Application of Sequential Probability Ratio Test to Computerized Criterion-Referenced Testing

Yuan-chin Ivan Chang
Institute of Statistical Science
Academia Sinica
Taipei, Taiwan, 115

March 10, 2003

Abstract

The adaptive testing is an important testing method in the modern educational/psychological testings. In adaptive mastery testing, the items are selectly adaptively according to the estimated information of the unknown latent ability levels, and given then to the test-takers, sequentially. Hence, the decision (master or non-master; pass or fail) for each test-taker is made sequentially based on each test-taker's responses to a particular sequence of items administered to him/her. Thus, statistically speaking, by the natural character of the adaptive mastery testing, it is a sequential problem with dependent observations. The Wald's (1947) SPRT (sequential probability ratio test) has been applied to this kind of mental testing problem by many researchers in the field of educational/psychological measurement theory; for example, Reckase (1978, 1983), Kingsbury and Weiss (1983) and Spray (1993). Most of their results are empirical studies of the performance of SPRT with different item selection schemes. In statistical literature, the SPRT with iid observations have been intensively studied by many statistician since Wald (1947). In this paper, we concentrate on the properties of stopping time of the SPRT under adaptive mastery testing situation; i.e. the test items (observations) are adaptively selected. Not only there are only few people discuss the properties of SPRT under non-iid setup, but also most of them are just large sample properties, which provide very little information for test-makers to design good mastery tests. Without independence property of the observations, it will require different approaches to analyze its performance. Here we apply some results of linear growth processes and a martingale extension of the Wald's equation to obtain a bound of the expectation of the stopping rules (i.e. the test length) used in adaptive mastery tests.

1 Introduction

In many applications of testing technology, such as licensure examination, require a testing device to classify the test-taker into one of the pre-defined categories according to the measurement of his/her skill level. In general, these kinds of tests are called the criterion-referenced tests. In some educational/psychological tests, we might just want to classify the test-takers into one of two categories; that is past/fail, master/nonmaster, certified/noncertified, etc. This kind of test is usually called the mastery test, as a special case of the criterion-referenced tests.

From test administrators' point of view, it is beneficial to have a test scheme such that the decision can be made precisely and efficiently. Researchers in educational/psychological testing believe that to construct an efficient test, the items must be selected adaptively, according to each test-taker's latent trait level (see Lord (1970), Lord (1980), Reckase (1983)). How to apply the idea of tailored/adaptive testing and the sequential methods of statistical analysis to the criterion-referenced (or mastery) testings has been studied by many people; for example, Load (1971a), Lord(1971b), Epstein (1978), Reckase(1978), Kingsbury and Weiss (1983), Reckase (1983), Spray (1993), Spray and Reckase (1996), Wainer (2000), etc. Among them, the Wald's(1947) sequential probability ratio test(SPRT) was one of the most popular statistical tools used and discussed in the literature.

If the idea of the tailored test is used, then the items selected for each test-taker will be all different, which will depend on his/her unknown latent trait level as well as the corresponding responses to the given items. Hence, to implement the tailored mastery testing will require an estimate of the unknown latent trait level. In addition, if the test items are selected adaptively (i.e. depending on each test-taker's responses to its own selected items), then we will no longer have independent observations. The traditional analysis tools used in Wald's(1947) SPRT are mostly based on independent assumptions, and might not be applicable to the current problem. That's why the statistical analysis

of adaptive mastery testing is more complicated than that of the traditional SPRT (see Reckase (1983)).

In Reckase (1983) and Spray (1993), they have done some numerical studies of the performance of the SPRT under tailored test set-up for both the mastery and the multiple-category criterion-referenced tests. For non-iid case, Lai (1981) has shown us a large sample result for the SPRT, when the observations satisfy a "slowly changing sequence" condition. It is an asymptotical result and provide only little information for how to make an efficient mastery test.

In this paper, we study the average of sample number (ASN) (i.e. the expected test length of mastery tests) of the SPRT under the adaptive item selection setup. We found that the expected sample size (or test length) can be bounded by applying a theory of Alsmeyer (1987) for the linear growth processes and an extension of Wald's identity for the martingale differences (see Chow and Teicher (1997)), which could provide more useful information to the test-makers for designing more efficient adaptive mastery tests.

