A SIMULATION STUDY OF COMPUTERIZED ADAPTIVE TESTING WITH A MISSPECIFIED MEASUREMENT MODEL

George Engelhard, Emory University

ABSTRACT

Computerized adaptive testing provides an approach to measurement in the social and behavioral sciences which can be used to individualize and tailor a unique set of test items for each individual. Computerized adaptive testing methods provide a means for selecting and administering test items that provide the best information about the examinee's location on a latent variable. The methods used for computerized adaptive testing are basically sequential designs for estimating a quantal response curve which relates the characteristics of test items to the probability of an examinee responding correctly. Two major measurement models, one proposed by Rasch (1960) and another proposed by Birnbaum (1968), can be used to represent this response curve. The effects of using a one-parameter item response model (Rasch) when the simulated responses of the examinees are generated by a two-parameter item response model (Birnbaum) were examined in this study. The results of the simulations suggest that the effects of "misspecifying" the measurement model - using the Rasch model when the simulated examinee responses were generated by the Birnbaum model - can be minimized by using a robust estimate of ability based on the principle of Tukey's biweight which was proposed by Mislevy and Bock (1982). Increasing the number of items administered in the computerized adaptive testing session does not appear to be as effective as the use of the robust estimate of ability.

1. INTRODUCTION

In the future, many of the standard paper-and-pencil tests that have played a major role in social science measurement will be administered by computer. When test items are administered by computer, it becomes possible to deal with some of the disadvantages which are inherent with paper-and-pencil tests. One such disadvantage is that each individual has to respond to all of the items on the test which involves administering clearly inappropriate items to some individuals. In the area of school achievement, the typical test includes items that are too easy for some individuals and too hard for other individuals. By requiring examinees to respond to inappropriate items, the probability of measurement error increases. Computerized adaptive testing provides an approach which can be used to individualize and tailor a unique set of items that are in the appropriate range of difficulty for all examinees. This provides the opportunity to achieve more reliable and valid measurement in the social sciences.

Many of the approaches that have been proposed for implementing computerized adaptive testing can be viewed as sequential designs for dichotomous data. These sequential designs have been used in a variety of dichotomous experiments that range from the testing of explosives (Dixon and Mood, 1948) to biological assays that model the probability of a test animal surviving at various doses levels (Finney, 1971). Wu (1985) provides a current review of the work on efficient sequential designs for binary data.

There are three major differences between the measurement model used in computerized adaptive testing and the typical quantal response model. The first difference is that the independent variable is generally known for quantal response models, while in the measurement model the independent variable is a latent variable and must also be estimated. A second difference is that each dichotomous experiment in the computerized adaptive testing session corresponds to a sample size of one. Each examinee takes a single item which is then scored right (1) or wrong (0), and on the basis of this response another item is selected - a harder item if the examinee succeeds on the item, and an easier item if the examinee fails. A third difference is that the design points are constrained within the typical computerized adaptive testing session. In other words, the value of the independent variable which is optimal in some sense may not be available in the form of a test item of the appropriate difficulty.

The purpose of this study is to examine the robustness of one approach to computerized adaptive testing. There are two major item response models which have been proposed and are used extensively in psychometrics. Both of these models are based on logistic item response curves for representing the relationship of several item parameters and examinee ability with the probability of success or failure on an item. Birnbaum (1968) proposed using two item parameters - item difficulty and discrimination - to model this relationship. Rasch (1960) argued for the use of only one item parameter - item difficulty - with the assumption that the item discrimination parameters are equal.

In this study, the effects of using the simpler one-parameter measurement model for computerized adaptive testing when the item response data is actually generated by the two-parameter model will be systematically explored. The measurement model is considered "misspecified" in the sense that the two-parameter model is the "correct" model, and the robustness of using a simpler model - one less item parameter - is examined. Previous research on the Rasch measurement model has suggested that it is generally robust to violations in the assumption of equal or homogeneous item discrimination parameters (van de Vijver, 1986; Dinero & Haertel, 1977). For example, van de Vijer (1986) concluded ". . . even in small samples and for short tests, heterogeneity of the item discriminations

hardly affects the accuracy of Rasch estimates" (p. 55). The effects of violating the assumption of equal item discrimination parameters in a computerized adaptive testing setting has not been systematically explored.

2. DESCRIPTION OF A COMPUTERIZED ADAPTIVE TESTING SYSTEM

There are four major components that must be included in any computerized testing system. These are (1) a measurement model, (2) a method for estimating the examinee's ability based on the responses to previous items, (3) a method for selecting the next best item and (4) a set of rules for starting and stopping the testing session. Each of these components will be described below, and the approach used in the current computerized adaptive testing system outlined.

