DOCUMENT RESUME

ED 269 437 TM 860 269

AUTHOR DeAyala, R. J.; Koch, William R.

TITLE A Computerized Implementation of a Flexilevel Test and Its Comparison with a Bayesian Computerized Adaptive Test.

PUB DATE [86]

NOTE 9p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 17-20, 1986).

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Ability; *Adaptive Testing; *Bayesian Statistics; *Comparative Analysis; *Computer Assisted Testing; Correlation; Estimation (Mathematics); Item Analysis; Item Banks; Latent Trait Theory; Scores; Scoring; Simulation; Testing Problems

IDENTIFIERS *Ability Estimates; Item Parameters; *Three Parameter Model; T Test

ABSTRACT

A computerized flexilevel test was implemented and its ability estimates were compared with those of a Bayesian estimation based computerized adaptive test (CAT) as well as with known true ability estimates. Results showed that when the flexilevel test was terminated according to Lord's criterion, its ability estimates were highly and significantly correlated with those of the Bayesian based CAT and with the known true thetas. Matched t-tests between the flexilevel and the Bayesian CAT and with the known true thetas were not significantly different from one another. The primary implication of this study is that it is feasible to implement a computerized adaptive test without having to create a calibrated item pool containing 200 or more items, as is the case with Item Response Theory (IRT) based CAT systems. Neither the software nor the hardware needed for flexilevel CAT methods present the rather serious obstacles that they often do for IRT-based CAT. This fact and the result that flexilevel CAT ability estimates compare favorably with the IRT-based CAT ability estimates indicate that, for the first time, CAT procedures can be made practical and feasible for ordinary classroom testing. (Author/PN)

ABSTRACT

A Computerized Implementation of a Flexilevel Test and its Comparison with a Bayesian Computerized Adaptive Test

R. J. DeAyala and William R. Koch
The University of Texas at Austin

A computerized flexilevel test was implemented and its ability estimates were compared with those of a Bayesian estimation based computerized adaptive test (CAT) as well as with known true ability estimates. Results showed that when the Flexilevel test was terminated according to Lord's criterion, its ability estimates were highly and significantly correlated with those of the Bayesian based CAT and with the known true thetas. Furthermore, matched t-tests between the Flexilevel and the Bayesian CAT and with the known true thetas were not significantly different from one another. Implications for classroom testing are presented.

OBJECTIVES

The objectives of this research were to (1) implement a computerized adaptive testing program based on the Flexilevel testing procedure, and (2) to determine how well this strategy compares with a Bayesian computerized adaptive testing procedure.

Adaptive testing has primarily been associated with two ability estimation techniques: Bayesian and maximum likelihood estimation. Both techniques utilize computationally intense procedures to determine which items should be administered to the examinee for the estimation of ability. Furthermore, these techniques require at least 100 and preferably more "calibrated" items for the formation of an item pool. The initial calibration of items requires a large subject pool, a large item pool and a sophisticated calibration computer program (e.g., MULTILOG, LOGIST). Therefore, the utilization of these techniques for adaptive testing requires access to mainframe computers, a calibration program, and personnel who are sophisticated in computer langauges such as FORTRAN, Pascal, MODULA-2, etc. Further, these personnel need a certain amount of knowledge of Item Response Theory (IRT).

In contrast, the adaptive testing technique known as Flexilevel Testing (Lord, 1972) was originally proposed as a paper and pencil method which, like CAT systems, attempted to match item difficulty with ability level. The Flexilevel technique begins by ordering the test items from easy to hard. The item of median difficulty is the first item administered to all examinees. The examinee then proceeds through the test taking an easier item each time the examinee gets an item wrong and a harder item each time he/she gets an item right. However, Lord discussed some potential problems with this method. For example, Lord was concerned with the extent to which different types of examinees might be confused by the flexilevel testing procedure, with the flexilevel test's loss of efficiency due to increased testing time per item (primarily a result of the examinees having to score their own items), with the scoring of an examinee who does not have time to finish the test, and with the scoring of an examinee who does not follow directions. Fortunately all of these problems can be corrected by using a computer to select, present and score the test items. More importantly, the programming necessary to implement a computerized Flexilevel test is much less sophisticated than, for instance, a maximum likelihood estimation based CAT system.

