DOCUMENT RESUME

ED 346 125 TM 018 393

AUTHOR Kirisci, Levent; Hsu, Tse-Chi
TITLE Estimation of Ability Level by Using Only Observable Quantities in Adaptive Testing.
PUB DATE Apr 92
NOTE 31p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; Bayesian Statistics; Comparative Analysis; *Computer Assisted Testing; Computer Simulation; Difficulty Level; Equations (Mathematics); *Estimation (Mathematics); *Mathematical Models; Monte Carlo Methods; *Predictive Measurement; Scoring; Test Construction; Test Format; Test Items
IDENTIFIERS *Ability Estimates; A Posteriori Index; Flexilevel Computerized Adaptive Testing

ABSTRACT

A predictive adaptive testing (PAT) strategy was developed based on statistical predictive analysis, and its feasibility was studied by comparing PAT performance to those of the Flexilevel, Bayesian modal, and expected a posteriori (EAP) strategies in a simulated environment. The proposed adaptive test is based on the idea of using item difficulty and past information (observed data) about an examinee to acquire the probability of answering further items correctly. Development of the PAT model is described with reference to: (1) initial items; (2) scoring method; (3) selection of subsequent items to be administered; and (4) terminating criteria. The model was compared to the Flexilevel, Bayesian modal, and EAP strategies in a Monte Carlo simulation study in which the ability levels of 999 examinees were generated using a 71-item test. The strategies performed similarly at the low ability level. At the medium level, the Bayesian modal and EAP strategies were the most efficient. At the high level, the Bayesian modal strategy required fewer items than did the PAT and the EAP strategies. The three strategies produced similar results in terms of error variance and ability estimates. The PAT is potentially useful, particularly in small classroom testing. There are 12 tables of study data, 2 figures, and a 14-item list of references. (SLD)

ESTIMATION OF ABILITY LEVEL

BY

USING ONLY OBSERVABLE QUANTITIES

IN ADAPTIVE TESTING

Levent Kirisci

and

Tse-Chi Hsu

University of Pittsburgh

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1992.

ESTIMATION OF ABILITY LEVEL BY USING ONLY OBSERVABLE QUANTITIES IN ADAPTIVE TESTING

The objectives of this study were (a) to develop the predictive adaptive testing (PAT) strategy which was based on statistical predictive analysis; and (b) to investigate its feasibility by comparing the performance of PAT to those of the Flexilevel (Lord, 1971, 1980), Bayesian modal (Assessment Systems Corporation, 1990) and expected a posteriori (EAP) (Bock & Aitken, 1981) strategies in a simulated environment.

MODEL

Predictive Statistical Analysis in Educational Testing

Much of statistical analysis is concerned with making inferences about the distributions of unknown parameters. In educational testing, the parameter θ usually represents the ability or trait of an examinee to be measured and an educational test is a tool that quantifies his/her ability level in some way to obtain a numerical score. This educational test could be a fixed-length paper-and-pencil conventional or an adaptive test.

The proposed adaptive test is based on the idea of using item difficulty p and past information (observed data) x about an examinee--in this case it will be the number of correct scores during the testing up to a certain point--to acquire his/her probability of answering future item(s) correctly.

The statistical predictive analysis is composed of two experiments: informative experiment e and future experiment f.

Each informative experiment e_i is an experiment that is performed in the past and its typical outcome is denoted by x_i, where

$$x_i = \begin{cases} 1 & \text{if response is correct,} \\ 0 & \text{otherwise,} \end{cases}$$

that is distributed as a Bernoulli variable with parameter θ,

$$f(x_i;\theta) = \theta^{x_i}(1-\theta)^{1-x_i}.$$

The informative experiment e involves responses to items that have already been administered. The future experiment also involves item(s) that will be administered to the examinee following the items already administered during the informative experiment e. Likewise, the outcome of the future experiment f_i, y_i, is a dichotomously scored item,

$$y_i = \begin{cases} 1 & \text{if response is correct,} \\ 0 & \text{otherwise.} \end{cases}$$

Then, the number of correct scores in future $y=\Sigma y_i$ is distributed binomially with parameter θ, $f(y_i,\theta)$, if items are independent and probability of $y_i=1$ is constant across the items.

The informative experiment e conveys information to the future experiment f about the performance of an examinee up to a particular point through the ability parameter θ that is assumed to be fixed (Aitchison & Dunsmore, 1975, p.19). This is the only link between these two experiments. The second assumption suggested that for a given examinee, his/her response to the previous items do not affect the response to the future item(s). This assumption is similar to the local independence assumption in item response theory (IRT). In simulation study, this can easily be met.

However, in real testing situations, an examinee's response to the previous item(s) may affect the response to the future item(s).

DEVELOPMENT OF THE MODEL

The development of the PAT model can best be described according to the components of adaptive testing. These components are: (a) initial (entry-level) item, (b) scoring method, (c) selection of the subsequent items to be administered, and (d) terminating criterion.