2.1 Two measurement models

One of the major problems in educational and psychological measurement is how to transform qualitative responses to a set of test items into a meaningful quantitative measure. An examinee's response is generally scored as correct (1) or incorrect (Ø). A measurement model is needed in order to represent the relationship between the dichotomous responses of the examinee to a set of items, and the underlying latent variable which the test items have been selected to represent. In education, the underlying latent variable might be reading comprehension, and a set of reading passages are selected to represent varying levels of difficulty on a reading comprehension scale.

Birnbaum (1968) proposed a measurement model giving the probability of success or failure on a test item as a function of two item parameters--item difficulty and item discrimination and the ability of the examinee. The probability of a correct response based on Birnbaum's measurement model can be expressed as

$$P_i = 1 / [1 + \exp(-B_i)] \qquad (1)$$

where

$$B_i = a_i (\theta - b_i) \qquad (2)$$

and a_i represents the discrimination parameter for item i, b_i is the difficulty parameter for item i, and θ represents the examinee's ability.

The measurement model proposed by Rasch (196Ø) for success on item i can be represented as follows

$$P_i = 1 / [1 + \exp(-R_i)] \qquad (3)$$

where

$$R_i = (\theta - b_i) \quad . \qquad (4)$$

The item discrimination parameters, a_i, which were included in the Birnbaum model, are viewed as being equal or homogeneous under the assumptions of the Rasch measurement model.

2.2 Estimation of ability

In the case of computerized adaptive testing, the item parameters have already been estimated on the basis of a large scale calibration study and can assumed to be known. If we assume that the responses of each examinee are independent, given θ, then the probability of a particular response vector for the examinee is

$$P = \prod_{i=1}^{n} P_i^{X_i} (1 - P_i)^{1 - X_i} \qquad (5)$$

where X_i represents the dichotomous response of the examinee to item i (1 for success and Ø for failure), P_i is given by (1) for the Birnbaum measurement model and by (3) for the Rasch measurement model, and n is the number of items previously administered to the examinee.

Since the item parameters are known and the examinee's ability is not known, (5) can be used to represent the likelihood function of θ for the examinee given the vector of item responses. The maximum likelihood estimate of the examinee's ability, θ, is the value which maximizes (5) with respect to the observed response vector.

In practice, the log of (5) is maximized and is given as

$$L = \sum_{i=1}^{n} X_i \ln(P_i) + (1 - X_i) \ln(1 - P_i). \qquad (6)$$

The first derivative of L with respect to examinee ability is

$$\frac{dL}{d\theta} = \sum_{i=1}^{n} (X_i - P_i) a_i \quad , \qquad (7)$$

while the second derivative is

$$\frac{d^2L}{d\theta^2} = - \sum_{i=1}^{n} P_i (1 - P_i) a_i^2 \quad . \qquad (8)$$

Newton-Raphson iterations can be used to obtain the maximum likelihood estimate of θ as follows

$$\theta^{K+1} = \theta^k - \frac{\sum_{i=1}^{n} (X_i - P_i^k) a_i}{- \sum_{i=1}^{n} (P_i^k (1 - P_i^k) a_i^2} \qquad (9)$$

where θ^k is an initial estimate of the examinee's ability and θ^{k+1} is the updated estimate. A good initial estimate is given by

$$\theta^k = \ln [s / (n - s)] + h \qquad (1Ø)$$

where s is the sum of the examinee's correct

responses, n is the number of items and h is the average difficulty of the preceding items. The Newton-Raphson iterations can be stopped after the differences between θ^{k+1} and θ^{k} become suitably small, such as .001 or less. Occasionally, the maximum likelihood estimates will not converge for some item response patterns and these examinees will require special treatment.

An estimate of the standard error for the ability estimate can be obtained as the negative reciprocal of the square root of the second derivative of L, after the estimate of θ has converged.

