MODERN SEQUENTIAL ANALYSIS AND ITS APPLICATIONS TO COMPUTERIZED ADAPTIVE TESTING

by

Jay Bartroff
Matthew Finkelman
Tze Leung Lai

Technical Report No. 2006-20
October 2006

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065
MODERN SEQUENTIAL ANALYSIS AND ITS APPLICATIONS TO COMPUTERIZED ADAPTIVE TESTING

by

Jay Bartroff
Department of Statistics
Stanford University

Matthew Finkelman
Measured Progress
Dover, NH 03820

Lze Leung Lai
Department of Statistics
Stanford University

Technical Reports No. 2006-20
October 2006

This research was supported in part by the National Science Foundation grant DMS 0305749

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
Modern Sequential Analysis and its Applications to Computerized Adaptive Testing

Jay Bartroff, Matthew Finkelman, and Tze Leung Lai

Abstract

After a brief review of recent advances in sequential analysis involving sequential generalized likelihood ratio tests, we discuss their use in psychometric testing and extend the asymptotic optimality theory of these sequential tests to the case of sequentially generated experiments, of particular interest in computerized adaptive testing. We then show how these methods can be used to design adaptive mastery tests, which are asymptotically optimal and are also shown to provide substantial improvements over currently used sequential and fixed length tests.

KEY WORDS: sequential analysis, computerized adaptive testing, mastery testing, generalized likelihood ratio statistics, item response theory.

1. Introduction

Sequential analysis of data is used in many types of psychometric tests. Some of these are computerized adaptive testing, classroom interaction assessment and intervention, psychological studies involving longitudinal data, depression diagnosis, and crime-suspect identification tests. The purpose of this article is to show how powerful techniques in modern sequential analysis can be used to design efficient testing procedures. In particular, we focus on computerized adaptive testing and show how these techniques can lead to substantial improvements over previous sequential procedures as well as conventional tests that do not incorporate early stopping.

Computerized adaptive testing (CAT) has been extensively studied in the psychometric literature as an efficient alternative to paper-and-pencil tests. By selecting an examinee’s kth test item based on his/her responses to items 1, . . . , k − 1, a CAT is tailored to the individual taking the examination and is thus intended to quickly home in on each examinee’s ability level. When the test is designed to measure only one trait, the ability level is typically denoted by θ to conform to the notation in standard item response theory. There is substantial literature on efficient estimation of θ in CAT applications (van der Linden and Pashley, 2000; Chang and Ying, 2003) and on the problem of classifying examinees as either masters or non-masters in a given content area (Reckase, 1983; Lewis and Sheehan, 1990; Chang, 2004). The latter problem, known as computerized mastery testing (CMT), can be
formalized by setting a cut point $\theta_0$ and defining an examinee as a master if and only if his/her ability level $\theta$ meets or exceeds that cut point.

Typically, a CMT assumes a so-called "indifference region" $(\theta_-, \theta_+)$ containing $\theta_0$, which may be thought of as the ability values which are close enough to the cut point that neither a decision of mastery nor a decision of non-mastery would result in a serious error. The statistical hypothesis of mastery is then given by $H_0 : \theta \geq \theta_+$, while the hypothesis of non-mastery is given by $H_1 : \theta \leq \theta_-$. In a CMT, it is often the case that an examinee can be quickly identified as a master or a non-master if that examinee's ability is substantially higher or lower than the cut point. Therefore, CMT often involves variable-length testing whereby the number of items administered varies by examinee. An important goal in CMT is to strike a balance between the confidence of a correct decision and the economy of the number of items administered. There are thus two essential components of any CMT: (i) the stopping rule that determines when to cease testing and make a classification decision; (ii) the method used to select items adaptively based on an examinee's item response pattern.

The sequential probability ratio test (SPRT; Wald, 1947) has been studied as a candidate stopping rule (Spray and Reckase, 1996; Eggen, 1999; Vos, 2000; Chang, 2004) for CMT. The SPRT has shorter average test lengths than fixed-length tests with the same type I and II error rates at two specific points along the $\theta$ scale. Although it has shorter average length, the SPRT does not constrain the maximum number of items administered. For a test to have no more than $N$ items, it is necessary to use a truncated SPRT (TSPRT), which halts testing and makes a classification decision once $N$ items have been administered. Suppose that $k$ items have been presented to an examinee, yielding the responses $u_1, \ldots, u_k$, where

$$u_i = \begin{cases} 
1, & \text{if the examinee answers the $i$th item presented correctly} \\
0, & \text{if the examinee answers the $i$th item presented incorrectly.}
\end{cases}$$

(1)

The classical theory of the SPRT assumes independence of responses so that the likelihood of $\theta$ is

$$L_k(\theta) = \prod_{i=1}^{k} [p_i(\theta)]^{u_i} [1 - p_i(\theta)]^{1-u_i},$$

where $p_i(\theta) = P\{u_i = 1\}$ for an examinee of ability $\theta$. The SPRT stops after the $k$th item and rejects $H_0 : \theta \geq \theta_+$ if

$$\log \frac{L_k(\theta_-)}{L_k(\theta_+)} \geq A,$$

(2)

or accepts $H_0$ if

$$\log \frac{L_k(\theta_-)}{L_k(\theta_+)} \leq -B,$$

(3)
where \( A, B > 0 \) are chosen so that \( P_{\theta+} \{ \text{reject } H_0 \} = \alpha \) and \( P_{\theta-} \{ \text{accept } H_0 \} = \beta \). Wald’s (1947) approximation yields

\[
A = \log((1 - \alpha)/\beta), \quad B = \log((1 - \beta)/\alpha).
\] (4)

The TSPRT stops with (2) or (3) for \( k < N \), and if stopping does not occur with the \((N-1)\)st item, it rejects \( H_0 \) if and only if

\[
\log \frac{L_N(\theta_-)}{L_N(\theta_+)} \geq C.
\] (5)

For the TSPRT, Spray and Reckase (1996) and Eggen (1999) still use (4) for the values of \( A \) and \( B \) and use for (5) the value

\[
C = (A - B)/2.
\] (6)

The motivation for (5) and (6) is that all examinees classified as non-masters at the \( N \)th item have a log-likelihood ratio no further from \( A \) than \(-B\), and those classified as masters have a log-likelihood ratio no further from \(-B\) than \( A \). Since (4) is based on the error rates of the untruncated SPRT, the true error rates of the truncated procedure, whose decision at truncation is given by (5) and (6), are often substantially inflated (see Table 2 below). This is of particular concern in CMT, where \( \alpha \) represents the percentage of proficient examinees who are failed.