Initial (Entry-Level) Item

In general, the prior distribution contains some information about the parameter θ. An investigator intends to generate more accurate inferences about the parameter θ by using the prior information. Since generation of a posterior distribution is simplified if the prior and likelihood densities belong to the same conjugate family, the prior distribution of ability is assumed to be a beta with a location parameter $g>0$ and a scale parameter $h>0$, in predictive adaptive testing:

$$f(\theta)=\frac{\Gamma(g+h)}{\Gamma(g)\Gamma(h)}\,\theta^{g-1}(1-\theta)^{h-1}, \qquad 0<\theta<1 \qquad (1)$$

where ability parameter θ is in the range of 0 and 1.

The selection of the entry-level item is closely related to the prior distribution of an ability. Since at the beginning of the testing there is no informative data, the total number of correct answers x and total number of items already administered n are 0 and 0, respectively. Therefore, the probability of answering

the initial item correctly given item difficulty p and no observed data, $f(y=1;p,x=0)$, equals the mean value of the beta (prior) distribution, $g/(g+h)$, where g and h are location and scale parameters, respectively. Thus, the initial item selected is the one whose item difficulty level is closest to the mean of the prior distribution.

Scoring Method

The likelihood function $L(\theta)$ of item responses $(x_1,\ldots,x_n)$ is the multiplication of Bernoulli distributions, $f(x_i;\theta)=\theta^{x_i}(1-\theta)^{1-x_i}$. Thus,

$$L(\theta)=\prod_i f(x_i;\theta)=\prod_i \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{x}(1-\theta)^{1-x} \qquad (2)$$

where $0<\theta<1$, $x=\Sigma x_i$ and $x=0,1,\ldots,n$.

Then, the posterior distribution is a beta distribution with density

$$f(\theta;x)=(\text{constant}).\theta^{x+g-1}(1-\theta)^{n+h-x-1}, \qquad (3)$$

where constant$=\Gamma(n+g+h)/(\Gamma(x+g)\Gamma(n+h-x))$. The mean of this distribution is $(x+g)/(n+g+h)$ and variance is $(x+g)(n+h-x)/(n+g+h)^2(n+g+h+1)$. As mentioned before, the probability assessment about the unknown parameter θ is not the final objective of the predictive analysis. The main purpose is to assess a probability about the future outcome y given informative data x without the unknown parameter θ. Thus, the predictive density function can be expressed as

$$f(y;x)=\int_\Omega f(y;\theta)\,f(\theta;x)\,d\theta \qquad (4)$$

where $f(y;\theta)$, which describes the future experiment, is distributed

binomially with parameter θ and sample size m (number of items to be administered in the future),

$$f(y;\theta) = \binom{m}{y}\theta^{y}(1-\theta)^{m-y}, \tag{5}$$

where y=0,1,...,m. Then, the predictive distribution, beta-binomial, for y given f(θ) and x, can be written as

$$f(y;x) = \binom{m}{y}\frac{\Gamma(u+v)\Gamma(y+u)\Gamma(m+v-y)}{\Gamma(u)\Gamma(v)\Gamma(m+u+v)}, \tag{6}$$

where y=0,1,...,m, u=x+g, and v=n+h-x (Ferguson, 1967). The mean of this distribution is mu/(u+v) and the variance is $muv(m+u+v)/(u+v)^2(u+v+1)$.

Figure 1

The Basic Steps Leading to the Predictive Distribution[1]

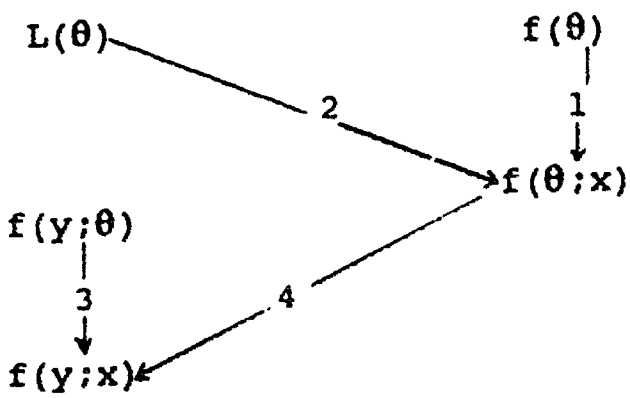


Figure 1 summarizes the basic steps leading to the predictive distribution. The arrows 1 and 2 converge to the f(θ;x) that is a

[1]The figure presented here is provided by Aitchison and Dunsmore, 1975.

result of Bayesian theorem. From that point, posterior distribution together with the distribution of future outcome y, arrows 3 and 4, are combined by using the definition of predictive distribution in (4).