In addition to the maximum likelihood estimate of examinee ability, a robust estimate of ability based on Tukey's biweight is used. Mislevy and Bock (1982) proposed the following modification of the Newton-Raphson iterations

$$\theta^{k+1} = \theta^{k} - \frac{\sum_{i=1}^{n} w_i^{k}(X_i - P_i^{k})\, a_i}{-\sum_{i=1}^{n} w_i^{k}\, P_i^{k}\,(1 - P_i^{k})\, a_i^{2}} \qquad (11)$$

where

$$w_i^{k} = \begin{cases} [1 - (u_i^{k})^2]^2 & \text{for } |u_i^{k}| \leq 1 \\ 0 & \text{otherwise} \end{cases} \qquad (12)$$

and

$$u_i^{k} = [(b_i - \theta^{k})\, a_i] \,/\, 3 \,. \qquad (13)$$

Mislevy and Bock (1982) have successfully used these biweight estimates of examinee ability to produce robust estimates when the observed responses are subject to errors such as quessing or carelessness on the part of the examinee. In this study, the use of these biweight estimates to compensate for assuming that the item discrimination parameters are equal is explored. In the case of the Rasch model, the item discrimination parameters, a_i, in (7) through (13) can be set equal to 1.0 and both the maximum likelihood and robust estimates of examinee ability obtained.

2.3 Item selection

Birnbaum (1968) has shown that the best item in terms of information about the examinee's position on the latent variable is obtained by administering an item which has a difficulty value as close as possible to the initial estimate of examinee ability. The item with the smallest absolute difference between the preliminary ability estimate and the available item difficulties was selected as the next item to be administered to the simulated examinee. Since an item can only be administered once, an index was developed to keep track of item use.

2.4 Starting and stopping rules

There are several approaches which can be used for starting the computerized adaptive testing session. In this study, an initial estimate of examinee ability was obtained by using a 5 item routing test with fixed item difficulties (-2.0, -1.0, 0.0, 1.0, 2.0) which was given to every simulated examinee.

In order to stop the computerized adaptive testing session, the number of items administered was fixed in advance. Since the effect of number of items on the estimates of ability was of interest, the estimates obtained on the basis on 10, 15 and 20 items were used.

In this study, the Rasch measurement model is used to estimate the examinees's ability based on the previously administered items, the Rasch model is used as the basis for selecting the next item to be administered to the examinee and the responses of the examinee to each item are simulated on the basis of the Birnbaum measurement model. The measurement model is misspecifed in the sense that the Rasch model is used to represent the responses of the examinees when the actual simulated responses are constructed on the basis of the Birnbaum model.

3.0 A SIMULATION STUDY

3.1 Description

There are always a variety of factors that can be varied in a simulation study. In this study, the focus is on how well a computerized adaptive testing system based on the Rasch model will recover the generating ability when the responses of the simulated examinees are generated on the basis of a different measurement model - the Birnbaum measurement model. In order to simulate this situation, a set of item parameters were generated. A set of 50 item difficulty parameters, b_i, uniformly distributed between -3.0 and +3.0 were generated and used throughout the study. Fifty item discrimination parameters, a_i, uniformly distributed between .5 and 1.5 were also generated and combined with the item difficulty parameters to yield a set of 50 items. The generating ability parameters were set at -2.0, -1.0, 0.0, 1.0 and 2.0.

Using the 50 item parameters described above, the responses of the simulated examinees were generated on the basis of the Birnbaum measurement model. The probability of succeeding on an item was based on (1) and the obtained probability was compared to a uniformly distributed random number between 0 and 1. If the probability was greater or equal to the random number, then the simulated examinee "succeeded" on the item (X = 1) and if the probability was less than the random number, then the simulated examinee "failed" on the item (X = 0).

3.2 Evaluation indices

Two indices were computed in order to examine how well the generating ability parameters were recovered by the computerized adaptive testing system. The first index is the root mean square error (RMSE) which is the square root of the average squared deviation of the estimate from the generating ability parameter. The average squared deviation or mean square error is the sum of the estimate's variance and the square of its bias. The second

index is the mean signed difference (MSD) which is simply the average signed difference of the estimate from the generating parameter. This MSD provides an index of the bias in the estimation procedure. There were 200 simulated computerized adaptive testing sessions with 10 replications per ability level.

3.3 Example

An example of a computerized adaptive testing session is given in Table 1. The generating ability was 0.0 and the item discrimination parameters were set to 1. A five item routing test was administered to the examinee (items 1 to 5) with item difficulties of -2.00, -1.00, 0.00, 1.00 and 2.00. The examinee succeeded on items 1 and 2 as expected because they are below the generating ability level of 0.0, and he was also able to succeed on item 4 which is unexpected because this item is above the ability level of the examinee. The initial estimate of the examinee's ability is (.59) and the standard error is (1.10). On the basis of this initial estimate the next best item is 27 with a difficulty of .71. The examinee succeeds on this item and the estimate of the ability level increases to 1.09 with a standard error of .99. The next item is 34 with a difficulty of 1.11 and the examinee fails on this item which results in a decrease in the estimate of the examinee's ability to .39. This process continues - estimation of ability, selection of the next most appropriate item and administration of this item - until the final estimate of the examinee's ability of 0.0 and a standard error of .49 is obtained.