We address herein this problem by using a new class of stopping rules, recently introduced in the sequential analysis literature for testing the composite hypotheses \( H_0 \) versus \( H_1 \) subject to type I and II error probability constraints and a prescribed maximum number of observations. These tests use the generalized likelihood ratio (GLR) statistics instead of simple likelihood ratios and have been shown to have certain optimality properties when the observations are independent and identically distributed (i.i.d.) and whose common distribution belongs to an exponential family. In a CAT, the successive responses \( u_1, u_2, \ldots \) of an examinee, however, are not identically distributed and may not even be independent if the items are chosen adaptively, since most CATs choose the next item to be an unused item in the available item pool according to some criterion. This is also another reason besides the truncation issue why the theory of the SPRT is not applicable to CMTs. We show in Section 2.2 that modern sequential testing theory can in fact accommodate this adaptive feature in sequential experimentation in addition to providing efficient stopping and terminal decision rules. In fact, the methodology developed in Section 2, which is illustrated by applications to CMTs, is applicable to a large variety of psychometric tests, allowing sequential choice of experiments (items in the CMT context) and providing a powerful test at
the conclusion of the study that satisfies the prescribed type I error probability constraint and whose expected sample size is nearly optimal and can be considerably smaller than the prescribed maximum sample size.

This paper is organized as follows. Section 2 first gives a review of recent developments in sequential GLR tests of composite hypotheses based on i.i.d. observations from an exponential family. Then the i.i.d. assumption is removed and the theory is extended to the case where experiments are chosen adaptively to generate an observation (response) at the next stage. The methodology is then applied to the design of efficient CMTs, in which the sequential choice of experiments corresponds to sequential selection of items to be administered to an examinee based on item response theory. Section 3 reports simulation studies of the performance of the proposed CMT and compares it with commonly-used fixed-length tests and TSPRTs. Section 4 gives some concluding remarks.

2. Modern Sequential Methods and Their Applications to CMT

2.1 Efficient Sequential GLR Tests for I.I.D. Observations

To summarize recent advances in sequential hypothesis testing in a general framework that is applicable to psychometric testing including CMTs, let \( X_1, X_2, \ldots \) be i.i.d. observations from an exponential family of densities \( f_\theta(x) = e^{\theta x - \psi(\theta)} \) and let \( L_k(\theta) \) denote the likelihood

\[
L_k(\theta) = \prod_{j=1}^{k} f_\theta(X_j).
\]

The SPRT, which uses the simple likelihood ratio \( \log(L_k(\theta_-)/L_k(\theta_+)) \) to test the hypotheses \( H_0 : \theta \geq \theta_+ \) versus \( H_1 : \theta \leq \theta_- \), is only optimal in the rare case that \( \theta \) is exactly \( \theta_- \) or \( \theta_+ \) and has to allow the possibility of many more than \( N \) observations being taken. A powerful technique in modern sequential analysis that allows the type I error probability to be controlled while having a maximum sample size \( N \) and preserving asymptotic optimality over the entire parameter space (instead of just at \( \theta_+ \) or \( \theta_- \)) is the modified Haybittle-Peto test (Lai and Shih, 2004). Let \( \hat{\theta}_k \) denote the maximum likelihood estimator (MLE) of \( \theta \) based on \( X_1, \ldots, X_k \). The modified Haybittle-Peto test involves replacing the simple likelihood ratios in (2), (3), and (5) by the GLR statistic \( L_k(\hat{\theta}_k)/L_k(\theta') \), which "self-tunes" to information about the true \( \theta \) accumulating in \( \hat{\theta}_k \) over the course of the test and in which \( \theta' \) denotes the appropriate alternative that will be specified below. Lai and Zhang (1994) and Lai (1997, 2001) have shown that sequential GLRs are efficient in many testing problems when the thresholds (e.g., \( A, B \) in (2), (3)) are appropriately adjusted, even when \( \theta \) is
multidimensional. However, the distribution of the GLR is generally more complicated than the simple likelihood ratio, and the classical approximations (4) do not apply. But with the modern computing power that is readily available to practitioners, Monte Carlo simulation or recursive numerical methods are viable and often the preferred methods for computing the thresholds, especially in light of the inflated error probabilities that result from using classical approximations with truncated tests; see Jennison and Turnbull (2000, Chapter 19) and Lai and Shih (2004).