2 Sequential Mastery Testing

The SPRT was initially developed by Wald (1947) for quality control problems during World War II. It has many extensions and applications; such as in clinical trial and in quality control. The original development of the SPRT is used as a statistical device to decide which of two simple hypotheses is more correct. The properties of it have been studied intensively by many researchers since Wald (1947). We will refer readers to Siegmund (1985) and Ghosh and Sen(1991) for more detail discussions. Here, we will only consider its application to the item response theory (IRT) based sequential-mastery testing procedures with adaptive item selection schemes (see Reckase(1983) and Kingsbury and Weiss (1983)).

For explanation purpose, we will temporally assume that all the items are randomly selected from an item pool. Let $\mathcal{B}$ denote the item bank, and $Y = 1$ (0) be a binary random

variable that denotes the test-taker's response is correct (incorrect) to a given test item $\gamma \in \mathcal{B}$. Assume further that probability of correct answer is

$$P(\theta) = P_\theta = P(Y = 1 | \, \theta, \gamma), \tag{2.1}$$

and $Q_\theta = 1 - P_\theta = P(Y = 0 | \, \theta, \gamma)$, where θ denote the unknown latent trait level of the test-taker.

Remark 2.1 *Note that $P(\theta)$ is sometimes called the Item Characteristic Curve (ICC) in educational/psychological testing. It is reasonable to assume that $P(\theta)$ is monotone in θ for any fixed item; i.e. the larger the latent trait level, the higher the probability of correct answer.*

Suppose that there are n items being administered to the test-taker, then the likelihood for a test-taker with latent trait level θ can be written as

$$L_n(\theta) = L_n(y_1, \cdots, y_n \, | \, \theta) = P_\theta^{\sum_1^n y_i} Q_\theta^{n - \sum_1^n y_i}.$$

Let θ_c be threshold chosen by the test-makers. In order to apply the Wald's SPRT to the mastery testing, we must specify an indifference region around θ_c. (Note that if there is no indifference region, then the SPRT might not stop with positive probability for some θ, which are very close to the cutting point θ_c (see Siegmund, 1985)). Let $[\theta_0, \theta_1]$ be the pre-described indifference region, where θ_0 and θ_1 are two constants such that $\theta_0 < \theta_c < \theta_1$. Then the sequential mastery testing can be formulated as a statistical testing problem of two hypotheses below:

$$H_0 : \theta \leq \theta_0 \text{ vs. } H_1 : \theta \geq \theta_1. \tag{2.2}$$

Let α and $\beta \in (0, 1)$ denote the given type I and type II errors, respectively. Because when the true θ is more extreme than θ_0 and θ_1, the errors α and β are smaller (see Reckase (1983) and Siegmund (1985)). Thus, due to the monotonicity of type I and type II

errors, instead (2.2), we might just consider the hypotheses testing problem of two simple hypotheses:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1. \tag{2.3}$$

By applying Wald's SPRT (1947) (see also Siegmund (1985)), we will reject the null hypothesis ($H_0 : \theta = \theta_0$), if $L_n(\theta_1)/L_n(\theta_0) \leq A$; and reject alternative hypothesis ($H_1 : \theta = \theta_1$), if $L_n(\theta_1)/L_n(\theta_0) \geq B$, where the boundaries A and B above are approximated by following formulas:

$$A \approx \frac{\beta}{1-\alpha} \tag{2.4}$$

and

$$B \approx \frac{1-\beta}{\alpha}. \tag{2.5}$$

If we assume further that the observations are i.i.d. (i.e. the items are selected randomly from an item pool), then it has been shown that the average sample size (ASN) under null and alternative hypotheses, respectively, are

$$E_0[T] = \mu_h^{-1} \left\{ \alpha \log \left(\frac{1-\beta}{\alpha} \right) + (1-\alpha) \log \left(\frac{\beta}{1-\alpha} \right) \right\}, \tag{2.6}$$

and

$$E_1[T] = \mu_h^{-1} \left\{ (1-\beta) \log \left(\frac{1-\beta}{\alpha} \right) + \beta \log \left(\frac{\beta}{1-\alpha} \right) \right\}, \tag{2.7}$$

where $\mu_h = E_h[\log f_1/f_0]$, $h = 0, 1$, and f_h are the probability density function under H_h, for $h = 0$, 1 (see Siegmund (1985), page 13).