METHOD

Programs: A computer program was written which simulated the implementation of (1) a computerized adaptive testing procedure based on Bayesian estimation and which selected items to be administered using Jensema's (1974) alpha technique and (2) a variation of this program which used the flexilevel method for item selection, but Bayesian estimation of ability. The former method will be referred to as the Bayesian CAT, whereas the latter will be known as the Flexilevel CAT.

Lord's presentation of Flexilevel testing uses the number correct score as an ability estimate. However, to compare the Flexilevel CAT's ability estimates with a traditional Bayesian CAT procedure's theta estimates requires a transformation of the scale of one of the ability estimates. Therefore, to compare the ability estimates from the Flexilevel CAT with the Bayesian CAT, Bayesian estimation of ability was used in the Flexilevel CAT. In addition to Bayesian estimation of ability, the Flexilevel CAT also calculated the number correct score using the scoring method presented in Lord (1980).

Data: Given a sample of 100 simulated examinees whose "true" ability levels were selected from a N(0,1) distribution, item parameters for 129 items (i.e., IRT discrimination, difficulty and guessing parameters) were used to generate the binary response strings corresponding to these "true" thetas. The response string was generated to the three parameter logistic model by comparing the probability of a correct response for each item with a random number generated from a rectangular (uniform) distribution between 0.0 and 1.0. For a given item and a given true theta, the probability of a correct response was calculated from the three parameter logistic model. When the probability of a correct response was greater than the random number, the examinee's response was correct; otherwise it was considered incorrect.

The responses for all the "examinees" were then used to calculate the proportion correct value for each item (a.k.a., the item's traditional difficulty) and the item's point biserial correlation (a.k.a., the item's discrimination). These traditional difficulty values were used to order the items from easy to hard.

According to Lord (1980) the items on a Flexilevel test should be ordered in terms of IRT difficulty. However, he states that "any rough approximation to this [ordering by difficulty] is adequate" (Lord, 1980, p. 115). The rationale for using traditional difficulty to order the items rather than IRT difficulty values will be presented below. A Pearson correlation coefficient between the traditional difficulty values and the IRT difficulty values was calculated.

Analyses: One Bayesian CAT and four Flexilevel CAT runs were performed. The Bayesian CAT was terminated when either of two criteria were met: (1) the standard error of estimate was less than or equal to 0.25; or (2) a maximum of 20 items was administered. The Flexilevel CAT runs were terminated when a predetermined number of items was administerd. The four termination criteria were 20, 30, 40, and 65 items. A paper and pencil Flexilevel Test is completed when 0.5 x (total number of items + 1) items have been administered (i.e., 65 items for this item pool). Therefore, the termination criterion of 65 items for the Flexilevel CAT was comparable to a paper and pencil Flexilevel test.

The ability estimates for each of the four Flexilevel CAT runs were correlated with the Bayesian CAT estimates. Further, these five estimates and the known true theta were also correlated. In addition, matched t-tests were performed between each Flexilevel CAT run's ability estimates and the "true thetas" as well as with the Bayesian CAT's ability estimates in order to determine whether the theta estimates from the Flexilevel CAT were significantly different from the true thetas.

RESULTS AND CONCLUSIONS

For the following discussion an alpha level of 0.05 was used for significance tests and all reported correlations are Pearson correlation coefficients. The correlation between the IRT difficulty parameter estimates and the traditional difficulty values (p levels) was a high negative correlation (r= -0.9033). The mean of the traditional difficulty values was 0.541 (S.D.= 0.20, minimum value= 0.14, maximum value=0.96) and the mean of the IRT difficulty values was 0.046 (S.D.=1.1967, minimum=-4.251, maximum=4.559). On the basis of this correlation it was concluded that the ordering of the items according to traditional difficulty values provided a sufficiently "rough approximation" to the ordering according to the IRT difficulty parameter estimates.