The modified Haybittle-Peto test of the hypotheses $H_0 : \theta \geq \theta_+ \text{ and } H_1 : \theta \leq \theta_-$ can be described as follows. If $N$ is the maximum number of observations and $\alpha, \beta$ are the desired type I and II error probabilities, then there is a value $\theta^{(N)}_\alpha < \theta_+$ such that the likelihood ratio test of $\theta = \theta_+$ versus $\theta = \theta^{(N)}_\alpha$ based on $N$ observations has type I and II error probabilities $\alpha$ and $\beta$; in this sense $\theta^{(N)}_\alpha$ is referred to as the implied alternative. Note that $\theta^{(N)}_\alpha$ is not necessarily equal to $\theta_-$, but it is the appropriate alternative to consider given the parameters $N, \alpha, \beta$, and $\theta_+$. In addition, focusing on the implied alternative $\theta^{(N)}_\alpha$ frees us from having to specify the alternative $\theta_-$, which is often chosen arbitrarily in practice. Let $0 < \rho < 1$. For $\rho N \leq k < N$, the modified Haybittle-Peto test stops after the $k$th item and rejects $H_0$ if

$$\hat{\theta}_k \leq \theta_+ \text{ and } \log \frac{L_k(\hat{\theta}_k)}{L_k(\theta_+)} \geq A,$$

or accepts $H_0$ if

$$\hat{\theta}_k > \theta^{(N)}_\alpha \text{ and } \log \frac{L_k(\hat{\theta}_k)}{L_k(\theta^{(N)}_\alpha)} \geq B,$$

for some constants $A$ and $B$. For $k = N$, the test is always terminated, with $H_0$ rejected if and only if

$$\hat{\theta}_N < \theta_+ \text{ and } \log \frac{L_N(\hat{\theta}_N)}{L_N(\theta_+)} \geq C$$

for some constant $C$. If both (7) and (8) hold for some $k$ (which can only happen when $A$ and $B$ are artificially small), then either decision can be made, for example, always accepting $H_0$ or deciding based on $\hat{\theta}_k$. In CMT, where the false negative rate is critical, a simple approach is to classify as proficient, i.e., accept $H_0$, when this occurs; we take this as the definition here.

Next the thresholds $A, B, \text{ and } C$ are chosen so that the false negative error rate does not exceed $\alpha$ and the false positive error rate, at the alternative $\theta^{(N)}_\alpha$ implied by the maximum
number \( N \) of observations, is close to \( \beta \). Specifically, \( A, B, \) and \( C \) will be chosen so that

\[
P_{\phi^N} \{(8) \text{ occurs for some } k < N\} = \varepsilon \beta, \tag{10}
\]
\[
P_{\phi^N} \{(7) \text{ occurs for some } k < N, (8) \text{ does not occur for any } j \leq k\} = \varepsilon \alpha, \tag{11}
\]
\[
P_{\phi^N} \{(7), (8) \text{ do not occur for any } k < N, (9) \text{ occurs}\} = (1 - \varepsilon) \alpha \tag{12}
\]
for some \( 0 < \varepsilon < 1 \). In practice any value of \( \varepsilon \) giving a test with desirable properties can be used, and Lai and Shih (2004) have shown that values \( 1/3 \leq \varepsilon \leq 1/2 \) work well in a variety of settings. The values of \( A, B, \) and \( C \) that satisfy (10)-(12) can be determined by Monte Carlo simulation, or by numerical methods based on the following normal approximation to the log-likelihood ratios: When \( \theta \) is the true parameter,

\[
Z_k = \text{sign}(\hat{\theta}_k - \theta) \left\{ 2k \log \frac{L_k(\hat{\theta}_k)}{L_k(\hat{\theta})} \right\}^{1/2} \approx N(0, k) \tag{13}
\]
for large \( k \), with independent increments \( Z_k - Z_{k-1} \) (with \( Z_0 = 0 \)). The normal approximation (13) suggests replacing the signed-root statistic \( Z_k \) by a sum of independent standard normal random variables \( S_k = Y_1 + \cdots + Y_k \sim N(0, k) \) so that, for example, the condition (7) becomes \( S_k/\sqrt{k} \leq -\sqrt{2}A \). Then, in place of (10)-(12), \( B, C, \) and \( A \) can be successively found by solving

\[
P\{S_k/\sqrt{k} \geq \sqrt{2}B \text{ for some } k < N\} = \varepsilon \beta \tag{14}
\]
\[
P\{S_k/\sqrt{k} \leq -\sqrt{2}A \text{ for some } k < N\} = \varepsilon \alpha \tag{15}
\]
\[
P\{S_k/\sqrt{k} > -\sqrt{2}A \text{ for all } k < N, S_N/\sqrt{N} \leq -\sqrt{2}C\} = \varepsilon(1 - \alpha) \tag{16}
\]
The left hand sides of (14)-(16) can be computed by recursive one-dimensional numerical integration; see Jennison and Turnbull (2000, Chapter 19) for a more detailed discussion.

Closed-form approximations to the probabilities in (10)-(12) have been developed by Siegmund (1985, Chapter 4) to compute them approximately without using Monte Carlo or numerical integration. Letting \( \phi \) and \( \Phi \) be the standard normal density and c.d.f. and \( m_0 \) the smallest integer \( \geq \rho N \), the normal approximation (13) used in conjunction with Siegmund's (1985) boundary crossing probability approximation yields

\[
(\sqrt{2}A - 1/\sqrt{2}A)\phi(\sqrt{2}A) \log(N/m_0) + 4\phi(\sqrt{2}A)/\sqrt{2}A \tag{17}
\]
as an approximation to (11), and

\[
1 + \phi(\sqrt{2}A) \log(N/m_0)/\sqrt{2}A + 4\phi(\sqrt{2}A)/\sqrt{2}A - 2\Phi(-\sqrt{2}C) + \sqrt{2}A\phi(\sqrt{2}A) \log(C/A) \tag{18}
\]
as an approximation to (12). The values of $A$ and $C$ can therefore be determined by first setting (17) equal to $\epsilon \alpha$ and solving numerically, and then setting (18) equal to $(1 - \epsilon)\alpha$ and solving numerically. Replacing $A$ by $B$ in (17) yields an analogous approximation for the probability in (10), which can be solved numerically to find $B$.

The modified Haybittle-Peto test with thresholds $A, B, C$ satisfying (10)-(12) has type I error rate $\alpha$ and never takes more than $N$ observations. It has asymptotically the smallest possible sample size of all tests with the same or smaller type I and II error probabilities. This was proved by Lai and Shih (2004, Theorem 2(i)) in the context of group sequential tests, and their proof can also be used to establish the following “fully sequential” version, of particular interest in CAT.