Predictive distribution f(y;x) is the best approximation to the f(y;θ) (Aitchison, 1975) that describes the examinee's future performance. To find the ability estimate of an examinee, i.e., the probability of answering next item correct given item difficulty p and number of correct scores x, f(y=1;p,x), the proportionality of f(y=1;p,x) to f(p;y=1,x)f(y=1;x) is used (Hacking,1965). f(y=1;x) is the predictive probability and f(p;y=1,x) is the posterior probability of item difficulty given past (observed data) and future information of an examinee. The item difficulty p is calculated as the proportion of total group responding an item incorrectly. To obtain the posterior distribution of item difficulty p, a prior distribution for item difficulty p is defined as a beta distribution with certain scale l>0 and location parameter k>0. The resulting posterior distribution is again distributed as a beta with parameters k+x and l-x, where $x=\Sigma x_i$. Therefore, after terminating the test, f(y=1;p,x) which is the probability of answering next item correctly, y=1, given item difficulty p and the number of correct response to items already administered, x, will be regarded as an ability estimate of an examinee. Thus, the probability f(y=1;p,x) combines the information from item difficulty, observed data and examinee's ability level.

Selection of Subsequent Items to be Administered

To find the most appropriate item to administer to an examinee, the following criterion is considered:

$$\min_{p} \mid f(y=1;x)-f(y=1;p,x) \mid, \qquad (7)$$

where f(y=1;x) is the predictive probability of answering the next item correctly given an examinee's number of correct scores to the items already administered. The item difficulty parameter p is calculated as the proportion of total group responding an item incorrectly. The above criterion is constructed by considering the following relations: (a) for a given adequately large item pool, almost perfect positive correlation between f(y=1;x) and f(y=1;p,x) that is the probability of answering the next item correctly given item difficulty and number of correct scores; and (b) also high negative correlation between f(y=1;p,x) and item difficulty p ($0 \le p \le 1$). According to the above criterion, the most appropriate item to be administered is the one with item difficulty that is closest to his/her predictive probability. In adequately large item pool, it can be shown that the values of f(y=1;x) and f(y=1;p,x) are similar for an item selected according to the criteria (7) specified above. Therefore, they both can be used as an ability estimate of an examinee. Thus, the most appropriate item to be administered is the one whose item difficulty is closest to the examinee's ability level.

Termination Criteria

There are two widely used termination criteria in literature:

(a) testing continues until a prespecified number of items are administered, (b) testing continues until a prespecified value of an information function or standard error of estimate is reached. In predictive adaptive testing, a combination of these two widely used termination criteria are employed. That is, testing will continue until either a prespecified number of items are administered or a prespecified value of standard error of estimate is reached.

The standard error of estimate obtained from a posteriori distribution of ability (3) is considered as a termination criterion. The following beta distribution in (3) is derived as a posterior distribution of ability in the process of extraction of predictive distribution, beta-binomial,

$$f(\theta;x)=(\text{constant}).\theta^{x+g-1}(1-\theta)^{n+h-x-1}, \qquad (8)$$

where (constant)= $\Gamma(n+g+h)/(\Gamma(x+g)\Gamma(n+h-x))$. The mean of this distribution is $(x+g)/(n+g+h)$ and variance is $(x+g)(n+hx)/(n+g+h)^2(n+g+h+1)$. The parameters g and h stand for the location and scale parameters of a prior distribution of ability, x denotes the number of correct scores out of n items already administered. Testing will continue until the square root of the variance of the above beta distribution reaches the prespecified value. As a result, after terminating the testing, predictive adaptive testing provides a final predictive probability, $f(y=1;p,x)$ or $f(y=1;x)$, both can be used as an ability estimate.

METHOD AND PROCEDURE

The performance of predictive adaptive testing was compared to those of the flexilevel, the Bayesian modal and EAP strategies. In order to show the feasibility of predictive adaptive testing, data were generated by the Monte-Carlo simulation technique.

Generation of Population

In this Monte-Carlo simulation study, each examinee was identified by a numerical value reflecting their ability level, θ. Ability levels of a total of 999 examinees were randomly generated from a standard normal distribution in the interval of 3.0 to +3.0.

The seventy-one-item test was generated by assuming that the discrimination parameter a was distributed uniformly in the interval of 0.19 to 1.69 (Hambleton & Traub, 1971). The difficulty level b was distributed normally with mean 0 and variance 1 in the interval of -3.0 to +3.0. Finally, the guessing parameter c was assumed to be uniformly distributed in the interval of 0 to 0.20. In order to simulate the responses of 999 examinees to the seventy-one-item test, the subprograms of IMSL (1984, version 9.2) library on PITT VAX/VMS system were used.

The dichotomous (0, 1) score of any examinee on any item was a probabilistic function of their ability level θ, the item difficulty b, and the parameters a and c. The probability $P_i(\theta_j)$ of a correct response under the 3-parameter logistic model item characteristic curve was calculated according to the following formula

$$P_i(\theta_j)=c_i + \frac{1-c_i}{1+e^{-Dai(\theta j-bi)}},$$

where i and j denoted item and examinee, respectively. In order to simulate dichotomous item response, each probability value $P_i(\theta_j)$ was compared with a random number r_{ij} which was generated from a uniform distribution in the interval of 0 to 1. The response was assumed correct and a score of 1 was assigned, if the probability value was equal or greater than the random number r_{ij}; otherwise a score of 0 was assigned.