But for a sequential tailored mastery testing, the items will be selected for each individual test-taker adaptively, i.e. the items are selected based on the estimate of the unknown latent trait level of each test-taker. Because of this natural character of tailored mastery testing, we are no longer to have independent observations.

It follows from the arguments of Wald (1947) (also see Siegmund (1985)), we already know that the above approximations of the boundaries ((2.4) and (2.5)) do not rely on the random sampling assumption. That is they are still valid boundaries even under the

tailored mastery testing setup. But the other asymptotic properties such as ASN and operation curve (OC), which might rely on the iid assumption, will require different analysis tools(see Reckase (1983)). Therefore, for both theoretical interest and practical purpose, it is important to study the following two questions:

1. Will the sequential procedure eventually stop under tailored-testing setup? (i.e. $P(T < \infty) = 1$?)

If the answer is "Yes", then

2. What is its expected test length? (i.e. $E[T\,|\theta = \theta_h] =?$, for $h = 0, 1$)

In this paper, a bound of ASN is obtained, which can be shown to depend on the item selection scheme, the length of the indifference region, as well as the given type I and II errors. Moreover, it can provide some information for the test-makers to design more efficient tailored mastery tests, and we will use a popular probability model to explain this.

3 Main Results

Suppose that θ_c is the threshold chosen by the test-makers as before. Originally, the classification problem of the mastery testing can be written as the following statistical hypotheses testing problem:

$$H_0 : \theta \leq \theta_c \text{ vs. } H_1 : \theta > \theta_c \tag{3.1}$$

Let the interval $[\theta_0, \theta_1]$ denote the in-difference region (see Reckase(1983)) with $\theta_0 < \theta_c < \theta_1$, where θ_0 and θ_1 are two fixed constants. Then, as discussed in the previous section, the above statistical hypotheses testing problem can be simplified to a hypotheses testing problem of two simple hypotheses:

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1 \tag{3.2}$$

Theorem 3.1 *Let* $[\theta_0, \theta_1]$ *be the indifference region. Suppose the item characteristic curve* $P(\theta)$ *is an increasing function, and differentiable for all* $\theta \in [\theta_0, \theta_1]$. *Then for* $h = 0, 1$, $P_h(T < \infty) = 1$.

Following the notations of Section 1, the *log-likelihood* function of a sample size (number of items) n can be written as

$$\ell_n(\theta) = \sum_{i=1}^{n} \{Y_i \log P_i(\theta) + (1 - Y_i) \log(1 - P_i(\theta))\} \tag{3.3}$$

and, the log-likelihood ratio of θ_1 to θ_0 can be written as

$$\begin{aligned} \lambda_n = \lambda(\theta_0, \theta_1) &= \sum_{i=1}^{n} Y_i \log\left(\frac{P_i(\theta_1)}{P_i(\theta_0)}\right) + (1 - Y_i) \log\left(\frac{1 - P_i(\theta_1)}{1 - P_i(\theta_0)}\right) \\ &= \sum_{i=1}^{n} (Y_i - P_{h,i}) \left[\log\left(\frac{P_i(\theta_1)}{P_i(\theta_0)}\right) - \log\left(\frac{Q_i(\theta_1)}{Q_i(\theta_0)}\right)\right] \\ &\quad + \sum_{i=1}^{n} \left\{P_{h,i} \log\left(\frac{P_i(\theta_1)}{P_i(\theta_0)}\right) + (1 - P_{h,i}) \log\left(\frac{Q_i(\theta_1)}{Q_i(\theta_0)}\right)\right\} \\ &= (I) + (II) \end{aligned} \tag{3.4}$$

where $P_{h,i} = P_i(\theta_h) = E_h[Y_i|\mathcal{F}_{i-1}]$ for $h = 0$, 1, and $i \in N$ (i.e. the conditional expectation of Y_i given $\mathcal{F}_{i-1}$ under either null ($h = 0$) or alternative hypothesis ($h = 1$)).