**Theorem 1.** Let $0 < \rho < 1$, and let $X_1, X_2, \ldots$ be i.i.d. observations from an exponential family with parameter $\theta$. Let $T_{\alpha, \beta, N}$ be the class of all tests of $H_0 : \theta \geq \theta_+$ taking no more than $N$ but no fewer than $\rho N$ observations and with error probabilities not exceeding $\alpha$ and $\beta$ at $\theta = \theta_+$ and $\theta_-(N)$, the alternative for which the likelihood ratio test of $\theta = \theta_+$ versus $\theta = \theta_-(N)$ based on $N$ observations has type I and II error probabilities $\alpha$ and $\beta$. If $M$ is the sample size of the modified Haybittle-Peto test, then as $\alpha \to 0$ and $\beta \to 0$ such that

$$log \alpha \sim log \beta,$$

for all $\theta$.

$2.2 \text{ Extension to Sequentially Generated Experiments}$

The primary motivation behind CAT is to reduce the length of the test by adaptively creating a test better suited to the individual examinee (see Bickel, Buyske, Chang, and Ying, 2001). This is accomplished by choosing an examinee’s $(k + 1)$st test item based on his/her previous responses $u_1, \ldots, u_k$. Hence the responses are no longer i.i.d., violating a basic assumption in Theorem 1 and also in the optimality theory of the SPRT.

Consider a set $J$ of experiments initially available, with $N \leq |J|$ (the number of elements of $J$). Let

$$f_{\theta, j}(x) = e^{\theta x - \psi_j(\theta)}, \quad j \in J,$$

(20)

denote the density of an observation coming from experiment $j$. Let $j_i$ denote the $i$th sequentially chosen experiment, which is selected according to some rule that involves only
the previous observations $X_1, \ldots, X_{i-1}$. The likelihood function still has the form

$$L_k(\theta) = \prod_{i=1}^{k} f_{\theta, j_i}(X_i)$$

since the density function of $X_i$ given $X_1, \ldots, X_{i-1}$ is $f_{\theta, j_i}$. Hence the computation of the error probabilities, and therefore also of the thresholds $A, B, C$, for the modified Haybittle-Peto test in the present case can proceed in the same way as in Section 2.1, in which $|J| = 1$ and $f_{\theta, j_i}(X_i)$ is simply $f_0(X_i)$.

**Theorem 2.** Suppose that experiments are sequentially chosen from (20) by a rule such that at stage $k$, the choice of $j_k$ depends only on $X_1, \ldots, X_{k-1}$, and that

$$\nu_j = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} P\{j_i = j\} \quad \text{exists for every } j, \quad (21)$$

$$\inf_{|\theta| \leq a} \sum_{j \in J} \nu_j \psi_j''(\theta) > 0 \quad \text{for all } a > 0. \quad (22)$$

Then (19) still holds for all $\theta$, where $M$ is the sample size of the modified Haybittle-Peto test and $T_{a, \beta, N}$ is the class of tests described in Theorem 1 that use this rule to select experiments at every stage prior to stopping.

The proof of Theorem 2 is given in the Appendix, which also gives the asymptotic theory of the MLE and GLR statistics in sequentially generated experiments from (20) under the assumptions (21) and (22). This theory allows us to use the approximation (13) to compute the probabilities in (10)-(12) and thereby determine the thresholds $A, B, C$ of the modified Haybittle-Peto test for the general exponential family considered here.

### 2.3 Application: Efficient Design of CMT

To apply Theorem 2 to the design of efficient CMTs, we use item response theory (IRT) to model the probability $p_j(\theta)$ that an examinee of ability $\theta$ gives the correct answer to item $j$. IRT is traditionally utilized in CMT to provide methods for adaptive item selection as well as to estimate and compare the respective abilities of examinees who were administered distinct sets of items. We assume in the sequel the three-parameter logistic (3-PL) model (Lord, 1980):

$$p_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta - b_j)}}, \quad (23)$$

with known parameters $(a_j, b_j, c_j)$ for all items $j$ in the available item pool.
Any CMT must have an item selection rule as well as a stopping rule. This item selection rule is adaptive in the sense that the choice of the \( k \)th question for an examinee depends on \( u_1, \ldots, u_{k-1} \), where the \( u_i \) are defined in (1) so that \( P_{\theta_j} \{ u_i = 1 \mid u_1, \ldots, u_{i-1} \} = p_j(\theta) \), in which \( j_i \) denotes the item chosen for the \( i \)th question. Most item selection rules in the literature maximize some index of psychometric information at a specified value of \( \theta \) to select the next item for a given examinee. One such index is the Kullback-Leibler (KL) information, which for the 3-PL model (23) is

\[
I_j(\theta, \theta') = p_j(\theta) \log \frac{p_j(\theta)}{p_j(\theta')} + [1 - p_j(\theta)] \log \frac{1 - p_j(\theta)}{1 - p_j(\theta')}.
\]  

(24)

The KL information \( I_j(\theta, \theta') \) is a measure of the distinguishability of the true ability level \( \theta \) from level \( \theta' \) provided by item \( j \). Another such measure used in CMT is the Fisher information, which for the 3-PL model is

\[
I_j(\theta) = \frac{a_j^2(1 - c_j)}{(c_j + e^{a_j(\theta - b_j)}(1 + e^{-a_j(\theta - b_j)})^2).
\]

Reckase (1983), Lewis and Sheenan (1990), Spray and Reckase (1996), and Chang and Ying (2003) use procedures that choose the next item in a test to be the unused item that maximizes the Fisher information at the cut point \( \theta_0 \) or at a current estimate of \( \theta \), like the MLE \( \hat{\theta}_k \). Spray and Reckase (1996) suggest maximizing information at \( \theta_0 \) rather than \( \hat{\theta}_k \) when using the SPRT. Eggen’s (1999) simulations showed that KL information outperforms both of these approaches based on Fisher information in some settings. These adaptive item selection rules satisfy (21)-(22).