Table 1 presents the Pearson correlations computed between the Bayesian IRT CAT theta estimates, the Flexilevel CAT estimates (for each termination criterion), and the true thetas. As can be seen from this table, all of the correlations between the Flexilevel CAT, regardless of number of items administered, and the true thetas are above 0.91 and are all significantly different from zero. Similarly, the association between the estimated thetas from the Bayesian CAT and those of the Flexilevel CATs are greater than 0.91 and are also statistically significant. Scatterplots of all 8 pairs show no "outliers" and, as would be expected, the points fall consistently along a line with a slope of 1.0

Table 1 : Correlation Between Flexilevel Thetas with IRT-CAT and True Thetas

Flexilevel	Bayesian CAT	True Thetas
65 Items	0.9256	0.9301
40 Items	0.9255	0.9238
30 Items	0.9203	0.9212
20 Items	0.9137	0.9140

There is more of a linear relationship between the Bayesian CAT and the true thetas (r=0.9514, see Table 2) than between the Flexilevel 65 item CAT and the true thetas (r=0.9301). The larger Bayesian CAT correlation with the true thetas is achieved by administering less than one-third as many items (average number of items administered by Bayesian CAT is 16.345).

Table 2 : Miscellaneous Statistics

Correlation between IRT difficulty & Traditional difficulty	=	-0.90235
Correlation between IRT CAT & True Theta	=	0.95140
Mean Number of Items administered in IRT CAT	=	16.34500

Based on the results of the present study it appears that selecting items according to the Flexilevel technique compares quite favorably with the Bayesian CAT procedure. All the correlations between the Bayesian CAT and the Flexilevel CAT, regardless of number of items administered, are high and significant. The Flexilevel CAT also provides comparably (high) correlations with the true thetas at the expense of administering more items.

The Pearson correlation coefficients computed between the number correct score for each Flexilevel CAT termination level and the true thetas, as well as with the IRT ability estimate generated from the corresponding Flexilevel CAT termination level, are presented in Table 3. As can be seen from this table the correlations between the true thetas and the number correct scores fall in the range of 0.92 to 0.93. The correlations between the Flexilevel CAT's estimated thetas and the number correct are all above 0.98. In short, using the number correct scoring method of Lord (1980) provides very good agreement with the known ability estimates as well as with the estimated abilities.

Table 3 : Correlation Between Flexilevel Number Correct with Flexilevel theta estimates and True Thetas

Flexilevel	Bayesian CAT	True Thetas
65 Items	0.9891	0.9351
40 Items	0.9878	0.9288
30 Items	0.9834	0.9214
20 Items	0.9872	0.9157

Table 4 presents the results of the matched t-tests between the estimated thetas from each of the Flexilevel CAT conditions and those from the Bayesian CAT as well as with the true thetas. As can be seen from this table, there is no significant difference between the true thetas and the Flexilevel 65 item CAT's theta estimates nor the Flexilevel 40 item CAT's estimates, t=-0.65 (p=0.515) and t=-1.62 (p=0.109), respectively. In contrast, the Flexilevel 30 item CAT produces theta estimates which are different from the known thetas; the same result holds for the Flexilevel 20 item CAT.

Table 4 : T-Tests (df=99; * denotes significant, alpha=0.05)

Comparison	Mean	SD	t	P
a. True Theta	-0.1111	1.101		
Flexilevel 65 items	-0.0762	1.373	-0.65	0.515
b. True Theta	-0.1111	1.101		
Flexilevel 40 items	-0.0295	1.294	-1.62	0.109
c. True Theta	-0.1111	1.101		
Flexilevel 30 items	-0.0000	1.240	-2.29*	0.024
d. True Theta	-0.1111	1.101		
Flexilevel 20 items	0.0129	1.139	-2.66*	0.009
e. IRT CAT	-0.0360	1.040		
Flexilevel 65 items	-0.0762	1.373	0.71	0.482
f. IRT CAT	-0.0360	1.040		
Flexilevel 40 items	-0.0295	1.294	-0.12	0.901
g. IRT CAT	-0.0360	1.040		
Flexilevel 30 items	-0.0000	1.240	-0.73	0.470
h. IRT CAT	-0.0360	1.040		
Flexilevel 20 items	0.0129	1.139	-1.06	0.294

As can also be seen from Table 4, the t-tests between the estimated thetas for the Bayesian CAT and the Flexilevel 65, 40, 30 and 20 item CATs are nonsignificant. The results show no difference between the Flexilevel CAT and the Bayesian CAT with respect to estimating the examinee's ability.