The Program ASCAL (Assessment Systems Corporation, 1990) was used to estimate ability and item parameters based on the generated item responses from 999 examinees. Chi-squared goodness-of-fit[2] tests for the true and estimated values of ability and for the true and estimated values of item parameters were carried out in order to provide an evidence for how well the data generation process worked.

The calculated chi-squared values are presented in Table 1 for ability parameter θ and item parameters a, b, c. It was concluded that the estimated values of the ability parameter θ and item parameters were not significantly different from their generated values.

---

[2] $\chi^2=\Sigma(O_i-E_i)/E_i$ is distributed as a chi-squared with k-1 degrees of freedom, where k is the number of categories and O_i and E_i are observed and expected values, respectively.

Table 1

Chi-Squared Goodness-of-Fit Test Results

Test	df	Chi-squared test value
Ability, θ	89	19.787
Discrimination, a	70	6.456
Difficulty, b	70	42.249
Guessing, c	70	8.317

Note: The goodness-of-fit tests are non-significant at the α=0.05 level.

Table 2

Conventional Item Statistics for Raw Scores

Number of items	71	Minimum	5
Number of examinee	999	Maximum	71
Mean	37.856	Median	38
Variance	167.929	Alpha	0.926
Std.Dev.	12.959	SEM	3.532
Skewness	0.036	Mean Bis	0.545
Kurtosis	-0.675		
Mean p	0.533		
Mean item-total	0.405		

The program ITEMAN (Assessment Systems Corporation, 1990) was employed to calculate the conventional item statistics such as proportion correct, biserial correlation, and point-biserial correlation. Furthermore, the alpha-reliability coefficient was calculated, 0.926. The results in Table 2 suggested that the 71-item test adequately represented examinees in the medium ability group.

Finally, the test for unidimensionality of the abilit. pace, which is assumed by IRT, was performed. According to the test proposed by Reckase (1979), inter-item correlation coefficients were calculated in order to find the eigenvalues. The test results showed that the first eigenvalue accounted for 37% of the total variance which was greater than the recommended 20% value. Therefore, the assumption of unidimensionality of the ability space appeared to be reasonable.

Sample

Examinees were grouped into three different ability levels based on their randomly generated true abilities. In order to assign each examinee to one of the three groups of low, medium, or high, examinees were ranked according to their generated true ability level. Then, the examinees were clustered into nine mutually exclusive groups in such a way that each section contained an equal number of examinees, i.e., 111. From each section, ten examinees were randomly selected. Thirty examinees from the top three ability sections were grouped into the high ability group. Similarly, the same number of examinees from the bottom three ability sections were classified as a low ability group. The remaining examinees formed the middle ability group.

Procedure

The Bayesian modal, EAP and PAT strategies required the specification of prior distributi-n about the examinee's ability level. The medium ability level assumption was the only one assumed for all strategies requiring the specification of prior

distribution. For the Bayesian modal and EAP strategies, the mean and variance of normal distribution were specified as 0 and 1, respectively. Since IRT-based adaptive testing strategies and PAT were based on different distributional assumptions, the prior distributions were not perfectly comparable. However, in this case, the prior was a beta distribution with the location and scale parameters g=2 and h=2, respectively. Since this beta distribution is symmetrical, its mean, mode and median values were all equal to 0.5.

Two termination criteria were used in the present study: In determining the ability estimate of an examinee and the final standard error of estimate, thirty-six items were administered to every examinee. This maximum number of items administered was required by the 71-item flexilevel test. Therefore, the comparison of the ability estimates and the final error variance of the ability estimates from different strategies were based on the same number of items. In determining the number of items required to reach the prespecified termination criterion, for the Bayesian modal, EAP and PAT strategies, the standard error of estimate that was calculated from the expected test information was set to 0.30.

To simulate the adaptive testing for the predictive and flexilevel strategies, Fortran IV computer programs were prepared. Items were selected according to the adaptive testing strategies and the corresponding response (correct or incorrect) was entered by the program itself. For the Bayesian modal and EAP strategies, MicroCAT was used to administer adaptive testing. When the program

selects the appropriate item to administer, that particular item was seen on the screen. The investigator then entered the response either correct or incorrect based on the examinee's simulated response.

In order to assess the accuracy of the performance of PAT, the correlation coefficients were calculated between ability estimate (final predictive probability) obtained from PAT and the generated ability score. Furthermore, the correlations between the generated ability score and the other ability estimates obtained from flexilevel, the Bayesian modal and EAP were computed as well. The test of equality for the above correlation coefficients were carried out in order to examinee the similarity between estimated and true ability scores in terms of order of scores.