3. Simulation Studies

3.1 Simulation of Proposed CMT

In this section we compare the fixed-length, TSPRT, and modified Haybittle-Peto tests of \( H_0 : \theta \geq \theta_+ \) versus \( H_1 : \theta \leq \theta_- \) about the ability level \( \theta \) in the 3-PL model. To isolate the effects of the different stopping rules, all tests use the same criterion – maximum Fisher information – to sequentially choose items. To simulate the tests in a realistic setting, a pool of 1136 real 3-PL item parameters was utilized; see Table 1. This item pool came from the Chauncey Group International, a subsidiary of the Educational Testing Service. The real-life cut point associated with the item pool is \( \theta_0 = -1.32 \). Mimicking simulations by Lin and Spray (2000), \( \theta_- \) and \( \theta_+ \) are taken to be \( \theta_0 \pm .25 = -1.57, -1.07 \). Following Spray and Reckase (1996), \( N \) was set to 50.
As mentioned above, the TSPRT with thresholds (4) and (6) usually has type I and II error probabilities substantially larger than the nominal values $\alpha$ and $\beta$. Using the Chauncey item pool, the type I error probability and the average length of the TSPRT for $\theta = \theta_+$ with $N = 50$ and thresholds (4), (6) is given in Table 2 for various values of $\alpha = \beta$. The actual type I error is roughly constant at about .16 for $\alpha \leq .1$. This is because the thresholds $A = B = \log((1-\alpha)/\alpha)$ are large enough that truncation occurs for nearly every examinee, evident through the large average test lengths, and consequently a large proportion of the examinees are misclassified at the truncation point. Since the type I error probability in CMT is the percentage of proficient $\theta_+$-level examinees who are misclassified as non-proficient, we propose a modification of the TSPRT by choosing $C$ suitably to make the type I error probability approximately equal to the nominal value $\alpha$, rather than use (6). Table 3 contains the average test length and percentage of examinees classified as non-master, i.e., the power, of the following tests at various values of $\theta$: The TSPRT using thresholds (4), (6) with $\alpha = \beta = .05$; the TSPRT, modified in the way described above (denoted by modTSPRT), with the same values of $A = B = \log(.95/.05)$ but with $C = 1.4$ to give type I error $\alpha = .05$; the fixed-length test with $N = 50$, using classification rule (5) with $C = 1.28$ that is chosen to give type I error probability $\alpha = .05$; the modified Haybittle-Peto test (denoted by modHP) with $A = 3.7, B = 3.3, C = 1.4$ that are chosen to satisfy (10)-(12) with $\alpha = \beta = .05$, $\varepsilon = 1/2, \rho N = 5$, and $\theta_{\land}^{(N)} = -1.95$ where the fixed-length test has power $1 - \beta = .95$. All four tests choose the next item to be the unused one that maximizes Fisher information at $\theta = \theta_0$ until there are at least one correct and one incorrect answers (so that the MLE is well defined), and at the MLE thereafter. The average test length and power are computed at eleven values of $\theta$ between $-2$ and $-0.5$, including $\theta_0, \theta_+$ (in bold), $\theta_-$, and $\theta_{\land}^{(N)}$ from 10,000 simulated tests each.

The fixed, modTSPRT, and modHP tests have very similar power functions for $\theta \leq \theta_+$. The TSPRT has high power but also greatly inflated type I error probability 16.1%, resulting from the use of the approximations (2), (3) in its definition, as discussed above. The modTSPRT has the same average test length as the TSPRT because they use the same thresholds $A$ and $B$, and both provide savings in test test length over the fixed-length test, particularly at ability levels outside the indifference region ($\theta_-, \theta_+$). The modHP test
provides substantial savings in test length over the fixed length as well as the TSPRT and modTSPRT. The self-tuning nature of the GLR allows modHP to dramatically shorten the tests of proficient examinees ($\theta \geq \theta_+$), for whom modHP is about half the length of modTSPRT. Moreover, the modHP tests are shorter on average even when $\theta = \theta_+$ or $\theta_-$, suggesting that the method of computing thresholds (10)-(12) contributes to its efficiency as well as the use of the GLR statistic.

3.2 Simulation of Proposed CMT with Exposure Control and Content Balancing

Even though the example in Section 3.1 utilizes a real item pool, the tests are compared under somewhat ideal circumstances where items can be selected purely due to their statistical properties. However, since the modified Haybittle-Peto test presented above relies on no specific item selection rule or IRT model, it has the flexibility to incorporate additional constraints on item selection that arise in typical CATs, such as exposure control and content balancing in the choice of items. In this section we illustrate this by presenting a second simulation study comparing the modified Haybittle-Peto test with the TSPRT and fixed length test, all using the following simple method for exposure control and content balancing.

Suppose that the exposure of the items in the pool needs to be controlled so that each item is administered to no more than a proportion $\pi$ of examinees, on average. Suppose also that the content of the test needs to be balanced in the sense that each item in the pool falls into one of $s$ categories, and these categories should be represented approximately in given proportions $q_1, \ldots, q_s$, where $\sum_i q_i = 1$. A simple way of satisfying these constraints when using a test of maximum length $N$ is the following. From each category $i = 1, \ldots, s$, first select the $Nq_i/\pi$ (neglecting rounding) items with the largest Fisher information at the cut-point $\theta_0$, then randomly select $Nq_i$ items from among these, resulting in a new item pool of $\sum_i Nq_i = N$ items, the proportion $q_i$ of which are in category $i$. The chance that a given item in category $i$ appears in the new pool is clearly no greater than $Nq_i/(Nq_i/\pi) = \pi$. If a test that allows early stopping is being used, like the TSPRT or modified Haybittle-Peto test, then the method of spiraling (Kingsbury and Zara, 1989) can be used so that the category proportions are close to $q_1, \ldots, q_s$ even when early stopping occurs; spiraling simply entails choosing at the $(k + 1)$st stage an item from the category $i$ whose proportion in the first $k$ items differs the most from $q_i$.

