P(R_{T'} \geq \theta_c \mid \theta < \theta_c) \geq 1 - P\{R_{T'} > \theta + \psi' \mid \theta < \theta_0\} \geq 1 - \alpha. \end{aligned} \tag{32}$$

Remark 5. In the decision rule of the single one-sided procedure, by checking whether $\tilde{\theta} \in [\theta_c, \theta_c + \psi']$ or not, we can have some idea of the reliability of the decision. Similarly, the "No Decision" part of the decision rule of the double one-sided procedure is used to indicate that the true θ might fall into the indifference region. Note that

$$R_T - L_T = \frac{z_\alpha + z_\beta}{\delta_n} \approx \psi'.$$

Hence, if the true θ is in $[\theta_c \pm \psi']$, then it is very likely that $L_T \leq \theta_c \leq R_T$, and there is no way to guarantee that the probability of classification will be as expected in this situation.

References

Anscombe, F.J. (1952). Large-sample theory of sequential estimation. *Proceedings of the Cambridge Philosophical Society, 48*, 600–607.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novic (Eds.) *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Chang, Y.-C.I. (1999). Strong consistency of maximum quasi-likelihood estimate of generalized linear models via a last time. *Statistics and Probability Letters, 45*, 237–246.

Chang, Y.-C.I. (2001). Sequential confidence regions of generalized linear models with adaptive designs. *Journal of Statistical Planning and Inference, 93*, 277–293.

Chang, Y.-C.I. (2003). *Application of sequential probability ratio test to computerized criterion-referenced Testing.* Technical Report 2003-01. Academia Sinica, Taiwan: Institute of Statistical Science.

Chang, Y-C. I. & Martinsek, A. (1992). Fixed size confidence regions for parameters of a logistic regression model. *Annals of Statistics, 20*, 1953–1969.

Chang, H.-H. & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 263–278.

Chang, Y.-C.I. & Ying, Z. (2003). Sequential estimation in variable length computerized adaptive testing. *Journal of Statistical Planning and Inference*, (in Press).

Chow, Y.S. (1965). Local convergence of martingales and the law of large numbers. *Annals of Mathematical Statistics, 36*, 552–558.

Chow, Y.S. & Robbins, H. (1965). On the asymptotic theory of fixed-width sequential intervals for the mean. *Annals of Mathematical Statistics, 36*, 457–462.

Chow, Y.S. & Teicher, H. (1988). *Probability theory: independence, interchangeability, martingales.* New York: Springer-Verlag.

Epstein, K.I. & Knerr, C.S. (1977). Applications of sequential testing procedures to performance testing. In D.J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 249–270).

Ferguson, R.L. (1969). *The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction.* Doctoral dissertation, University of Pittsburgh. Dissertation Abstracts International, 30-09A, 3856. (University Microfilms No. 70–4530).

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

Ghosh, B.K. & Sen, P.K. (1991). *Handbook of sequential analysis.* New York: Marcel Dekker.

Ghosh, M., Mukhopadhyay, N. & Sen, P.K. (1997). *Sequential estimation.* New York: Wiley.

Glas, C.A.W. & Vos, H.J. (1998). *Adaptive mastery testing using the Rasch model and Bayesian sequential decision theory.* Research Report 98-15. Department of Educational Measurement and Data Analysis, University of Twente, Enschede, The Netherlands.

Juhlin, K.D. (1985). *Sequential and non-sequential confidence intervals with guaranteed coverage probability and beta-protection.* PhD. Dissertation. University of Illinois.

Kingsbury, G.G. & Weiss, D.J. (1983). A comparison of IRT-based adaptive mastery testing and 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.

Lord, F.M. (1971). Tailored testing, an application of stochastic approximation.*Journal of the American Statistical Association, 66*, 336, 707–711.

Pollard, D. (1984). *Convergence of stochastic processes.* New York: Springer-Verlag.

Rasch, G. (1960). *Probabilitistic models for some intelligence and attainment tests.* Copenhagen:Danmarks Paedagogiske Institut.

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.

Siegmund, D. (1985). *Sequential analysis.* New York: Springer-Verlag.

Spray, J.A. (1993). *Multiple-category classification using a sequential probability ratio test.* ACT Research Report Series 93-7. The American College Testing Program, Iowa.

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.

Tartakovsky, A. (1998). Asymptotic optimality of certain multihypothesis sequential tests: non-i.i.d. case. *Statistical Inference for Stochastic Processes, 1*, 265–295.

Wainer, H. (2000). *Computerized adaptive testing: A primer,* (2nd ed.), Hillsdale, NJ: Erlbaum.

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

Weiss, D.J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Wijsman, R. (1981). Confidence sets *Communications in Statistic—heavy and Methods, Series A*, based on sequential tests. 10, 2137–2147.

Wijsman, R. (1982). Sequential confidence sets: Estimation-oriented versus test-oriented construction. In S.S. Gupta & J.O. Berger (Eds.), *Statistical Decision Theory and Related Topics* (Vol.III, 2nd ed. pp. 435–450), New York: Academic Press.

Wijsman, R. (1986). Sequential confidence intervals with β-protection in one-parameter families. In J. Van Ryzin (Ed.) *Adaptive statistical procedures and related topics.* Lecture Notes—Monograph Series Vol. 8, Hayward, CA: Insitute of Mathematical Statistics.

Woodroofe, M. (1982). *Nonlinear renewal theory in sequential analysis.* CBMS-NSF Regional Conference Series in Applied Mathematics. Philadephia: SIAM.

Manuscript received 20 OCT 2003
Final version received 20 SEP 2004