Data Collection

The following data were collected for each strategy:

1. item identifier;

2. subject's response; (0,1),

3. flexilevel test score, ability estimate scores obtained from the Bayesian modal, EAP and PAT--final predictive probability was used as an ability estimate for PAT-- strategies;

4. the final error variance of ability estimate, i.e., standard error of posterior distribution for the Bayesian modal, EAP and PAT strategies; and

5. the number of items required to reach a prespecified terminating criterion.

Data Analysis

The independent variables that were considered are as follows:

1. adaptive testing strategy, and

2. ability levels.

The ability level (high, medium, low) was regarded as a between-subjects variable. On the other hand, the adaptive testing strategy (flexilevel, PAT, the Bayesian modal, the EAP) was considered to be a within-subjects variable. In this study, a two-way mixed factorial design with repeated measures on one of the factors was used.

The dependent variables that were considered are:

1. The number of items required for each strategy to reach a prespecified terminating criterion. This dependent variable was the indicator of efficiency in adaptive testing;

2. The absolute value of the difference between generated true ability and estimated ability scores obtained from flexilevel, the Bayesian modal, EAP and PAT strategies. Since the ability estimates obtained from IRT-based adaptive testing strategies, flexilevel and PAT could not be compared on the same metric--due to the difference in distributional assumptions, the difference was calculated between the standardized scores. Thus, the comparisons, in some sense, were made possible. Furthermore, the absolute value of the difference was taken in order to show the accuracy and the similarity of the obtained scores; and

3. The absolute value of the difference between error variance of the final estimate obtained from adaptive testing

strategy and error variance when the complete test was considered. Due to the reasons mentioned in the previous paragraph, all the error variances of ability estimates were transformed to standardized scores before taking the differences. This dependent variable was an indicator of similarity between error variances obtained from adaptive testing strategy and error variance of the complete test (true error variance).

The following hypotheses were tested:

H_{01}: There is no significant difference between means of examinees for different adaptive testing strategies for each of the dependent variables 1-3,

H_{02}: There is no significant difference between means of examinees for three different ability levels for each of the dependent variables 1-3,

H_{03}: There is no significant interaction effect of the adaptive testing strategy and ability level for each of the dependent variables 1-3.

Since the flexilevel test administers the same number of items to each examinee, it was excluded from hypotheses testing when the first dependent variable was considered. For the second dependent variable, all four adaptive testing strategies were included. However, for the third dependent variable, since the error variance could not be calculated for the flexilevel test, it was excluded from hypotheses testing.

RESULTS AND DISCUSSIONS

Number of Items Required to Reach the Prespecified Termination Criterion

The strategies that were considered here were the Bayesian modal, EAP, and PAT. The preliminary studies showed that the raw scores of the number of items required did not meet the assumptions to carry out F-tests in mixed factorial design (Kirk, 1982, p.74), i.e., the observations were not normally distributed and variances were not equal. Therefore, an angular transformation (Kirk, 1982, p. 83) of the observed scores was performed. The cell and marginal means corresponding to the adaptive testing strategies and ability groups are summarized in Table 3. The results of two-way mixed factorial design in Table 4 revealed that the interaction effect between adaptive testing strategy and ability group was statistically significant at α=0.01. In order to show the nature of the interaction effect, Figure 2 was plotted by considering cell means provided in Table 3. The plot indicated that, at the low ability level, the PAT strategy required more items to reach the prespecified termination criterion than the Bayesian modal and EAP. However, the pairwise mean differences calculated according to the Scheffe post-hoc method, at the low ability level, were not statistically significant (see Table 5).

Table 3

Cell and Marginal Means of the Number of Items Required to Reach the Prespecified Termination Criterion

	Low	Medium	High	Marginal
Modal	17.43	10.87	12.80	13.70
EAP	19.97	12.40	14.90	15.76
PAT	17.47	19.63	19.73	18.94
Marginal	18.29	14.30	15.81	

Table 4

Results of the Mixed Factorial Design on the Number of Items Required to Reach the Prespecified Termination Criterion in Terms of Angular Transformation

Sources	SS	df	MS	F-ratio	p-value
Mean	0.05	1	0.05	0.11	0.744
Ability(A)	3.08	2	1.54	3.42	0.037
Error(A)	39.17	87	0.45		
Strategy(S)	31.26	2	15.63	42.38	0.000
S X A	11.17	4	2.79	7.57	0.000
Error(S)	64.18	174	0.37		

If the starting point matched with the actual ability level, the medium ability level, the Bayesian modal and EAP required less number of items than the PAT strategy. As can be noticed in Table 5, the pairwise mean differences between PAT and EAP and also PAT and the Bayesian modal were statistically significant at α=0.01 level.

