DOCUMENT RESUME

ED 352 367 TM 019 120

AUTHOR De Ayala, R. J.
TITLE The Nominal Response Model in Computerized Adaptive Testing.
PUB DATE [92]
NOTE 39p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; Comparative Testing; *Computer Assisted Testing; Computer Simulation; Equations (Mathematics); Estimation (Mathematics); Item Banks; *Item Response Theory; *Mathematical Models; Test Items; Test Length
IDENTIFIERS Ability Estimates; *Nominal Response Model; Three Parameter Model

ABSTRACT

One important and promising application of item response theory (IRT) is computerized adaptive testing (CAT). The implementation of a nominal response model-based CAT (NRCAT) was studied. Item pool characteristics for the NRCAT as well as the comparative performance of the NRCAT and a CAT based on the three-parameter logistic (3PL) model were examined. Ability estimates were generated at test lengths of 10, 15, 20, 25, and 30 items from item pools of 90 items. Abilities were generated for 1,300 examinees in 1 study and for 900 examinees in the other study. Results show that for 2-, 3-, and 4-category items, items with maximum information of at least 0.16 produced reasonably accurate ability estimation for tests with a minimum test length of about 15 to 20 items. Moreover, the NRCAT was able to produce ability estimates comparable to those of the 3PL CAT. Implications of these results were discussed. Eight tables and six graphs illustrate the discussion. (SLD)

The Nominal Response Model in Computerized Adaptive Testing

R.J. De Ayala

University of Maryland

Please forward correspondence to :

R.J. De Ayala

Department of Measurement, Statistics and Evaluation

Benjamin Building, Rm 1230F

University of Maryland

College Park, MD 20742

ABSTRACT

One important and very promising application of item response theory (IRT) is computerized adaptive testing (CAT). Although most CATs use dichotomous IRT models, research on the use of polytomous IRT models in CAT has shown promising results. This study concerned the implementation of a nominal response model-based CAT (NR CAT). Item pool characteristics for the NR CAT as well as the comparative performance of the NR CAT and a CAT based on the three-parameter logistic (3PL) model were examined. Results showed that for two-, three-, and four-category items, items with maximum information of at least 0.16 produced reasonably accurate ability estimation for tests with a minimum test length of about 15 to 20 items. Moreover, the NR CAT was able to produce ability estimates comparable to those of the 3PL CAT. Implications of these results were discussed.

One important and very promising application of item response theory (IRT) is computerized adaptive testing (CAT). Unlike the conventional paper-and-pencil test in which an examinee, regardless of ability, is administered all test items, CAT is a procedure for administering tests which are individually tailored for each examinee. The advantage of IRT-based CAT over paper-and-pencil testing have been well documented (e.g., Wainer, 1990; Weiss, 1982).

Although not necessary (cf., De Ayala, Dodd, & Koch, 1990), a CAT system typically uses an IRT model in combination with test item characteristics to estimate the examinee's ability. Typically, either the dichotomous three-parameter logistic (3PL) or Rasch models (e.g., McBride & Martin, 1983; Kingsbury & Houser, 1988) have been used in CAT. These models do not differentiate between an examinee's incorrect answer and other incorrect alternatives for purposes of ability estimation. In short, dichotomous models and dichotomous model-based CATs operate as if an examinee either knows the correct answer or randomly selects an incorrect alternative.

The operation of dichotomous model-based CATs do not incorporate findings from human cognition studies (e.g., Brown & Burton, 1978; Brown & VanLehn, 1980; Lane, Stone, & Hsu, 1990; Tatsuoka, 1983). For instance, Tatsuoka's (1983) analysis of student misconceptions in performing mathematics problem showed that wrong responses could be of more than just one kind, however, dichotomous scoring uniformly assigned a score of zero to all the wrong responses. Moreover, it has been demonstrated by Nedelsky (1954), from a classical test theory (CTT) perspective, and Levine and Drasgow (1983), from an IRT perspective, that the distribution of wrong answers over the options of multiple-choice items differed across ability levels. In this regard, an item's incorrect alternatives may augment our estimate of an examinee's ability by providing information about the examinee's level of understanding (i.e., provide diagnostic information). Both Bock (1972) and Thissen (1976) have found that for examinees with ability estimates in the lower half of the ability range the nominal response (NR) model provided from one third to nearly twice the information furnished by a dichotomously scored two-parameter model; there was no difference in information yield between these two models for ability estimates above the median θ. It should be noted that in an application to multiple-choice and free-response items, Vale and Weiss (1977) found that the NR model provided more information for middle ability examinees than that shown in the Bock (1972) and Thissen (1976) studies. In CTT, the use of proper scoring techniques to assess this partial knowledge yields increases in the reliability of multiple choice tests (e.g., Coombs, Milholland, and Womer, 1956). Frary (1989), Haladyna and Sympson (1988), and Wang and Stanley (1970) all provide a review of the literature on option scoring strategies. It is obvious that the