INSERT TABLE 4 ABOUT HERE

Table 4 contains the average test length and power of the fixed-length ($N = 50$), TSPRT,
modified TSPRT, and modified Haybittle-Peto tests using this method of exposure control and content balancing, for various values of $\theta$ (with $\theta_*$ in bold). For this study, the Chauncey item pool used in Section 3.1 was randomly divided into $s = 3$ “content” categories, $\pi$ was set to .25, and $q_1 = .4$, $q_2 = .3$, $q_3 = .3$. Each entry in Table 4 was computed from 10,000 simulated tests. The fixed-length ($N = 50$) test uses classification rule (5) with $C = 1.33$, chosen to achieve type I error probability about $\alpha = .05$. The TSPRT uses the stopping rule (2)-(6) with $\alpha = \beta = .05$, and the modified TSPRT (denoted by modTSPRT) uses the same values of $A$ and $B$ but with $C = 1.3$ to ensure type I error probability of $\alpha = .05$, as discussed in Section 3.1. The modified Haybittle-Peto test (denoted by modHP) uses $A = 3.7$, $B = 3.8$, $C = 1.47$ that are chosen to satisfy (10)-(12) with $\alpha = \beta = .05$, $\varepsilon = 1/2$, $\rho N = 5$, and $\theta_0(N) = -2.11$, where the fixed-length test has power $1 - \beta = .95$. The tests show very similar relative performance to those in Table 3. The modTSPRT and modHP tests have power functions very similar to the fixed-length test, while the TSPRT is overpowered, including an inflated type I error probability of 19.3% that results from use of the approximations (2), (3), (6) in its definition, as discussed above. The modHP tests are substantially shorter than the TSPRTs for all values of $\theta$ considered, and particularly for $\theta \geq \theta_*$ where the reduction was around 40% to 50%. Note that the tests in Table 4 are less powerful and on average longer than the corresponding ones in Table 3; this is because they do not always choose the most informative item available in order to satisfy the exposure control and content balancing constraints.

4. Conclusion

This paper shows how efficient sequential tests that use “self-tuning” sequential GLR statistics can be extended from the i.i.d. setting to incorporate sequentially designed experiments. The tests are also sufficiently general to handle practical issues that arise in computerized adaptive testing applications, like the method used in Section 3.2 to satisfy the constraints on exposure control and content balancing or the more complex methods proposed by Sympson and Hetter (1985) and Stocking and Swanson (1993). These tests have potential applications in psychometric testing with sequentially generated experimental designs and data-dependent stopping rules, as illustrated in Sections 2.3 and 3 for CMT.
Appendix: Proof of Theorem 2 and Related Asymptotic Theory

In order to prove Theorem 2, we modify the basic arguments of Lai and Shih (2004) that prove Theorem 1, and whose key ingredients are the following.

(a) Hoeffding’s (1960) lower bound for the expected sample size \( E_\theta T \) of a test that has error probabilities \( \alpha \) and \( \beta \) at \( \theta = \theta_+ \) and \( \theta_-(N) \), which simplifies asymptotically to

\[
E_\theta(T) \geq (1 + o(1)) \log \alpha / \max\{I(\theta, \theta_+), I(\theta, \theta_-(N))\}
\]

\[ (25) \]

as \( \log \alpha \sim \log \beta \), where \( I(\theta, \lambda) = E_\theta \{ \log[ f_\theta(X_i) / f_\lambda(X_i)] \} = (\theta - \lambda) \psi'(\theta) - \{ \psi(\theta) - \psi(\lambda) \} \) is the Kullback-Leibler information.

(b) The sample size \( N \) of the fixed-sample-size likelihood ratio test of \( \theta = \theta_+ \) versus \( \theta = \theta_-(N) \) with error probabilities \( \alpha \) and \( \beta \) at \( \theta = \theta_+ \) and \( \theta_-(N) \), which satisfies

\[
N \sim | \log \alpha / I(\theta^*, \theta_+) | \]

\[ (26) \]

as \( \log \alpha \sim \log \beta \), where \( \theta_-(N) = \theta^* < \theta_+ \) is the unique solution of \( I(\theta^*, \theta_+) = I(\theta^*, \theta_-(N)) \). Moreover, \( \max\{I(\theta, \theta_+), I(\theta, \theta_-(N))\} \) attains its minimum at \( \theta = \theta^* \).

(c) \( \lim_{n \to \infty} P \{ \max_{m \leq m \leq n} | \hat{\theta}_m - \theta | \geq \delta \} = 0 \) for every \( \delta > 0 \).

To extend this to Theorem 2, we need analogs of (a), (b), and (c) to hold for the case of sequentially generated experiments. Without assuming the \( X_i \) to be independent, Lai (1981, Theorem 2) has derived a Hoeffding-type lower bound which in our case takes the form

\[
E_\theta(T) \geq (1 + o(1)) \log \alpha / \max \left\{ \sum_{j \in J} \nu_j I_j(\theta, \theta_+), \sum_{j \in J} \nu_j I_j(\theta, \theta_-(N)) \right\},
\]

\[ (27) \]

where \( \nu_j = \lim_{n \to \infty} n^{-1} \sum_{i=1}^n P \{ j_i = j \} \) exists by (21) and

\[
I_j(\theta, \lambda) = (\theta - \lambda) \psi'_j(\theta) - \{ \psi_j(\theta) - \psi_j(\lambda) \}.
\]