Note that for any given θ_0 and θ_1, and for $i = 1, 2, \cdots$,

$$\left\{\log\left(\frac{P_i(\theta_1)}{P_i(\theta_0)}\right) - \log\left(\frac{Q_i(\theta_1)}{Q_i(\theta_0)}\right)\right\}$$

is $\mathcal{F}_{i-1}$-measurable, and $\{(Y_i - P_{h,i}),\ i = 1, 2, \cdots\}$ is a sequence of martingale differences with respect to $\mathcal{F}_i$, for both $h = 0$ and 1. This implies that (I) of (3.4) is a martingale with respect to $\mathcal{F}_i$ for each h.

Let the stopping rule T be defined:

$$\begin{aligned} T &= \text{first } n \text{ such that } \lambda_n \leq a \text{ or } \lambda_n > b, \\ &\quad \text{and} \\ &= \infty, \text{ if there is no such } n \text{ exists.} \end{aligned} \tag{3.5}$$

When the sampling stopped, we will accept H_0 if $\lambda_T \leq a$ and accept H_1 if $\lambda_T > b$. Note that the choice of a and b will depend on the type I and type II errors (see (3.6) and the equations below it).

Let the constants α and $\beta \in (0, 1)$ be the Type I and Type II errors as given before. Then $\alpha = P_{H_0}(\lambda_n > b)$, and $\beta = P_{H_1}(\lambda_n \leq a)$. It follows from the same arguments of Wald (1947)(c.f. Siegmund (1985)), and without *iid* assumption the boundaries a and b in the stopping rule T (see 3.5) will be shown to satisfy the following two inequalities:

$$a \leq \log(\frac{\beta}{1-\alpha}) \text{ and } b \leq \log(\frac{1-\beta}{\alpha}). \tag{3.6}$$

These inequalities suggest that we can use

$$a = \log(\frac{\beta}{1-\alpha})$$

and

$$b = \log(\frac{1-\beta}{\alpha})$$

as the approximated boundaries for our test.

Remark 3.1 *The arguments for the above inequalities of the test boundaries only depend on the expected likelihood functions under either null and alternative hypotheses, and do not depend on the i.i.d. assumption (see Wald (1947), Reckase(1983) and Siegmund(1985)). Hence, the same arguments can be applied to the current setup. But the other properties of the non-i.i.d. case, such as the expected sample size, under both null and alternative hypotheses, will require different analysis tools from those used in the i.i.d. case. In the rest of this Section, we will concentrate on the expected sample size of the SPRT under the IRT-based adaptive mastery test.*

By the mean-value theorem,

$$\log\left(\frac{P(\theta_1)}{P(\theta_0)}\right) = (\theta_1 - \theta_0)\frac{P'(\theta^*)}{P(\theta^*)} = (\theta_1 - \theta_0) \cdot (A), (say) \tag{3.7}$$

and

$$\log\left(\frac{Q(\theta_1)}{Q(\theta_0)}\right) = (\theta_1-\theta_0)\frac{Q'(\theta^{**})}{Q(\theta^{**})} = (\theta_1-\theta_0)\frac{-P'(\theta^{**})}{1-P(\theta^{**})} = (\theta_1-\theta_0)\cdot(B), (say) \tag{3.8}$$

where P' and Q' denote the first derivative of P and Q with respect to θ, respectively; and θ^* and θ^{**} are lying on the segment of θ_0 and θ_1. Note that both of them are $\mathcal{F}_{i-1}$-measurable. For each $i \in N$, let

$$Z_i = Z_{h,i} = (Y_i - P_{h,i})[(A)-(B)] + P_{h,i}\cdot(A) + (1-P_{h,i})\cdot(B) \tag{3.9}$$

(For simplicity, the index h in $Z_{h,i}$ might be omitted, when it is clear that we are under either null or alternative hypotheses.)