3.4 Results

The root mean square errors (RMSE) and mean signed differences (MSD) are given in Table 2. In order to set up a standard of comparison for the root mean square errors (RMSE), the results of the 10 item computerized adaptive testing sessions using the maximum likelihood estimates of ability with the item discrimination set equal to one were used. The results of the simulated computerized adaptive testing sessions are discussed in terms of five ability groupings--very low, low, average, high and very high.

For examinees with very low ability (θ = -2.00), the RMSE for the maximum likelihood estimates is .47 and increasing the number of items does not seem to have much of an effect. The biweight estimates for the very low ability examinees have a higher RMSE when 10 items are administered, but by the time 15 items have been administered the RMSE is comparable to the standard. When the item discrimination parameters are allowed to vary between .5 and 1.5, the maximum likelihood estimates for the very low ability examinees have RMSEs which are almost twice as large as the standard of comparison and the RMSEs do not decrease as the number of items increases. In the case of the biweight estimates of ability ($.5 \leq a_i \leq 1.5$), the RMSEs for the very low ability examinees are essentially the same as the maximum likelihood estimates of ability when the item discrimination parameters are equal.

The RMSE for examinees with low abilities (θ = -1.00) is .51 and this remains relatively constant as the number of items is increased. The biweight estimates for the low ability examinees when ten items are administered is lower than the standard, but as the number of items increases the RMSEs become comparable. When the item discrimination parameters are allowed to vary, the RMSEs for the maximum likelihood estimates of low ability examinees are almost twice as large when ten items are administered and become comparable by the time 15 items have been administered. The biweight estimates ($.5 \geq a_i \geq 1.5$) also have a larger RMSE for 10 items, but it is not quite as high as the ML estimates. The RMSEs for 15 and 20 items are comparable to the standard.

The RMSE for the examinees of average ability (θ = 0.0) is .80 and this decreases as the number of items increases. The biweight estimates (a_i = 1) for the examinees with average ability perform better than the standard when 10 items are administered, and the RMSEs decrease as the number of items is increased. The ML estimates of ability ($.5 \geq a_i \geq 1.5$) for examinees of average ability perform slightly better than the standard when 10 items have been administered, and significantly better as the number of items increases. The biweight estimates for the examinees of average ability perform almost the same as the standard.

Turning to the high ability examinees (θ = 1.00), the RMSE is .59 for the maximum likelihood estimates (a_i = 1) based on 10 items and this decreases slightly as the number of items increases. The biweight estimates (a_i = 1) have comparable RMSEs. The RMSEs for the ML estimates ($.5 \geq a_i \geq 1.5$) are similar to the standard, while the RMSEs for the biweight estimates ($.5 \geq a_i \geq 1.5$) are comparable when 10 items are administered, but the RMSEs seem to decline significantly as the number of items is increased.

The RMSE for the very high ability examinees (θ = 2.00) using maximum likelihood estimation (a_i = 1) is .75, and the RMSE decreases as the number of items increases. The biweight estimates (a_i = 1) perform much better than the ML estimates. The ML estimates when the item discrimination parameter is allowed to vary have RMSEs very similar to the standard, although there does seem to be a slight increase in the RMSEs as the number of items increases. The biweight estimates ($.5 \geq a_i \geq 1.5$) for very high ability examinees perform much better than the ML estimates even when the assumption of equal item discriminations is true.

4. CONCLUSIONS

Computerized adaptive testing provides an approach to measurement which allows each examinee to respond to a unique set of test items which are individually tailored to be in the appropriate range of difficulty. Research

on adaptive tests has indicated that this approach to testing can yield significant improvements in measurement quality and efficiency which can result in equal measurement precision at all ability levels (Weiss, 1982). This increase in precision is usually obtained with fewer items than are typically administered in the case of standard paper-and-pencil tests.

In this study, the robustness of a computerized testing system developed on the basis of a one-parameter model item response model (Rasch, 1960) was examined. The effects of "misspecifying" the item response model--simulated responses were generated on the basis of a two-parameter item response model (Birnbaum, 1968)--on the estimation of ability was systematically explored. The results of this small scale simulation study suggest that the maximum likelihood estimates obtained by using the one-parameter model were sensitive to a lack of homogeniety in the item discrimination parameters. This finding contrasts with the conclusions reached when the robustness of the Rasch model has been explored in a non-adaptive testing situation. Even though the maximum likelihood estimates did not perform well in terms of the RMSE, a robust estimator based on Tukey's biweight performed very well.