dichotomization of the examinee's response ignores any partial knowledge that the examinee may have of the correct answer and, as a result, this information cannot be used for ability estimation.

Some research has explored the benefits and operating characteristics of CATs based on polytomous IRT models (e.g., Dodd, Koch, & De Ayala, 1989; Koch & Dodd, 1989; Sympson, 1986). Research on the use of polytomous IRT models in CAT has shown promising results. For instance, Sympson (1986) found that adaptive tests based on a polytomous model (Model 8) could be shortened by 15-20% without sacrificing test reliability. In addition, these studies have shown that item pools smaller than those used with dichotomous model-based CATs have led to satisfactory estimation, that the use of the ability's standard error of estimation for terminating the adaptive test is preferred to the minimum item information termination criterion, and that the use of a variable stepsize instead of a fixed stepsize tends to minimize nonconvergence of trait estimation; the models under study were Masters's (1982) partial credit (PC), Andrich's (1978) rating scale (RS), and Samejima's (1969) graded response (GR) models.

Bock's (1972) NR model is appropriate for items with unordered responses, such as multiple-choice aptitude and achievement test items. In addition, the NR model may be used with testlets (Wainer & Kiely, 1987) to solve various testing issues, such as multidimensionality (Thissen, Steinberg, & Mooney, 1989), with items which do not have a "correct" response, such as demographic items (e.g., to provide ancillary information), and items whose alternatives provide educational diagnostic information. Moreover, innovative computerized item formats may be specifically developed for use with polytomous models and adaptive testing environments. Presently, CATs typically present simple paper-and-pencil item formats.

The objectives of this study concerned the implementation of an NR model-based CAT (NR CAT) and were three-fold. First, because the NR model is written in terms of slope and intercept parameters, a form not typically used (cf., Hambleton & Swaminathan, 1985; Lord, 1980; Weiss, 1983), formulae for the location parameters were derived in order to facilitate understanding the model's formulation. In this regard, the NR model's relationship with the dichotomous two-parameter logistic (2PL) model was presented. Moreover, because of the importance of item information in CAT, the effect of varying the location parameters on the distribution of item information was examined. Second, paramount to CAT performance is the quality of the item pool. Two factors which determine the item pool's quality are the locations of the item and their discrimination indices. Because it is accepted that items should be evenly and equally distributed throughout the θ continuum of interest (Patience & Reckase, 1980; Urry, 1977; Weiss,

1982) and there is no reason to believe that this would not hold for the NR model, this factor was not studied. However, the minimum item information (i.e., the discrimination indices' effect) which would allow reasonably accurate ability estimates by the NR CAT was investigated. This investigation (referred to as Study 1) was limited to the 2-, 3-, and 4-category cases. Third, the comparative performance of the NR CAT and a CAT based on a dichotomous (3PL) model was assessed (referred to as Study 2). Furthermore, because of the existence of option information an exploratory simulation was conducted in which items were selected on the basis of option information.

Model

The NR model assumes that item alternatives represent responses which are unordered. The NR model provides a direct expression for obtaining the probability of an examinee with ability θ responding in the j-th category of item i as:

$$p_{ij}(\theta) = \frac{\exp(c_{ij} + a_{ij}\theta)}{\sum_{h=1}^{m_i} \exp(c_{ij} + a_{ij}\theta)} , \qquad (1)$$