$$\lambda_n = (\theta_1-\theta_0)\sum_{i=1}^{n} Z_i \tag{3.10}$$

Since $d = \theta_1 - \theta_0 > 0$, the stopping time T can be re-written as

$$T = \inf\{n \geq 1 : \sum_{1}^{n} Z_i \leq a/d \text{ or } \sum_{1}^{n} Z_i > b/d\}. \tag{3.11}$$

As mentioned in Remark 2.1, for the IRT based educational/psychological tests, it is common to assume that the item characteristic curve (ICC) is monotonic, increasing in the latent variable. That is we can assume further that

(A.1) The ICC, $p(t) \in (0,\, 1)$, is a continuous, increasing function for $t \in [\theta_0,\, \theta_1]$.

The (A.1) implies that $p(t)$ is differentiable with $p'(t) > 0$ for all $t \in [\theta_0,\, \theta_1]$.

For a fixed θ, and for $i = 1, 2\cdots$, define

$$f_{i,\theta}(t) = f_i(t) = p_i(t)\log\frac{p_i(t)}{p_i(\theta)} + (1-p_i(t))\log\frac{1-p_i(t)}{1-p_i(\theta)}.$$

Then, when $h = 1$, the (II) of (3.4) can be written as $\sum_{i=1}^{n} f_{i,\theta_0}(\theta_1)$. Similarly, if $h = 0$, the (II) of (3.4) can be written as $\sum_{i=1}^{n}(-1)f_{i,\theta_1}(\theta_0)$. (Note that we will omit the subscript i of f above, when there is no ambiguity.)

By definition, it is clear that $f_\theta(\theta) = 0$, and its derivative with respect to t is

$$\begin{aligned} f'_\theta(t) = \frac{d}{dt} f_\theta(t) &= p'(t)\log(p(t)) + p'(t) - p'(t)\log(p(\theta)) \\ &\quad -p'(t)\log(1-p(t)) - p'(t) + p'(t)\log(1-p(\theta)) \\ &= p'(t)\log\left(\frac{p(t)}{p(\theta)}\right) - p'(t)\log\left(\frac{1-p(t)}{1-p(\theta)}\right). \end{aligned} \tag{3.12}$$

By (A.1), the equation (3.12) implies that

$$\begin{aligned} f'_\theta(t) > 0 \qquad & \text{for all } t > \theta, \\ f'_\theta(t) = 0 \qquad & \text{for all } t = \theta, \text{ and} \\ f'_\theta(t) < 0 \qquad & \text{for all } t < \theta. \end{aligned}$$

It follows, by the mean-value theorem again, that $f(t) = f(\theta) + (t-\theta)f'(t^*) = (t-\theta)f'(t^*)$, where t^* is lying in the line segment of θ and t. Hence, for $\theta_1 > \theta_0$, there exist θ^* and $\theta^{**} \in [\theta_0, \theta_1]$, such that

$$f_{\theta_0}(\theta_1) = (\theta_1 - \theta_0) f'_{\theta_0}(\theta^*) > 0 \tag{3.13}$$

and

$$f_{\theta_1}(\theta_0) = (\theta_0 - \theta_1) f'_{\theta_1}(\theta^{**}) > 0. \tag{3.14}$$

Therefore, $Z_{h,i}$ can be re-written as

$$Z_{h,i} = (Y_i - P(\theta_h))[(A) - (B)] + (-1)^{1-h} f_{i,\theta_{1-h}}(\theta_h) \tag{3.15}$$

This implies that Z_i, for all i, $-E_0 Z_i > 0$ and $E_1 Z_i > 0$ under the null H_0 and the alternative H_1 hypotheses, respectively.