Lai’s (1981, Theorem 2) bounds are derived for sequential tests of \( H_0 : P = P_0 \) versus \( H_1 : P = P_1 \) with type I and type II error probabilities \( \alpha \) and \( \beta \), based on random variables \( X_1, X_2, \ldots \) from a distribution \( P \) such that \( (X_1, \ldots, X_m) \) has joint density function \( p_m(x_1, \ldots, x_m) \), under the assumptions that for \( k = 0, 1 \),

\[
n^{-1} \log[ p_n(X_1, \ldots, X_n) / p_{n,k}(X_1, \ldots, X_n) ] \ 	ext{converges in probability to} \ \eta_k,
\]

\[ (28) \]

\[
\lim_{n \to \infty} P \left\{ \max_{m \leq n} \log \left[ \frac{p_m(X_1, \ldots, X_m)}{p_{m,k}(X_1, \ldots, X_m)} \right] \geq (1 + \delta)n\eta_k \right\} = 0 \ \text{for every} \ \delta > 0,
\]

\[ (29) \]
where \( p_{n,k} \) denotes the joint density function under \( H_k, k = 0, 1 \). These conditions can be shown to hold in the present case, for which
\[
\log L_n(\theta) = \theta \sum_{i=1}^{n} X_i - \theta \sum_{i=1}^{n} \psi_j(\theta), \tag{30}
\]
\[
\eta_0 = \lim_{n \to \infty} n^{-1} \log \frac{L_n(\theta)}{L_n(\theta_+)} = \lim_{n \to \infty} n^{-1} \{ (\theta - \theta_+) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} [\psi_j(\theta) - \psi_j(\theta_+)] \} = (\theta - \theta_+) \sum_{j \in J} \nu_j \psi_j(\theta) - \sum_{j \in J} \nu_j [\psi_j(\theta) - \psi_j(\theta_+)] = \sum_{j \in J} \nu_j I_j(\theta, \theta_+), \tag{31}
\]
and similarly \( \eta_1 = \sum_{j \in J} \nu_j I_j(\theta, \theta_-^{(N)}) \). The uniform convexity assumption (22) can be used in conjunction with (21) to show that (c) still holds in the setting of Theorem 2. Moreover, modification of the proof of Theorem 3 and equation (15) in Lai and Shih (2004) can be used to show that as \( \log \alpha \sim \log \beta, N \sim |\log \alpha|/ \sum_{j \in J} \nu_j I_j(\theta^*, \theta_+) \) analogous to (26), where \( \sum_{j \in J} \nu_j I_j(\theta^*, \theta_+) = \sum_{j \in J} \nu_j I_j(\theta^*, \theta_-^{(N)}) \), and that
\[
E_0 M \sim |\log \alpha|/ \max \left\{ \sum_{j \in J} \nu_j I_j(\theta, \theta_+), \sum_{j \in J} \nu_j I_j(\theta, \theta_-^{(N)}) \right\} \sim \inf_{T \in T_0, \theta, N} E_0 T, \tag{32}
\]
proving Theorem 2.

In the sequel we let \( \theta_0 \) denote the true parameter value to study the asymptotic properties of the MLE \( \hat{\theta}_m \) and the GLR statistics in sequentially designed experiments that satisfy (21) and (22). Strictly speaking, since \( |J| \geq N \) in Theorem 2, we have implicitly assumed for notational simplicity that \( J = J_N \) with \( J_N \subset J_{N'} \) for \( N \leq N' \), so the set \( J \) in (21), (22), (31), and (32) actually refers to \( \lim_{n \to \infty} J_n = \cup_{n=1}^{\infty} J_n \). Note that (c) ensures that with probability approaching 1, \( \hat{\theta}_m \) is near \( \theta_0 \) for all \( \rho n \leq m \leq n \). A standard argument involving martingale central limit theorems (Durrett, 2005, p. 411) and Taylor’s expansion of \( \log L_n(\theta) \), given by (30), around \( \theta_0 \) can be used to show that as \( n \to \infty \)
\[
\left\{ n \sum_{j \in J} \nu_j I_j(\theta) \right\}^{1/2} (\hat{\theta}_n - \theta_0) \text{ has a limiting standard normal distribution,} \tag{33}
\]
and that the signed-root likelihood ratio statistics in (13), with \( \theta \) replaced by \( \theta_0 \), are asymptotically normal with independent increments, generalizing (13) from the i.i.d. case to sequentially generated experiments.
References


Table 1: Five-number summary of Chauncey item pool parameters

<table>
<thead>
<tr>
<th></th>
<th>min</th>
<th>1st quartile</th>
<th>median</th>
<th>3rd quartile</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_j$</td>
<td>0.289</td>
<td>0.683</td>
<td>0.862</td>
<td>1.074</td>
<td>2.372</td>
</tr>
<tr>
<td>$b_j$</td>
<td>-5.531</td>
<td>-1.998</td>
<td>-0.943</td>
<td>0.253</td>
<td>5.426</td>
</tr>
<tr>
<td>$c_j$</td>
<td>0.048</td>
<td>0.211</td>
<td>0.232</td>
<td>0.255</td>
<td>0.529</td>
</tr>
</tbody>
</table>

