Deriving a Stopping Rule for Sequential Adaptive Tests

Draft

Irina Grabovsky, Hua-Hua Chang*

January 13, 2003

1 Introduction

Computerized adaptive testing has become increasingly popular in the large-scale educational testing due to advancement of modern computer technology. The most recognizable distinction between adaptive and the usual computer-based testing is that the selection of test items is tailored to individual examinee's ability level. Items are selected sequentially according to the examinee's performance on the earlier questions. For example, a student with a pattern of successful responses may be presented with the more difficult items in the course of the test. He or she will be exposed to a smaller number of items and potentially finish the test sooner. For the less able examinee, it may not make sense to continue on with the sequence of more difficult questions. Perhaps the discrimination and difficulty parameters should be adjusted according to a certain selection rule. Potentially it could take longer to estimate ability but of course the goal is to make sure

*This research was supported by National Board of Medical Examiners, Project No. CTTR 40200.

that every student's ability is estimated very accurately. If this goal is satisfied, the major advantage of sequential adaptive tests is that it provides more precise ability estimates with fewer items than that required in the conventional tests. However, methodological and theoretical development in CAT presents a number of open problems. One of the areas of interest in adaptive testing is the problem of the length of the test.

For the fixed length CAT, each examinee is offered an individualized test that fits his/her ability the best, but number of items is determined in advance. Most applications implementing adaptive tests are fixed length tests (GRE, GMAT). However, it is desirable to reduce the test length represented by the number of items offered to an examinee while maintaining the same level of precision in measuring examinee's ability trait. Able examinees can avoid responding too many easy items, and less able examinees can avoid being exposed to too many difficult items. In the adaptive test with a variable length, the number of items administered to an examinee is not specified in advance, but rather determined during the course of the test. We consider the problem of deciding on the number of items needed to be offered to an examinee so that his or her ability is measured effectively.

2 Maximizing information approach to the varied length tests

Most existent stopping rules for the varied test length are based on maximizing the Fisher information evaluated at the current estimate of the ability. In particular, setting a threshold to the amount of the information collected (Fisher information number) can serve as an indicator that there had been enough items collected to assess examinee's ability efficiently. The conventional stopping rule is based on the fact that information function is the reciprocal of the asymptotic sampling variance of the estimator [Lord, 1980]:

$$I(\theta) = \frac{1}{Var(\hat{\theta}|\theta)}.$$

Let σ_e be the standard deviation of the measurement of the ability in the process of adaptive test. The requirement of achieving some small standard error σ_e for each examinee can be translated in terms of the information function as the requirement of exceeding certain positive bound:

$$\sigma_e \leq \delta_0$$

is approximately equivalent to

$$I(\theta) \geq 1/\delta_0.$$

For instance, choosing $\delta_0 = 0.1$ as a desired degree of precision translates into comparing information function to the number 100. As pointed out in [H.Wainer, 2000] the additivity of the information can be used to estimate the score precision. We computed the value of the information function $I_j(\theta_j)$ after the item with parameters (a_j, b_j) item had been administered using current MLE ability estimate $\hat{\theta_j}$. The cumulative information function was computed as a sum:

$$I_n = \sum_{j=1}^{n} I_j(\hat{\theta_j}).$$

The information-based stopping rule is a reliable method that guarantees that all examinees are tested to the same level of accuracy. However, the right hand side here may actually exceed the value of the true information function at a stage n due to the measurement error in estimates θ_n. Thus, the method does not give us a feel for how many items in truth were needed to achieve this level of accuracy. Also, this approach is based on the assumption of local independence. As we know, in CAT the selection of the next item is based on the performance of examinee on the previous items [Chang, Ying, 2000]. We need to incorporate the knowledge of the sequential ability estimate in our stopping rule. As Thissen and Mislevy noted in the chapter on testing algorithms for CAT [Wainer, 2000], "simulations may be used to determine in advance how long the test is likely to be for examinees of various levels of proficiency".

3 The logistic regression based Stopping Rule for adaptive testing

3.1 Fixed size confidence region idea.

Our idea is to derive the stopping rule that provides us with the desired confidence level that the sequential estimates for the given item selection mechanism fall into the vicinity of true ability of the given size. Thus we came to formulate the question of finding the right stopping time in the process of ability estimation in terms of the mathematical problem of fixed size confidence interval.