Define two new the stopping times T_0 and T_1:

$$T_0 = \inf\{n \geq 1 : \sum_1^n Z_i \leq a/d\} = \inf\{n \geq 1 : -\sum_1^n Z_i \geq -a/d\}, \tag{3.16}$$

and

$$T_1 = \inf\{n \geq 1 : \sum_1^n Z_i \geq b/d\}. \tag{3.17}$$

(Note that if α and β are both less than 1/2, then $a = \log(\beta/(1-\alpha)) < 0$. So $-a/d > 0$.) Thus, by the definition of T_0, T_1 and T,

$$T = \min(T_0, T_1).$$

From equation (3.12), it is clear that $f'_{\theta_h}(t)$ are bounded for all $t \in [\theta_0, \theta_1]$ and for all h. Therefore, we can assume that for any fixed $\theta_0 < \theta_1$, there exist two constants C_l and C_u such that for all $t \in [\theta_0, \theta_1]$

$$C_l \leq |f'_{\theta_h}(t)| \leq C_u.$$

Since f_i is $\mathcal{F}_{i-1}$-measurable, it follows from (3.13) and (3.14), for $h = 0, 1$

$$d \cdot C_l \leq E_h[(-1)^{1-h} f_{1-h}(\theta_h)|\mathcal{F}_{i-1}] = (-1)^{1-h} f_{1-h}(\theta_h) \leq d \cdot C_u \tag{3.18}$$

It implies that, with probability one, for all h, and for all $i = 1, 2, \cdots$,

$$0 < d \cdot C_l \leq (-1)^{1-h} E_h[Z_i|\mathcal{F}_{i-1}] \leq d \cdot C_u < \infty. \tag{3.19}$$

Now, we can apply Alsmeyer (1987) Proposition 2.1(c) to show that for any $p > 1$ the expected stopping times under null and alternative hypotheses will satisfy the following inequalities:

$$E_0[T]^p = E_0[\min(T_0, T_1)]^p \leq E_0[T_0]^p < \infty \tag{3.20}$$

and

$$E_1[T]^p = E_1[\min(T_0, T_1)]^p \leq E_1[T_1]^p < \infty. \tag{3.21}$$

This implies that $P_h(T < \infty) = 1$ for all h, and it completes the proof of Theorem 3.1.

Remark 3.2 *As long as there is an indifference region, d is always a positive number. The equation (3.20) and (3.21) will imply that the stopping rule will eventually stop. On the other hand,* T_0 *and* T_1 *will both go to infinity as d goes to* 0*. It implies that the expected sample size (or the test length) will also go to infinity as d goes to* 0*. This also means that for practical purpose it is necessary to introduce an indifference region. For more detail discussion, please see Reckase(1983) and Siegmund (1985).*

3.1 Expectation of Stopping Time

Suppose that there is no overshoot, then

$$E_h\left[\sum_{i=1}^{T} Z_i\right] = P_h(\text{ Accept } H_0) * (\frac{a}{d}) + P_h(\text{ Accept } H_1) * (\frac{b}{d}) \tag{3.22}$$

Thus, it follows from (3.22) that if $h = 0$, then

$$K_0 \equiv E_0\left[\sum_{i=1}^{T} Z_i\right] = \alpha \log\left(\frac{1-\beta}{\alpha}\right) + (1-\alpha)\log\left(\frac{\beta}{1-\alpha}\right);$$

and if $h = 1$, then

$$K_1 \equiv E_1\left[\sum_{i=1}^{T} Z_i\right] = \beta \log\left(\frac{\beta}{1-\alpha}\right) + (1-\beta)\log\left(\frac{1-\beta}{\alpha}\right)$$

From (3.15), for all $h = 0, 1$

$$\sum_{i=1}^{T} Z_{h,i} = \sum_{i=1}^{T}(Y_i - P_{h,i})[(A) - (B)] + \sum_{i=1}^{T}(-1)^{1-h} f_{i,\theta_{1-h}}(\theta_h). \tag{3.23}$$

It is known that first summation of the right hand side of (3.23) is a martingale with respect to $\mathcal{F}_i$, and $E_h[(Y_i - P_{h,i})|\mathcal{F}_{i-1}] = 0$, for all i. From (3.20), (3.21) and Alsmeyer (1987), we already have that $E_h[T^p] < \infty$ for $p > 1$, and for all h. Thus, by applying Chow and Teicher (1997), Corollary 11.2.3,

$$E_h\left[\sum_{i=1}^{T}(Y_i - P_{h,i})\right] = E_h[Y_1 - P_{1,h}] = 0 \tag{3.24}$$