To summarize, the Flexilevel CAT's estimated thetas are highly and significantly correlated with the known thetas as well as with the Bayesian theta estimates, regardless of number of items administerd. Using number correct as an ability estimate, one finds a high and significant correlation with the known true levels of theta.

The estimated thetas from the 65 item Flexilevel CAT (i.e., the termination criterion espoused by Lord (1980)) are significantly correlated with the known true thetas. Further, the matched t-tests show that these estimated thetas are not significantly different from the known true thetas nor from the Bayesian CAT's estimated thetas. The same is true for the 40 item Flexilevel CAT. Flexilevel 30 and 20 item CATs produce theta estimates which are significantly different from the known true thetas. However, the Flexilevel 30 and 20 item CAT's estimated thetas are not significantly different from the Bayesian CAT's theta estimates.

EDUCATIONAL IMPLICATIONS

The purpose of this study was to examine the feasibility of implementing a computerized flexilevel test. This implementation alleviates four of six concerns about Flexilevel testing expressed by Lord's (1980, p. 126). The other two concerns are: (1) is the examinee's attitude and performance improved when the test "tailors" the difficulty of the items administered to match the examinee's ability, and (2) what other serious inconveniences and complications are there in Flexilevel testing. Because this was a simulation study these issues were not addressed. The first of these two concerns applies to all computerized adaptive testing procedures. Lord's second concern is a topic for further empirical investigation.

In the authors' opinion the primary implication of this study is that it is feasible to implement a computerized adaptive test without having to create a calibrated item pool containing 200 or more items, as is the case with IRT based CAT systems. There is no need for a large subject pool, e.g., 1000 examinees, to have taken test items previously to obtain estimates of the item parameters. Given the nonsignificant t-tests between the theta estimates from the Bayesian CAT and the Flexilevel 65 item CAT as well as between the Flexilevel 65 item CAT and the known true thetas, it appears that traditional difficulty values are sufficient for ordering the items from easy to hard. This would permit the use of a previously administered classroom test to form the "item pool" for a Flexilevel test. However, because traditional difficulty values are sample dependent it would still be desirable to administer the items to a large number of students to obtain "stable" difficulty values.

It is apparent from the above mentioned results of the Flexilevel 65 item CAT with the Bayesian CAT, as well as with the true thetas, that the Flexilevel 65 item CAT yields the same rank ordering of examinees as the other technique. In this regard, it may be considered a matter of indifference whether the examinee is administered a Bayesian CAT, a Flexilevel 65 item CAT, or a conventional 129 item test.

The results showed that number correct score was highly related to the known true thetas. Therefore, it is reasonable to continue to use this method of estimating examinee ability (as Lord uses in his paper-and- pencil version of a Flexilevel test).

The programming required to implement a computerized Flexilevel test requires nowhere near the sophistication necessary for implementing a Bayesian or maximum likelihood estimation based CAT system. An introductory computer science course's textbook would present all the programming concepts necessary to code a Flexilevel CAT. Furthermore, one could implement a Flexilevel CAT on a microcomputer with two disk drives (320K minimum) and a 128K of RAM (a standard IBM PC comes with 256K of RAM and two 360K drives). Of course more memory and/or a hard disk would allow greater flexibility with respect to implementation, but the point is that a Flexilevel CAT can be implemented on a very (inexpensive) basic system.

Therefore, neither the software nor the hardware needed for Flexilevel CAT methods present the rather serious obstacles that they often do for IRT-based CAT. This fact and the result that Flexilevel CAT ability estimates compare favorably with the IRT-based CAT ability estimates indicate that, for the first time, CAT procedures can be made practical and feasible for ordinary classroom testing.

REFERENCES

Jensema, C.J. (1974). The Validity of Bayesian Tailored Testing. Educational and Psychological Measurement, 34, 757-766.

Lord, F. M. (1971). The Self-Scoring Flexilevel Test. Journal of Eductional Measurement, 8, 147-151.

Lord, F. M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale, N.J.: Erlbaum, 1980.