where a_{ij} is the slope parameter, c_{ij} is the intercept parameter of the nonlinear response function associated with the j-th category of item i, and m_i is the number of categories of item i (i.e., $j = 1, 2, ..., m_i$). For convenience the slope and intercept parameters are sometimes represented in vector notation, where $\mathbf{a} = (a_{i1}, a_{i2}, ..., a_{im})$ and $\mathbf{c} = (c_{i1}, c_{i2}, ..., c_{im})$, respectively. As an aide to interpreting these parameters a logistic space plot of the (multivariate) logit (i.e., $c_{ij} + a_{ij}\theta$) against θ for a three-category ($m = 3$) item with $\mathbf{a} = (-0.75, -0.25, 1.0)$ and $\mathbf{c} = (-1.5, -0.25, 1.75)$ is shown in Figure 1. As can be seen, the c_{ij}'s value is the y-intercept (i.e., $\theta = 0.0$) and a_{ij} is the slope of the category's response function. The a_{ij}s are analogous to and have an interpretation similar to traditional option discrimination indices. That is, a crosstabulation of ability groups by item alternatives shows that a category with a large a_{ij} reflected a response pattern in which as one progressed from the lower ability groups to the higher ability groups there was a corresponding increase in the number of persons who answered the item in that category and for categories with negative a_{ij}s this pattern was reversed. Moreover, it appears that, in general, large values of c_{ij} are associated with categories with large frequencies and as the value of c_{ij} becomes increasingly smaller the frequencies for the corresponding categories decrease.

Insert Figure 1 about here

The probability of responding in a particular category as a function of θ is depicted by the category or option characteristic curve (OCC). Figure 2 contains the OCCs corresponding to the three category item presented in Figure 1.

Insert Figure 2 about here

The intersection of the OCCs can be obtained by setting adjacent category multivariate logit equal to one another and solving for θ. Therefore,

$$\theta = \frac{c_1 - c_2}{a_2 - a_1} \quad . \tag{2}$$

In general, for any item with $m_i \geq 2$ and because θ and b are on the same scale:

$$b = \frac{c_{(j-1)} - c_j}{a_j - a_{(j-1)}} \quad . \tag{3}$$

This formulation is analogous to that of the PC model in which step difficulties are defined at the intersection of adjacent category characteristic curves.

In Bock (1972) the NR model is compared with a binary version (i.e., the item consists of correct and incorrect categories). When $m_i = 2$ then (1) becomes,

$$p_2(\theta) = \frac{\exp(c_2 + a_2\theta)}{\exp(c_1 + a_1\theta) + \exp(c_2 + a_2\theta)} \quad , \tag{4}$$

Given (4) and noting that the two linear constraints imposed on the item parameters, $\sum a=0$ and $\sum c=0$ (to address the indeterminacy of scale), imply that in the two-category case

$$a_1 = -a_2 \text{ and} \tag{5}$$

$$c_1 = -c_2. \tag{6}$$

Therefore, given (5) and (6) one obtains that for $m_i = 2$

$$b = -\frac{c_2}{a_2} \quad . \tag{7}$$

Solving (7) for c_2 and substituting the equality into (4),

$$p_2(\theta) = \frac{\exp(-2a_2b + a_2\theta)}{\exp(-2a_2b + a_2\theta) + \exp(a_1\theta)} \quad . \tag{8}$$

By substitution of (5) into (8), and simplifying, one obtains

$$p_2(\theta) = \{1 + \exp(-2a_2(\theta - b))\}^{-1} \quad . \tag{9}$$

Therefore, if one casts the NR model's discrimination parameters in terms of the 2PL model's discrimination parameter, a, and because a is typically positive:

$$a = |-2a_2| = |2a_1| \quad ; \tag{10}$$

for $m_i = 2$ the 2PL and NR models are equivalent. For example, Figure 3 shows the NR model's OCCs for an item with $a_2 = 0.40$, $a_1 = -0.40$, $c_2 = 0.2$ and $c_1 = -0.20$ and the item characteristic curve (ICC) for the 2PL model with $a = 0.80$ and $b = -\frac{0.4}{0.8} = -0.5$.