Table 5

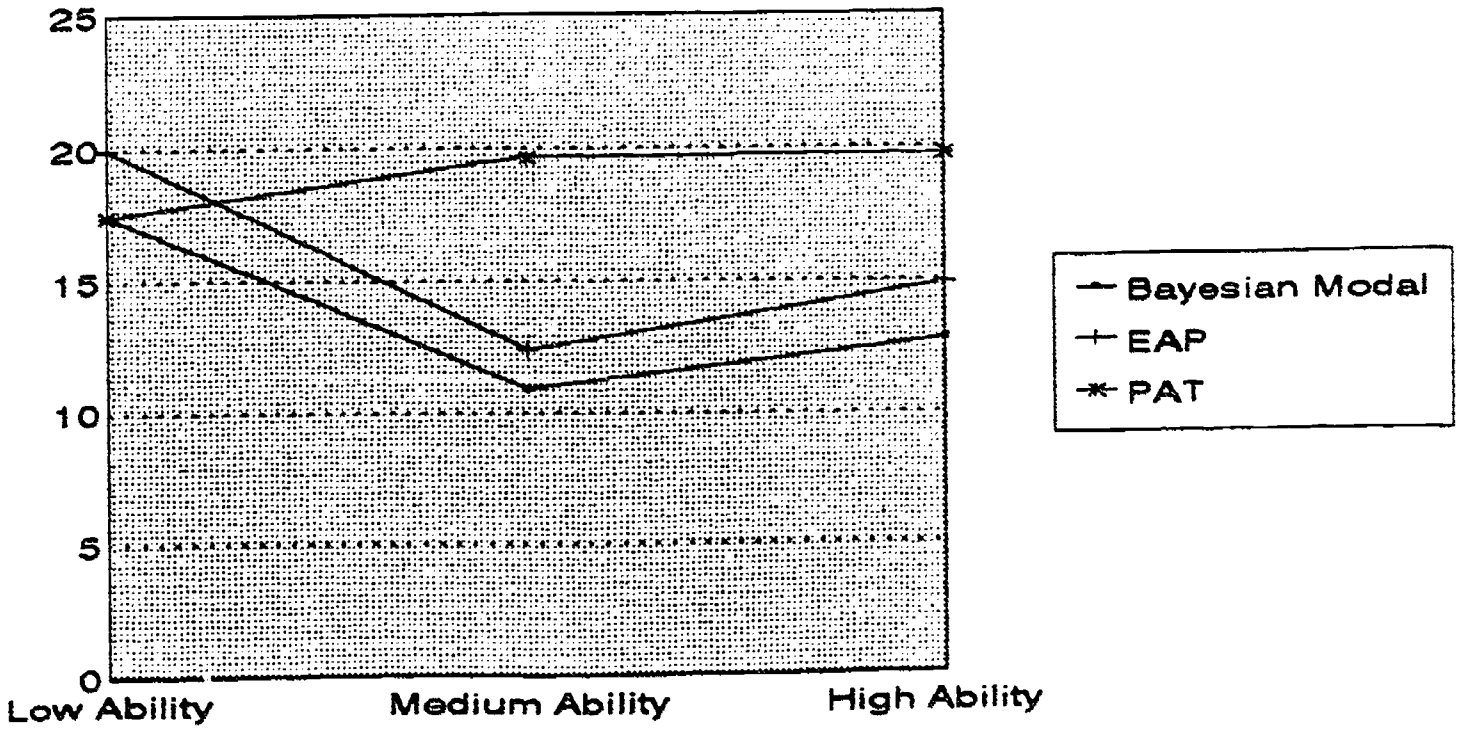
F-Values of Scheffe Test for Adaptive Testing Strategy and Ability Group on the Number of Items Required to Reach the Prespecified Termination Criterion in Terms of Angular Transformation

Ability Group		EAP	PAT
Low	Modal	2.043	2.177
	EAP		0.134
Medium	Modal	1.521	7.505*
	EAP		5.983*
High	Modal	5.461*	6.194*
	EAP		0.732

* $p<0.01$

Figure 2

Interaction Effect Between Adaptive Testing Strategy and Ability Group on the Number of Items Required to Reach the Prespecified Termination Criterion



At the high ability level, the number of items required increased for the Bayesian modal and EAP strategies. However, for PAT strategy, the number of items required was higher than those of Bayesian modal and EAP. The post-hoc comparison for pairwise mean difference between PAT and EAP strategies, at the high ability level, was not significant at $\alpha=0.01$. The Bayesian modal strategy required significantly fewer number of items than the PAT and EAP.

In summary, the results revealed that at the low ability level the number of items required by the three adaptive testing strategies were not significantly different. The Bayesian modal and EAP strategies required significantly fewer number of items than the PAT when the starting point matched with the actual ability level. At the high ability level, the Bayesian modal strategy required significantly less number of items than the PAT and EAP.