The problems of this kind were considered in various statistical models. The earliest work on sequential estimation in order to reduce the sample size was due to Stein (1945, 1949) who proposed the two-stage estimation procedure for the mean parameter in a normal population with variance unspecified. A fully sequential estimation rule was later developed by Anscombe (1953) and Chow and Robins (1965) which lead to many subsequent investigations. The most relevant work to the logistic IRT model is, perhaps, in Chang and Martinsek (1992) which deals with the fixed-width sequential estimation for the logistic regression model.

The corner stone of our investigation is the fact that model explaining measurements and ability estimation process (Item Response Theory) is closely related to the well studied in statistics multinomial logistic regression model. In the next section we establish the connection between the sequential adaptive testing model and the logistic regression.

3.2 IRT as a logistic regression model

Both theoretical and implementation aspects of adaptive testing largely rely on the item response theory models, which relate examinee's ability to their response to the test items. Suppose that an examinee's ability level can be characterizes by a single parameter θ. A basic assumption of IRT is that for a given item the probability of producing a correct answer depends only on examinee's ability parameter θ. The resulting probability curve , as θ varies, is known as the item characteristic curve of the item. Let Y be the response of the examinee , θ - the true ability. In this paper we consider the so-called 2-PL model whose item characteristic curve is defined by the following equation:

$$P(Y=1|\theta)=\frac{e^{1.7a(\theta-b)}}{1+e^{1.7a(\theta-b)}}.$$

We note that this model is a particular case of the multiple logistic regression model. Indeed, to introduce the p-parameter multinomial logistic model consider the set of independent identically distributed observations Y_i, and the sequence of the so-called risk factors X_i. In our study the analog of X_i is the set of pre-calibrated item parameters. The model considered in statistical literature that is relevant to our IRT model is described by the following equation:

$$P(Y_i = 1|X_i) = \frac{\exp X_i^T \beta}{1 + \exp X_i^T \beta},$$

where β and X_i are p-dimensional vectors, and the linear combination $X_i^T \beta$ is their scalar product. To see that IRT is a particular case of the 2-dimensional logistic model above it is enough to notice that the exponent $1.7a(\theta - b)$ can be presented in the form of the scalar product $X_i^T \beta$, where

$$X_i = (1.7a_i, -1.7a_i b_i), \beta = (\theta, 1).$$

The parameter β is usually estimated via maximization of the likelihood function, and we will also consider the MLE method of sequential estimation of ability.

3.3 Stopping rule for logistic regression as applied to IRT

As we noted, the general problem of constructing the fixed size and prescribed accuracy confidence interval had been studied by several authors. Let d be the desired width of the confidence interval around the true value of θ, and $z_{\alpha/2}$ be the $\alpha/2$ quantile of the normal distribution. For instance, for $\alpha = 0.05$ we get that

$$P(-z_{\alpha/2}, z_{\alpha/2}) = 0.95.$$

Statisticians were interested in precisely how many MLE estimates of the parameter β one has to obtain in order to be confident with probability $1-\alpha$ that the estimate falls into the confidence set of a predetermined size d.

For the purposes of investigating the IRT model, we will use here the stopping rule proven by Y.I. Chang and A. Martinsek (p.1967) for estimating the linear combination of $C^T \beta$. The stopping rule consists of computing the real number r_n for each step n of the MLE estimation process, and comparing it to the bound δ. The number r_n is computed based on the components of the vector X_i,and the current MLE estimate β_n. Precisely, the bound δ is given by

$$\delta = \frac{d^2}{z_{\alpha/2}^2},$$

while the sequence r_n is computed as follows:

$$r_n = C^T \Sigma_n^{-1} C,$$

where the matrix Σ_n is computed as the sum of matrices

$$\Sigma_n = \sum_{i=1}^{n} \frac{\exp X_i^T \beta_n}{1 + \exp X_i^T \beta_n} X_i X_i^T.$$

The rule commands to proceed with the MLE estimation until for the first time

$$r_n \leq \delta.$$

Mathematically, the rule is expressed as follows:

$$T = inf\{C^T \Sigma_n^{-1} C \leq \frac{d^2}{z_{\alpha/2}^2}\}.$$

In the context of IRT model, we want to find out how many MLE estimates of the true ability, and consequently, how many items an examinee has to be presented with, so that we could be sure with probability 0.95 that the final estimate falls into the confidence interval of width d around the true ability.