Insert Figure 3 about here

Information

For the NR model, the item information ($I_i(\theta)$) is equal to the sum of the option informations, where option information may be defined as (Bock, 1972)

$$I_{ij}(\theta) = \mathbf{aWa}'p_{ij}(\theta)\,, \tag{11}$$

and item information is

$$I_i(\theta) = \sum_{h=1}^{m_i} \mathbf{aWa}'p_{ij_j}(\theta) = \mathbf{aWa}'\,. \tag{12}$$

Where for a given item i,

$$\mathbf{W} = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_m \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_m \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ -p_mp_1 & -p_mp_2 & \cdots & p_m(1-p_m) \end{bmatrix}$$

For the $m_i = 2$ case, the location of maximum item information (I_{max}) is $\theta_{max} = \frac{c_1 - c_2}{a_2 - a_1}$ with $I_{max} = 0.25(a_2 - a_1)^2$. Due to the number of unknowns a formula for the location of maximum item information cannot be determined for $m_i > 2$. When $m_i = 2$ and for a given a changing the values of **c** forces the location of I_{max} to shift along the θ continuum, but the maximum amount of information remains constant.

For the $m_i = 3$ case and for a given **a**, if the *b*s are in ascending order, then the item information function becomes comparatively more leptokurtic as the difference between *b*s become less extreme. When the *b*s are in descending order, then item information function becomes comparatively more platykurtic as the difference between *b*s become less extreme. In both cases there is also a shifting in the location of I_{max}.

For the $m_i = 4$ case and for a given **a**, if the *b*s are in ascending order, then the item information function becomes comparatively more platykurtic as the difference between *b*s become less extreme. This pattern holds if one reverses the last two *b*s. When the *b*s are in descending order, then relative to the item information function when the *b*s are in ascending order, the function becomes more leptokurtic as the difference between *b*s become less extreme. This is also true if one transposes the first two *b*s. For the other two possible *b* patterns, the information function becomes comparatively more leptokurtic as the distance among the *b*s decreases. Moreover, it is possible to obtain bimodal item information functions. For instance, Figure 4 contains the information function for an item where **a** = (1, 0.1, -0.1, -1) and **c** = (0.1, 2.4, -2.6, 0.1).

Insert Figure 4 about here

As (12) implies item information is a function of the magnitude of the elements of **a** and the order of the elements of **a** (i.e., for a given **c**, **a** = (-0.25, 1.0, -0.75), **a** = (-0.25, -0.75, 1.0) and **a** = (-0.75, -0.25, 1.0) will produce three different I_{max}s at three different θ_{max}s. For a given **a** the signs of the elements are irrelevant as long as $\sum a=0$ (and $\sum c=0$). For instance, given two items with the same **c** (e.g., **c** = (0.25, -0.15, -0.1)) but **a**s which differ only in the sign of the elements, such as **a** = (0.4, 0.25, -0.65) and **a** = (-0.4, -0.25, 0.65), the items will have the same I_{max} = 0.245 but at different θ_{max}s; specifically, θ_{max} = 0.83985 for **a** = (-0.4, -0.25, 0.65) and for **a** = (0.4,0.25,-0.65) θ_{max} = -0.83985. This is also true in the four category case. Given the same **c**, two items whose **a**s differ only in the sign of the elements (and satisfy $\sum a=0$), such as **a** = (0.55, 0.4, -0.35, -0.6) and **a** = (-0.55, -0.4, 0.35, 0.6) will yield I_{max} = 0.258679 at θ_{max} = 0.059 and θ_{max} = -0.059, respectively.

METHOD

Study 1: Determination of Minimum Item Information for use in NR CAT

Programs: A program for performing adaptive testing with the NR model was written (NR CAT). The program used expected a posteriori (EAP) estimation (Bock & Mislevy, 1982) of ability and item selection was on the basis of information. The adaptive testing simulation was terminated when a maximum of thirty items was reached. Ability estimates at test lengths of 10, 15, 20, 25 and 30 items were recorded. The initial ability estimate for an examinee was the population's mean and a uniform prior with ten quadrature points was used. An additional program for generating the data according to the NR model was written and is discussed below.

Data: A series of item pools were created. The item pools differed from one another on the basis of two factors, maximum item information, I_{max}, and the number of item alternatives, m: 2, 3, and 4 options. The item pool size was 90 items (cf., Dodd, Koch, & De Ayala, 1989; Koch & Dodd, 1989).

Although Urry's (1977) guidelines for the discrimination parameter were stated in terms of *a*'s magnitude, the importance of an item's *a* value is its effect on I_{max}. Because when the number of categories is three or more different combinations of **a** and **c** can produce the same I_{max} value, establishing guidelines in terms of the magnitude elements of these vectors was not pursued. Rather, specified values for I_{max} were set a priori and the **a** vector to obtain a specific I_{max} was determined. The I_{max} values studied were 0.25, 0.16, 0.09, and 0.04.