Absolute Value of the Difference Between Standardized Ability Estimate and Generated Ability

The second dependent variable was the absolute value of the difference between standardized ability estimate obtained from the adaptive testing strategies and generated ability scores. The data were analyzed by two-way mixed factorial design. The first factor was the adaptive testing strategy (the Bayesian modal, EAP, and PAT). The second factor was the ability group, i.e., low, medium, and high.

For the same reasons mentioned in preceding section, a transformation of data was necessary to meet the assumptions of

normality and homogeneity of variances. The data were transformed by using the square-root method (Kirk, 1982, p. 82). The cell and marginal means are presented in Table 6. The results of two-way mixed factorial design were summarized in Table 7.

Table 6

Cell and Marginal Means of the Obtained Ability Estimate and Generated True Ability in Terms of Raw Scores

	Low	Medium	High	Marginal
True	-1.22	-0.05	0.94	-0.11
Flex	0.38	0.53	0.68	0.53
Modal	-1.15	-0.06	0.96	-0.08
EAP	-1.17	-0.06	0.93	-0.10
PAT	0.32	0.41	0.55	0.42

The test results showed that the interaction effect between adaptive testing strategy and ability group in terms of square-root of the absolute value of the difference between standardized ability estimate and standardized generated ability score was not significant at the $\alpha=0.01$ level. Therefore, the next step in data analysis was to test the main effects due to the adaptive testing strategy and ability group.

Table 7

Results of the Mixed Factorial Design on the Absolute Value of Difference Between Standardized Ability Estimate and Generated Ability in Terms of Square-Root Transformation

Source	SS	df	MS	F-ratio	p-value
Mean	70.11	1	70.01	1496.52	0.000
Ability(A)	0.15	2	0.08	1.65	0.198
Error(A)	4.07	87	0.05		
Strategy(S)	0.62	3	0.21	6.06	0.001
S X A	0.29	6	0.05	1.42	0.208
Error(S)	8.93	261	0.03		

Table 7 revealed that the main effect of adaptive testing strategy was significant at α=0.01. The post-hoc comparisons of pairwise mean differences were calculated by using the Scheffe method and are summarized in Table 8. According to the results presented in Table 8, the pairwise mean differences between adaptive testing strategies were all non-significant at the α=0.01 level. The pairwise comparisons computed by the Scheffe method were not able to detect any significant mean differences between adaptive testing strategies. On the other hand, the main effect of ability group was found to be non-significant at the α=0.01 level.

Table 8

F-Values of Scheffe Test for the Main Effect of Adaptive Testing Strategy on the Absolute Value of Difference Between Standardized Ability Estimate and Generated Ability Score in Terms of Square-Root Transformation

	Modal	EAP	PAT
Flex	0.548	0.629	2.164
Modal		0.082	2.711
EAP			2.793

In summary, the main effects of adaptive testing strategy and ability group were additive. Although the main effect of adaptive testing strategy was significant, the post-hoc comparisons did not reveal any significant pairwise differences between the means of adaptive testing strategies.

Absolute Value of Difference Between Standardized Error Variances of Ability Estimate and Complete Test

Since the flexilevel test did not yield any error variance of ability estimate, it was not included into the statistical analysis. The strategies which were considered here were the Bayesian modal, EAP, and PAT.

Due to the procedural differences among adaptive testing strategies, IRT-based adaptive testing strategies and PAT did not produce comparable error variances. All the error variances of abilities were transformed to z scores before taking the absolute value of differences. These absolute value of differences that were taken between the error variance obtained from the complete

test and error variance obtained from adaptive tests showed the accuracy.

For the same reasons mentioned in previous sections, a transformation of data was necessary to meet the assumptions for normality and homogeneity of variances. The scores were transformed by using the logarithmic transformation.

The cell and marginal means are presented in Table 9. The results of mixed factorial design were summarized in Table 10. The test results of mixed factorial design showed that the interaction effect between adaptive testing strategy and ability group in terms of logarithmic transformation of the absolute value of the difference between standardized error variances was not significant at α=0.01 level. The tests for the main effects due to the adaptive testing strategy and ability group revealed that the main effects of adaptive testing strategy and ability group were not significant at α=0.01. According to the above results, the means of the error variances produced by the Bayesian modal, EAP and PAT, were statistically similar.

Table 9

Cell and Marginal Means of the Error Variances Obtained from Complete and Adaptive Tests in Terms of Raw Scores

	Low	Medium	High	Marginal
True	.03	.04	.03	.03
Modal	.07	.05	.05	.06
EAP	.08	.05	.06	.06
PAT	.05	.06	.06	.06

Table 10

Results of the Mixed Factorial Design on the Absolute Value of Difference Between Standardized Error Variances Obtained from Adaptive Testing Strategies and Complete Test in Terms of Logarithmic Transformation

Source	SS	df	MS	F-ratio	p-value
Mean	14.58	1	14.58	11.86	0.001
Ability(A)	2.59	2	1.29	1.05	0.354
Error(A)	106.94	87	1.23		
Strategy(S)	2.25	2	1.12	1.23	0.295
S X A	8.51	4	2.13	2.33	0.058
Error(S)	158.98	174	0.91		

Corrslation Coefficients Between Ability Estimate and Generated Ability

In the final section, the correlation coefficients between ability estimates obtained from adaptive testing strategies and generated ability scores were computed. The results were summarized in Table 11.