Table 2: Type I error probability $P_{\theta^+} \{\text{reject } H_0\}$ and average test length $E_{\theta^+}T$ of TSPRT using Chauncey item pool with maximum test length $N = 50$ and thresholds $A = B = \log((1 - \alpha)/\alpha)$, $C = 0$.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = .001$</th>
<th>.005</th>
<th>.010</th>
<th>.050</th>
<th>.100</th>
<th>.200</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\theta^+} {\text{reject } H_0}$</td>
<td>.165</td>
<td>.163</td>
<td>.163</td>
<td>.161</td>
<td>.165</td>
<td>.193</td>
</tr>
<tr>
<td>$E_{\theta^+}T$</td>
<td>50.0</td>
<td>50.0</td>
<td>49.7</td>
<td>44.2</td>
<td>36.7</td>
<td>22.4</td>
</tr>
</tbody>
</table>
Table 3: Average test length and power (in parentheses) of the fixed-length, TSPRT, modified TSPRT (modTSPRT), and modified Haybittle-Peto (modHP) tests using the Chauncey item pool.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>fixed (% of $\theta$)</th>
<th>TSPRT (% of $\theta$)</th>
<th>modTSPRT (% of $\theta$)</th>
<th>modHP (% of $\theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.50</td>
<td>50.0 (0.04%)</td>
<td>27.6 (0.23%)</td>
<td>27.6 (0.04%)</td>
<td>12.6 (0.13%)</td>
</tr>
<tr>
<td>-0.75</td>
<td>50.0 (0.28%)</td>
<td>34.7 (1.53%)</td>
<td>34.7 (0.24%)</td>
<td>15.8 (0.46%)</td>
</tr>
<tr>
<td>-1.00</td>
<td>50.0 (2.70%)</td>
<td>42.6 (10.5%)</td>
<td>42.6 (2.57%)</td>
<td>22.1 (3.27%)</td>
</tr>
<tr>
<td>$\theta_+$ = -1.07</td>
<td>50.0 (5.00%)</td>
<td>44.2 (16.1%)</td>
<td>44.2 (5.00%)</td>
<td>24.5 (5.00%)</td>
</tr>
<tr>
<td>-1.25</td>
<td>50.0 (17.5%)</td>
<td>46.5 (39.0%)</td>
<td>46.5 (17.1%)</td>
<td>30.5 (17.1%)</td>
</tr>
<tr>
<td>$\theta_0$ = -1.32</td>
<td>50.0 (25.5%)</td>
<td>46.6 (49.2%)</td>
<td>46.6 (24.0%)</td>
<td>32.4 (25.1%)</td>
</tr>
<tr>
<td>-1.50</td>
<td>50.0 (51.7%)</td>
<td>44.2 (75.6%)</td>
<td>44.2 (49.7%)</td>
<td>35.0 (49.1%)</td>
</tr>
<tr>
<td>$\theta_-$ = -1.57</td>
<td>50.0 (62.9%)</td>
<td>42.3 (83.3%)</td>
<td>42.3 (60.3%)</td>
<td>34.7 (59.4%)</td>
</tr>
<tr>
<td>-1.75</td>
<td>50.0 (83.3%)</td>
<td>36.3 (94.8%)</td>
<td>36.3 (82.6%)</td>
<td>30.5 (80.4%)</td>
</tr>
<tr>
<td>$\theta_-(N) = -1.95$</td>
<td>50.0 (95.0%)</td>
<td>29.3 (99.0%)</td>
<td>29.3 (93.5%)</td>
<td>23.6 (92.2%)</td>
</tr>
<tr>
<td>-2.00</td>
<td>50.0 (95.4%)</td>
<td>27.8 (99.3%)</td>
<td>27.8 (94.3%)</td>
<td>22.1 (93.2%)</td>
</tr>
</tbody>
</table>

Table 4: Average test length and power (in parentheses) of the fixed-length, TSPRT, modified TSPRT (modTSPRT), and modified Haybittle-Peto (modHP) tests with exposure control and content balancing.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>fixed (% of $\theta$)</th>
<th>TSPRT (% of $\theta$)</th>
<th>modTSPRT (% of $\theta$)</th>
<th>modHP (% of $\theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.50</td>
<td>50.0 (0.03%)</td>
<td>36.7 (0.45%)</td>
<td>36.7 (0.00%)</td>
<td>17.6 (0.14%)</td>
</tr>
<tr>
<td>-0.75</td>
<td>50.0 (0.43%)</td>
<td>42.1 (2.71%)</td>
<td>42.1 (0.31%)</td>
<td>21.7 (0.64%)</td>
</tr>
<tr>
<td>-1.00</td>
<td>50.0 (3.04%)</td>
<td>46.7 (14.1%)</td>
<td>46.7 (3.27%)</td>
<td>28.0 (3.12%)</td>
</tr>
<tr>
<td>$\theta_+$ = -1.07</td>
<td>50.0 (5.00%)</td>
<td>47.5 (19.2%)</td>
<td>47.5 (5.00%)</td>
<td>29.8 (5.00%)</td>
</tr>
<tr>
<td>-1.25</td>
<td>50.0 (14.7%)</td>
<td>48.4 (39.8%)</td>
<td>48.4 (15.4%)</td>
<td>34.3 (13.9%)</td>
</tr>
<tr>
<td>$\theta_0$ = -1.32</td>
<td>50.0 (21.1%)</td>
<td>48.3 (49.1%)</td>
<td>48.3 (21.6%)</td>
<td>35.7 (20.3%)</td>
</tr>
<tr>
<td>-1.50</td>
<td>50.0 (40.9%)</td>
<td>47.0 (71.8%)</td>
<td>47.0 (42.9%)</td>
<td>38.1 (38.0%)</td>
</tr>
<tr>
<td>$\theta_-$ = -1.57</td>
<td>50.0 (50.1%)</td>
<td>46.1 (78.9%)</td>
<td>46.1 (51.4%)</td>
<td>37.9 (47.5%)</td>
</tr>
<tr>
<td>-1.75</td>
<td>50.0 (72.0%)</td>
<td>42.7 (91.1%)</td>
<td>42.7 (73.0%)</td>
<td>36.4 (67.3%)</td>
</tr>
<tr>
<td>-2.00</td>
<td>50.0 (91.2%)</td>
<td>35.9 (98.3%)</td>
<td>35.9 (91.7%)</td>
<td>30.2 (86.2%)</td>
</tr>
<tr>
<td>$\theta_-(N) = -2.11$</td>
<td>50.0 (95.0%)</td>
<td>33.1 (99.2%)</td>
<td>33.1 (95.3%)</td>
<td>27.3 (91.2%)</td>
</tr>
</tbody>
</table>