When m_i = 2, the **a** vectors may be specified a priori. For the I_{max} values of 0.25, 0.16, 0.09, and 0.04, the corresponding **a**s were (0.50, -0.50), (0.40, -0.40), (0.30, -0.30), and

(0.20, -0.20), respectively. (For the 2PL model these as are equivalent to as of 1.0, 0.8, 0.6, and 0.4, respectively.) Because Urry (1977) has recommended the use of items with $a \geq 0.80$ in CAT, for the $m_i = 2$ condition the use of $\mathbf{a} = (0.40, -0.40)$ was expected to be equivalent to the use of $a = 0.80$ with a 2PL model-based CAT. For each I_{max} level of the $m_i = 3$ and $m_i = 4$ conditions the $\mathbf{a}$ vectors for the items were chosen through a trial-and-error procedure to approximate the relevant I_{max} value.

A number of researchers have stated that the item bs should be evenly distributed throughout the θ range of interest (e.g., Patience & Reckase, 1980; Urry, 1977; Weiss, 1982). Therefore, item b(s) were distributed at nine scale points between -4.0 to 4.0 in increments of 1 logit (i.e., for item 1 $b = -4.0$, for item 2 $b = -3.0$, etc.); for the $m_i > 2$ conditions the average location for an item was set at one of the nine scale points.

Once the $\mathbf{a}$ vector for a given I_{max} level was determined, then the $\mathbf{c}$ vector to locate the items, in terms of its b (for $m_i = 2$) or average b (for $m_i > 2$), at the specified scale points could be calculated. Therefore, these item sets consisted of 9 items with a constant maximum information which were distributed to encompass the examinee ability range. These 9 items were replicated to produce a 90-item pool for each of the 12 combinations of the 4 I_{max} levels crossed by the 3 m_i levels. De Ayala, Dodd, & Koch (1990) found that multiple items with the same parameters were administered to an examinee as the CAT estimation algorithm approaches its final ability estimate.

Thirteen hundred examinees' abilities were generated to be evenly distributed between -3.0 and 3.0 using a one-half logit interval between successive θ levels (i.e., for 100 examinees $\theta = -3.0$, for 100 examinees $\theta = -2.5$, etc.). These true θs (θ_Ts) plus the 90 item parameters for each condition were used to generate polytomous response strings with a random error component for each simulated examinee (i.e., 12 response data sets were created). Generation of an examinee's polytomous response string was accomplished by calculating the probability of responding to each alternative of an item according to the NR model. Based on the probability for each alternative, cumulative probabilities were obtained for each alternative. A random error component was incorporated into each response by selecting a random number from a uniform distribution [0,1] and comparing it to the cumulative probabilities. The ordinal position of the first cumulative probability which was greater than the random number was taken as the examinee's response to the item.

Analysis: The focus of Study 1 was to determine the minimum I_{max} value which would result in a significant improvement in the estimation of ability. The accuracy of ability estimation was assessed by root mean square error (RMSE) and Bias. RMSE and Bias were calculated according to:

$$RMSE(\theta) = \sqrt{\frac{\sum (\hat{\theta}_k - \theta_T)^2}{n_f}} \quad (13)$$

$$Bias(\theta) = \frac{\sum (\hat{\theta}_k - \theta_T)}{n_f} , \quad (14)$$

where $\hat{\theta}_k$ is the ability estimate for examinee k with latent ability θ_T, and n is the number of examinees at interval f (i.e., n_f = 100).

The analysis of the 2-, 3- and 4-category cases were treated as separately. Therefore, the basic design is a one-group repeated measures with two dependent variables, RMSE and Bias, with I_{max} as the between subjects factor and test length as the within subjects factor. The test length factor was included because the accuracy of ability estimation is influenced by both the adaptive test length as well as the information content of the items administered. Because the Bonferroni method was used to control for familywise Type I error, α was set at 0.008 (i.e., 0.05/6). Post hoc analysis was performed with the Scheffé test using a critical F of 13.2595 ($=(\upsilon_1)F_{0.008,\ 3,\ 48}$). Descriptive statistics on the adaptive tests were calculated.