Table 11

Correlation Coefficients Between True Ability and Ability Estimates and Also Between Ability Estimates

	Flex	Modal	EAP	PAT
True	0.966	0.971	0.976	0.933
Flex	1.000	0.934	0.925	0.916
Modal		1.000	0.980	0.889
EAP			1.000	0.876
PAT				1.000

Note: "True" stands for generated true ability score.

The results showed that all the correlation coefficients presented in Table 11 were statistically significant at the α=0.001 level. The correlation coefficients between generated true ability and ability estimates obtained from adaptive testing strategies were all above 0.93. This revealed that all the ability estimates obtained from adaptive testing strategies were highly correlated with the generated true ability scores. The EAP strategy had the highest correlation coefficient (0.976).

The test for equality of the above correlation coefficients (Glass & Stanley, 1970, p.313) such as corr(True,Flex)=corr(True,Modal) are summarized in Table 12.

Table 12

Test for Equality of Correlation Coefficients Between True Score and Ability Estimates

	(True,Modal)	(True,EAP)	(True,PAT)
(True,Flex)	-0.7348	-1.5481	2.7247*
(True,Modal)		-0.8051	2.5797*
(True,EAP)			3.6108*

* p<0.01

The results showed that, in terms of correlations with true score, PAT is significantly different from the flexilevel and EAP at 0.01 level. However, all the other correlations coefficients between adaptive test scores and true scores were not significantly different.

SUMMARY AND CONCLUSION

A model using predictive statistical analysis was developed. The feasibility of the model was compared with other adaptive testing strategies in a simulation study. The results of the data analysis can be summarized as follows:

1. In terms of number of items administered to reach the prespecified termination criterion, all the three adaptive testing strategies performed similar at the low ability level. At the medium ability level, the Bayesian modal and EAP strategies were the most efficient ones. At the high ability level, the Bayesian modal strategy required significantly less number of items than the PAT and EAP.

2. In terms of the absolute value of the difference between standardized ability estimate and generated ability score, all the strategies yielded statistically comparable estimates.

3. In terms of the absolute value of the difference between standardized error variances, all the adaptive testing strategies, the Bayesian modal, EAP, and PAT, produced equally comparable and similar results.

4. As a final analysis, the correlation coefficients were calculated between ability estimates obtained from adaptive testing strategies and generated true ability score. The results showed that all the correlation coefficients were comparable and highly significant. The tests for the equality of the correlation coefficients, mentioned above, revealed that the PAT and Bayesian modal strategies produced significantly similar ability estimates

to the true ability scores in terms of the order of scores.

The performance of PAT was not quite as efficient as the Bayesian modal and EAP strategies at the middle ability level in terms of number of items required. However, PAT produced similar results in terms of error variance. When ability estimates were considered, all the adaptive testing strategies produced equally comparable results.

Based on the results of this study, it can be concluded that PAT has a potential to be utilized. Since IRT-based adaptive testing strategies require a larger sample size to calibrate item parameters and some assumptions to be met, the implementation of PAT into small classroom testing is more practical.

REFERENCES

Aitchison, J. (1975). Goodness of prediction fit. Biometrika, 62, 547-554.

Aitchison, J., & Dunsmore, I.R. (1975). Statistical prediction analysis. Cambridge: Cambridge University Press.

Aitchison, J., & Sculthorpe, D. (1965). Some problems of statistical prediction. Biometrika, 52, 469-483.

Assessment Systems Corporation (1990). User's manual for the MicroCAT testing system, (3rd. ed.). Assessment System Corporation. St. Paul, Minnesota: Author.

Bock, R.D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika, 46, 443-459.

Ferguson, T.S. (1967). Mathematical statistics: a decision theoretic approach. New York: Academic Press.

Glass, G.V. & Stanley, J.C. (1970). Statistical methods in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall.

Hambleton, R.K., & Traub, R.E. (1971). Information curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 24, 273-281.

Hacking, I. (1965). Logic of statistical inference. Cambridge: Cambridge University Press.

IMSL (1984). Library user's manual, Fortran subroutines for mathematics and statistics, (Edition 9.2). IMSL, Inc., Houston, Texas: Author.

Kirk, R.E. (1982). Experimental design, (2nd ed.). Belmont, CA: Brooks/Cole.

Lord, F.M. (1971). The self-scoring flexilevel test. Journal of Educational Measurement, 8, 3, 147-151.

Lord, F.M. (1980). Applications of item response theory to practical problems. Hillsdale, N.J.: Erlbaum Publishers.

Rechase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. Journal of Educational Statistics, 4, 207-230.