Recall that for our IRT application we chose $\beta = (\theta, 1)$, so it suits us to choose $C = (1, 0)$. Note that in these notations $C^T \beta_n = \theta_n$ so this brings us directly to computing the number of steps in the process of sequential MLE estimation of θ. In the formula for Σ_n above the components of X_i are $(1.7a_i, -1.7a_i b_i)$, and also $\beta_n = (\theta_n, 1)$. Thus, all parameters of the stopping rule are identified through the individual sequential MLE estimates θ_n and the sequence of item parameters a_i, b_i.

4 Results of the simulation study on the stopping rules

In this project we conducted simulations for two stopping rules: the conventional method, that we call “information based method”, and the “logistic regression rule”. The data set that we had on hands consisted of item parameters (discrimination and difficulty) used for the National Assessment of Educational Progress. The item pool consisted of 252 questions and we ran

Table 1: Information-based method				
Bound for Stand Dev σ	Limit for Information	Mean Stop. Time	Median Stop. Time	Mean Sq. Error
0.3	10	8	6	0.09
0.25	16	12	8	0.05
0.2	25	22	12	0.04
0.15	45	47	27	0.02
0.1	100	127	105	0.01
0.05	400	200+	200+	0.01

adaptive process generating 200 items for each examinee. Thus, we were able to obtain mean square error estimates for the fixed length test. We chose the sample of size 100 from standard normal distribution to represent simulated true ability estimates for hypothetical sample of 100 students. Next, we simulated the sequential adaptive test process for which items are selected based on maximizing the Fisher information. The ability is estimated by maximum likelihood method. For each examinee, we recorded the step number after which the information would reach the set in advance bound. We then compared the number of steps required by each method. The results for the information-based method are presented in the Table 1. A similar series of runs had been conducted for the logistic regression based method. Instead of the desired bound to the standard deviation we impose here the size of the confidence interval we wish ability estimates to fall. We will keep the confidence level at 0.05 throughout these simulations. The results are presented in the Table 2. As we see from the tables, both stopping rules produce a very small mean square error. The smaller the confidence interval size or the smaller the desired bound for standard deviation, the smaller the mean square error is. This is a good evidence that both stopping rules give effective ways to determine the length of the test for each examinee individually.

To be able to compare the effectiveness of both rules, we need to bring the results in both tables to the same scale. The first method is formulated in terms of the standard deviation σ of the ability estimates around the mean, while the second method is formulated in terms of the 95% confidence interval of the size d. As we know, for the normal distribution the 95% confidence interval around the mean μ is approximately equal to $(\mu - 2\sigma, \mu + 2\sigma)$. Thus,

Table 2: ”Logistic regression” stopping rule confidence level = 0.05

d - size of the Confidence Interval	Mean Stopping Time	Median Stopping Time	Mean Square Error
0.6	11	9	0.08
0.5	13	10	0.08
0.4	16	13	0.07
0.3	22	18	0.05
0.2	31	26	0.03
0.1	60	55	0.02
0.01	109	104	0.02

Table 3: Comparison of the effectiveness of the two rules

Pairs σ, d: $d = 2\sigma$	Fisher Information		Logistic Regression	
	Mean Stop. Time	Mean Sq. Error	Mean Stop. Time	Mean Sq. Error
$\sigma = 0.1, d = 0.2$	127	0.01	31	0.03
$\sigma = 0.15, d = 0.3$	47	0.02	22	0.05
$\sigma = 0.2, d = 0.4$	22	0.04	16	0.07
$\sigma = 0.3, d = 0.6$	8	0.09	11	0.08

the symmetric interval of length 4σ around the mean is in fact the 95% confidence interval. So the approximate correspondence that we will establish between the two rules is to compare the number of required steps for σ and d for which the equality $2\sigma = d$ holds. The Table 3 above illustrates the results of comparison. It shows the required number of steps for both methods for two pairs of corresponding values of σ, d.