Study 2: Comparative performance of the NR and 3PL CATs

Programs: The NR CAT program from Study 1 was used in Study 2; the NR CAT could select items on the basis of either item or option information. An additional CAT program based on the 3PL model (3PL CAT) was written. The 3PL CAT program estimated ability through EAP and selected items on the basis of information. The adaptive testing simulation was terminated when either of two criteria were met: a maximum of thirty items was reached or when a predetermined standard error of estimate (SEE) was obtained (SEE termination criteria of 0.20, 0.25, 0.30 were used). The initial ability estimate for an examinee was the population's mean. Both CATs used a ten point uniform prior distribution.

A data generation program based on a linear factor analytic model (Wherry, Naylor, Wherry, & Fallis, 1965) was written and is discussed below. The linear factor analytic approach for generating the data was used to minimize any bias in favor of either the 3PL or NR model; this procedure has been used previously (De Ayala, Dodd, & Koch, in press; Dodd, 1984; Koch, 1981; Reckase, 1979).

Calibration: MULTILOG (Thissen, 1988) was used to obtain item parameter estimates for the NR and 3PL models using default program parameters.

Data: Thirteen hundred examinees' abilities were generated to be evenly distributed between -3.0 and 3.0 using a one-half logit interval between successive θ levels. The examinees' responses to 150 4-alternative items were generated according to the linear factor analytic model:

$$z_{ki} = a_i\theta_{T_k} + \sqrt{1 - h_i^2}\, z_{e_{ki}} \quad (15),$$

where θ_{T_k} was examinee k's latent ability, a_i was item i's factor loading, h_i^2 was item i's communality, and z_{eki} was a random number generated from a N(0,1) distribution to be the error component of examinee k and item i. All factor loadings were uniformly high and ranged from 0.62 to 0.84. Subsequent to the calculation of z_{ki}, z_{ki} was compared to pre-specified category boundaries to determine the category response for examinee k to item i.

These data were submitted to MULTILOG to obtain item parameter estimates for both the NR and 3PL models. Given the results of Study 1, item pools for the NR and the 3PL CATs were constructed by identifying items with values of $I_{max} \geq 0.16$ and whose θ_{max} values were evenly distributed throughout the -2.0 to 2.0 ability range. These items were replicated to produce item pools of 152 items.

Analysis: The focus of Study 2 was to determine whether there were any psychometric advantages to be achieved by using the polytomous NR model as oppose to the dichotomous 3PL model. The quality of the ability estimation provided by the two CATs was analyzed by calculating RMSE and Bias. Moreover, the number of items administered (NIA) in obtaining $\hat{\theta}$ was also used for comparing the two types of CATs. The design was a one-group repeated measures design with three dependent variables: RMSE, Bias, and NIA; type of CAT (NR, 3PL) was the between subjects factor and SEE termination criterion (0.20, 0.25, 0.30) was the repeated measures or within subjects factor. Because the Bonferroni method was used to control for familywise Type I error, α was set at 0.0056. Post hoc analysis was performed with the Scheffé test using a critical F of 10.223 $(=(\upsilon_1)F_{0.0056,\ 1,\ 16})$.

Because of the item pool characteristics only examinees with $-2.0 \leq \theta_T \leq 2.0$ were used in the CATs. For each of these 900 examinees an adaptive test was simulated using the NR and 3PL CATs, the relevant item pool and SEE termination criterion. Descriptive statistics on the adaptive tests were calculated.

RESULTS

Study 1

Table 1 contains descriptive statistics on the NR adaptive tests. As would be expected, there was a direct relationship between the fidelity coefficient, $r_{\hat{\theta}\theta_T}$, and I_{max} as well as between $r_{\hat{\theta}\theta_T}$ and test length. For $I_{max} = 0.25$ there was a slight increase in $r_{\hat{\theta}\theta_T}$ as the number of categories increased for a given test length: 10, 15, 20, or 25 items; this increase in $r_{\hat{\theta}\theta_T}$ tended to diminish with increasing test length.