The use of robust estimators in the case of computerized adaptive testing within the context of the Rasch measurement model appears to be a promising approach for adjusting for bias which is introduced when the assumption of homogeneous item discrimination parameters is not met. One potential problem with the use of robust estimators in computerized adaptive testing situations is that the initial estimates of ability become very important. If a poor starting value is chosen, then the biweight estimator is "conservative" and it may take more items to converge on the generating ability. Further research is needed on the use of robust estimators in the context of computerized adaptive testing.

REFERENCES

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing Company.

Dinero, T.E. & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. Applied Psychological Measurement, 1, 581-592.

Dixon, W.J. & Mood, A.M. (1948). A method for obtaining and analyzing sensitivity data. Journal of American Statistical Association, 43, 109-126.

Finney, D.J. (1971). Probit analysis. Third edition. London: Cambridge University Press.

Mislevy, R.J. & Bock, R.D. (1982). Biweight estimates of latent ability. Educational and psychological measurement, 42, 725-737.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: Dansmarks Paedagogiske Institut (Reprinted 1980, Chicago: The University of Chicago Press.)

van de Vijver, F. (1986). The robustness of Rasch estimates. Applied Psychological Measurement, 10, 45-57.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.

Wu, C.F.J. (1985). Efficient sequential designs with binary data. Journal of the American Statistical Association, 80, 974-984.

Table 1

Example of a Computerized Adaptive Testing Session with Maximum Likelihood Estimates of Examinee Ability (Generating Values - Theta = 0.0, a = 1)

Index	Item	Difficulty (b_i)	Disc (a_i)	Response	Theta (Θ)	SE
1	1	-2.00	1.00	1		
2	2	-1.00	1.00	1		
3	3	0.00	1.00	0		
4	4	1.00	1.00	1		
5	5	2.00	1.00	0	(.59)	(1.10)
6	27	.71	1.00	1	1.09	.99
7	34	1.11	1.00	0	.71	.87
8	28	.71	1.00	0	.39	.80
9	26	.45	1.00	0	.12	.76
10	24	.11	1.00	1	.36	.70
11	25	.31	1.00	0	.14	.67
12	23	-.21	1.00	1	.30	.63
13	29	.74	1.00	0	.16	.61
14	22	-.37	1.00	0	-.06	.59
15	21	-.44	1.00	1	.07	.56
16	30	.78	1.00	1	.27	.54
17	31	.91	1.00	0	.17	.52
18	32	.91	1.00	0	.09	.51
19	20	-.69	1.00	0	-.08	.50
20	19	-.80	1.00	1	.00	.49

Table 2

Root Mean Square Error (RMSE) and Mean Signed Difference (MSD) by Generating Ability and Number of Items

		RMSE			MSD		
Method	Disc	10	15	20	10	15	20
a. Very low ability (Theta = -2.0)							
ML	1.0	.47	.49	.47	.13	.01	-.18
BIW	1.0	.77	.55	.61	.01	-.01	-.01
ML	.5 - 1.5	.91	.87	.86	.22	.33	.10
BIW	.5 - 1.5	.47	.42	.52	.14	.11	.03
b. Low ability (Theta = -1.0)							
ML	1.0	.51	.63	.50	.15	.07	.10
BIW	1.0	.38	.54	.52	-.20	-.21	-.16
ML	.5 - 1.5	.90	.57	.61	-.10	-.06	.10
BIW	.5 - 1.5	.68	.53	.50	.36	.18	.06
c. Average ability (Theta = 0.0)							
ML	1.0	.80	.51	.53	.45	.16	-.10
BIW	1.0	.61	.45	.48	.13	.08	.10
ML	.5 - 1.5	.73	.49	.37	-.08	-.02	.03
BIW	.5 - 1.5	.71	.53	.57	.16	.14	.26
d. High ability (Theta = 1.0)							
ML	1.0	.59	.52	.53	-.22	-.27	-.31
BIW	1.0	.52	.44	.35	.03	.10	.18
ML	.5 - 1.5	.63	.68	.48	-.12	-.27	-.16
BIW	.5 - 1.5	.69	.37	.30	-.21	.10	.07
e. Very high ability (Theta = 2.0)							
ML	1.0	.75	.71	.56	.08	-.08	-.08
BIW	1.0	.42	.46	.28	-.25	-.01	-.03
ML	.5 - 1.5	.74	.73	.79	.05	.22	.31
BIW	.5 - 1.5	.55	.44	.45	-.30	-.12	-.03