Therefore, it follows from (3.18) and (3.23),

$$d \cdot C_l\, E_h[T] \leq E_h[\sum_{i}^{T} Z_i] \leq d \cdot C_u\, E_h[T].$$

It implies that

$$\frac{K_h}{d \cdot C_u} \leq E_h[T] \leq \frac{K_h}{d \cdot C_l} \tag{3.25}$$

The above result can be summarized as following Theorem:

Theorem 3.2 *Suppose the item characteristic curve satisfies (A.1), and the length of indifference region, $d > 0$. Then for all h,*

$$\frac{K_h}{d \cdot C_u} \leq E_h[T] \leq \frac{K_h}{d \cdot C_l}.$$

(The proof follows from previous arguments and will be omitted here.)

Note that the equation (3.12) can be re-written as

$$f'_\theta(t) = P'(t) \log \left(\frac{P(t)/(1-P(t))}{P(\theta)/(1-P(\theta))} \right) \tag{3.26}$$

Thus,

$$f'_{\theta_0}(\theta_1) = P'(\theta_1) \log \left(\frac{P(\theta_1)/(1-P(\theta_1))}{P(\theta_0)/(1-P(\theta_0))} \right) \tag{3.27}$$

The $P'(\theta_1)$ of (3.27) is the "discrimination power" of the item at θ_1. The second term of right hand side of (3.27) is the log-odds-ratio at θ_0 and θ_1. From (3.19), it is obvious that the bounds C_l and C_u exist and depend on this odds-ratio. In addition, it follows from (3.25), the larger the log-odds-ratio of θ_0 and θ_1, the smaller the average of test length (expectation of stopping time). This will provide some information for test-maker to design a more efficient mastery test. In the next section, we will use a 2-PL model as an example to explain it.

4 3-Parameter Logistic Model

Let $Y \in \{0, 1\}$ be a binary random variable as defined in the previous section, and $\mathcal{B}$ denote an item bank. Suppose an item with parameter $\gamma = (a, b, c) \in \mathcal{B}$ is selected. The 3-parameter logistic (3-PL) model:

$$P(Y = 1|\theta, \gamma) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}} \tag{4.1}$$

is a commonly used probability model in mental testing to describe the probability of getting a correct answer ($Y = 1$) for a test-taker with a latent trait level θ for a given item with

parameter $\gamma = (a, b, c)$. The parameters a, b and c are called the discrimination, difficulty, and guessing parameters, respectively. Usually, we will assume that $a > 0$, b belonging to a compact set, and $c \in [0, \epsilon]$ for a small positive number ϵ. If $c = 0$, then it is called a 2-PL model. If we assume further that the a is a positive constant, then it becomes the Rasch model (Rasch (1960)). Our method here can be applied to all these three models. In this section, we will use a 2-PL model (i.e. $c = 0$) for illustration.

If a 2-PL model is used, then the item response curve becomes

$$P(\theta) = P\left(Y = 1 | \, \theta, \gamma = (a, b)\right) = \frac{e^{a(\theta - b)}}{1 + e^{a(\theta - b)}}. \tag{4.2}$$

Then

$$P'(\theta) = \frac{a \, e^{a(\theta - b)}}{(1 + e^{a(\theta - b)})^2}.$$

The log-odds-ratio at θ_0 and θ_1 is

$$\log\left(\frac{P(\theta_1)/1 - P(\theta_1)}{P(\theta_0)/(1 - P(\theta_0))}\right) = a\,(\theta_1 - \theta_0) = a \cdot d \tag{4.3}$$

Hence, there exist two constants C_l and C_u such that for all $\theta \in [\theta_0, \, \theta_1]$

$$C_l \leq P_i'(\theta) \log\left(\frac{P(\theta_1)/1 - P(\theta_1)}{P(\theta_0)/(1 - P(\theta_0))}\right) = \frac{d \, a_i^2 \, e^{a_i(\theta - b_i)}}{(1 + e^{a_i(\theta - b_i)})^2} \leq C_u \tag{4.4}$$