We also were able to compare the effectiveness of both methods with the fixed length method. The code for our CAT simulation process is designed to compute MSE for the fixed length tests for all consecutive lengths 1 - 200. The Table 4 presents the MSE for the stopping times for the Logistic Regression based method and the MSE for the corresponding fixed length method.

Table 4: ”Logistic regression” stopping rule Versus Fixed Length test			
d - size of the Confidence Interval	Mean Stopping Time	Varied length MSE	Fixed Length MSE
0.6	11	0.08	0.07
0.4	16	0.07	0.05
0.3	22	0.05	0.04
0.2	31	0.03	0.03
0.1	60	0.02	0.02
0.01	109	0.02	0.01

5 Discussion of the results and future direction

As evidenced from the Table 3, to achieve the same degree of accuracy for each individual given confidence level $\alpha = 0.05$, one needs to administer less items according to the ”Logistic regression” based method than if one follows the conventional ”Fisher Information” method. This may suggest that the old method is certainly very accurate method of estimation of the ability but perhaps it requires too many extra items administered. The ”Logistic Regression” rule advises that considerably smaller number of steps is needed to achieve the prescribed degree of accuracy: 11 versus 66, 16 versus 168. The price that one has to pay is reflected in the mean square error term. We see that MSE decreases from 0.08 to 0.02 and from 0.07 to 0.01 if the more conventional method is applied. Intuitively, the MSE is smaller for the old method than for the new one since more items are administered and the ability is estimated more precisely across the student population. However, we believe that MSE are sufficiently small for both methods and one needs to decide whether the decrease in mean square error overweighs the fact more resources are needed to administer the extra items. The smaller number of items would reduce overlap rates significantly and thus increase security of the test.

Comparing the results for the fixed length test and the new stopping rule we should note that both methods are very comparable in terms of the accuracy across the examinee pool. The Table 4 implies that the MSE for the series of varied length tests and the corresponding fixed length tests show a difference of 0.01 in the MSE. This is not really a significant difference in overall accuracy. The existing length of the NAEP test (about 30 items)is in line with the requirement of the size confidence interval of the ability estimate to be about 0.2.

The varied length tests allow for a different number of items to be administered to each individual. Our simulations show a wide range of stopping times depending on the individual ability and the MLE estimation process. If one wants to take advantage of the simpler logistics of administering fixed length test, our new stopping rule can be used as an effective way to predetermine the length of the test.

On the other hand, the stopping rule that we suggest allows to decide on the number of items to be presented to an examinee during the course of the test and guarantees the prescribed accuracy of ability estimation will be achieved. The rule suggests on average fewer items to be administered that the conventional stopping rule.

The limitation of our study is that the rule had been tested on just one dataset, and further simulations are needed to confirm the findings. As a future direction of research, we want to expand the technique to the 3-parameter IRT model. This would require the mathematical derivation of the stopping rule since such model does not exactly falls into the class of logistic regression models. Also, we would like to apply our results to other item selection mechanisms such as expected a posteriori and some stratification methods.

References

Bickel, P. and Doksum, K. Mathematical Statistics, 1977.

Birnbaum, A. Some latent trait models. In F.M.Lord and M.R. Novick, Statistical theories of mental test scores. 1968

Chang, H. & Ying, Z. Nonlinear Sequential Designs for logistic Item Response Theory Models with applications to computerized Adaptive tests. Accepted for Publication in *Annals of Statistics.*

Chang, I. and Martinsek, A. (1992). Fixed size confidence regions for parameters of a logistic regression model. The Annals of Statistics, Vol.20, No.4, 1953-1969.

Chow, Y.S. and Robbins, H. (1965). On the asymptotic theory of fixed-width confidence intervals for the mean. Ann. Math. Statist. Vol.36, 457-462.

Lord, F. Applications of the Item Response Theory to Practical testing Problems.

Stein, C. (1945) A two sample test for linear hypothesis whose power is independent of the variance. Ann. Math. Statist. 16 243-252

Wainer, H. Computerized Adaptive Testing: A primer. 1990 Lawrence Erlbaum Associates, Publishers