Then, by (3.25), we have both the upper and lower bounds of the expected test length (stopping time). If we assume the discrimination parameter a is bounded (say $m \leq a \leq M$) for all items $\gamma \in \mathcal{B}$ then it becomes

$$\frac{K_h}{d^2 M^2} \left\{ \sup_{b, \theta \in [\theta_0, \, \theta_1]} \frac{e^{m(\theta - b)}}{(1 + e^{m(\theta - b)})^2} \right\}^{-1} \leq E_h[T] \leq \frac{K_h}{d^2 M^2} \left\{ \inf_{b, \theta \in [\theta_0, \, \theta_1]} \frac{e^{M(\theta - b)}}{(1 + e^{M(\theta - b)})^2} \right\}^{-1} \tag{4.5}$$

As mentioned before, in the adaptive sequential mastery testing, items are selected depending on the estimate of the information of the latent variable θ. The equation (4.4) and (4.5) provide us a good information for designing a mastery test. For example, for a given length of indifference region d, if we can choose a_i and b_i such that (4.4) (*or the odds-ratio*) is "maximized", then we will have the shortest test length mastery test.

Remark 4.1 *The probit model is another popular probability model used in mental testing. It assumes that*

$$P(Y = 1|\theta, \gamma) = c + (1 - c) \int_{-\infty}^{a(\theta - b)} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

For the arguments used in previous sections, we only assume the monotonicity and differentiability of the item characteristic curve within the indifference region, which are very common and reasonable assumptions in the mental testing. So, it can be applied to the above probit models as well as many other probability models, which satisfy the assumption above.

References

Alsmeyer, G. (1987). On The Moment of Certain First Passage Times for Linear Growth Processes, Stochastic Processes and Their Applications 25, 109 – 136.

Chow, Y. S. and H. Teicher (1988). *Probability Theory: Independence, Interchangeability, Martingales* Springer–Verlag, New York.

Kingsbury, G. Gade and Weiss, David J. (1983). A Comparison of IRT-Based Adaptive Mastery Testing and a Sequential Mastery Testing Procedure, 258-288, *New Horizons In Testing – Latent Trait Test Theory and Computerized Adaptive Testing*, Ed. David J. Weiss, Academic Press, New York.

Ghosh, B.K. and Sen, P.K. (1991). *Hand book of Sequential Analysis*, Marcel Dekker, Inc., New York.

Lai, T.L. (1981). Asymptotic Optimality of Invarianet Sequential Probability Ratio Tests, *Annals of Statistics*, 9, 318-333.

Load, F. (1970) Some test theory for tailored testing. In W.H. Holzman (Ed.), *Computer-Assisted Instruction, Testing, and Guidance* (pp. 139-183). Harper & Row, New York.

Load, F. (1971a) Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement.* **31**, 3–31.

Load, F. (1971b). A theoretical study of two-stage testing. *Psychometrika* **36**, 227-242.

Load, F. (1980) *Applications of item response theory to practical testing problems*, Lawrence Erlbaum, Hillsdale, New Jersey.

Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests*, Copenhagen: The Danish Institute of Educational Research. (Extended edition, 1980. Chicago: The University of Chicago Press.

Reckase, M. (1983). A procedure for Decision Making Using Tailored Testing, 238-257, *New Horizons In Testing – Latent Trait Test Theory and Computerized Adaptive Testing*, Ed. David J. Weiss, Academic Press, New York.

Spray, J. (1993). Multiple-Category Classification Using a Sequential Probability Ratio Test, ACT Research Report Series 93-7.

Spray, J. and Reckase, M. (1996). Comparison of SPRT and Sequential Bayse Procedures for Classifying Examinees Into Two Categories Using a Computerized Test, Journal of Educational and Behavioral Statistics Winter 1996, V. 21, No. 4, pp 405-414.

Siegmund, D. (1985). *Sequential Analysis*, Springer-Verlag, New York Inc.

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

Wainer, H. (2000) *Computerized Adaptive Testing: A Primer*, 2nd Edition, Lawrence Erlbaum Associates, New Jersey.