Insert Table 1 about here

The repeated measures analyses are presented in Table 2. As can be seen for the two category condition (Table 2a) the average RMSE improved significantly as both test length and I_{max} increased. Post hoc analysis of the I_{max} factor showed that for the two category case there was a significant reduction in RMSE as I_{max} increased from 0.04 to 0.09 to 0.16 for tests of 15-, 20-, 25- and 30-items in length. Increasing the item information content from 0.16 to 0.25 did not produce a significant improvement in ability estimation as assessed by RMSE. For the 10-item test there was, in addition to the above finding, a significant improvement in accuracy of estimation from 0.16 to 0.25. That is, for the shorter test length of 10 items more informative items were needed than at longer test lengths.

Insert Table 2 about here

For all I_{max} values there was a significant improvement in the accuracy of estimation as tests increased in length from 10 to 15 to 20 items. As would be expected, at higher item information levels (e.g., 0.16 and 0.25) increasing the length of the tests from 20 to 25 items or from 25 to 30 items did not yield a significant reduction in RMSE; for I_{max} = 0.09 estimation accuracy was significantly improved by increasing the test length from 20 to 25 items, but not from 25 to 30 items. In short, it appears that the use of items with $I_{max} \geq 0.16$ (i.e., $a \geq 0.80$) provides reasonable ability estimation for tests of 20 (possibly 15) or more items. With shorter length tests more informative items are required than at longer test lengths. Test length and I_{max} did not have a significant effect on Bias. This is, in part, a function of the way Bias is calculated and the potential for cancellation of negative Bias by positive bias. Figure 5 contains RMSE and Bias plots for selected NR CATs; these plots are typical of all the NR CAT plots.

Insert Figure 5 about here

For the three category condition (Table 2b) and test lengths of 20 or more items the results were similar to the two category condition. That is, there was a significant reduction in RMSE as I_{max} increased from 0.04 to 0.09 to 0.16, but not from 0.16 to 0.25. However, for the 10- and 15-item test lengths the results were the reverse those of the two category condition. In general, results for the four category condition (Table 2c) parallel those of the two- and three-category condition. That is, there was a significant reduction in RMSE as I_{max} increased from 0.04 to 0.09 to 0.16 to 0.25 for tests of 20 or fewer items. There was no significant reduction in RMSE as I_{max} increased from 0.16 to 0.25 for tests of 25 or 30 items.

Study 2

Table 3 contains descriptive statistics on the NR and 3PL adaptive tests. The results for the NR and 3PL CATs tended to be comparable with the only meaningful difference in $r_{\hat{\theta}\theta_T}$ appearing at a termination SEE of 0.30. However, the NR CAT tended to administer adaptive tests which, on average, were shorter than those of the 3PL CAT.

Insert Table 3 about here

Table 4 contains the source tables for the repeated measures analysis. With respect to RMSE and Bias there were no significant differences between the 3PL and NR CATs. Although the NR CAT did administer, on average, fewer items than did the 3PL CAT to achieve the same accuracy in estimation, this difference was not significant using the Bonferroni criterion. That is, the ability estimation of the NR CAT was comparable to that of the 3PL CAT.

Insert Table 4 about here

Because with a polytomous model item information is the sum of the information functions for individual responses (a.k.a., category or option information function) an exploratory study selecting items on the basis of category information was conducted (i.e., which item provided the maximum information for the particular alternative chosen by the examinee). It was believed that selecting items on the basis of category information would be more consistent with the concept of polytomous scoring of examinee responses than selecting items on the basis of item information which ignores which particular response an examinee provided. (Of course, the likelihood function is a function of an examinee's particular responses.) This exploratory study used the same simulated data and programs as Study 2, except that items were selected on the basis of category information rather than on the basis of item information. These results are provided in Table 5 and as can be seen parallel those presented in Table 4. Specifically, the NR CAT which selected items on the basis of category information provided ability estimation which, in terms of RMSE and Bias was comparable to that of the 3PL CAT. However, unlike the NR CAT results presented previously, selecting items on the basis of category information did result in the NR CAT administering significantly shorter tests, on average, than did the 3PL CAT for all SEE termination conditions. The post hoc comparison Fs for NIA were all significant at an overall $\alpha = 0.05$ and were 12.074, 16.225, and 11.357 for the SEE termination criteria of 0.20, 0.25, and 0.30, respectively. As can be from Table 6, despite

this reduction in test length the NR CAT yielded fidelity coefficients comparable to those of the 3PL CAT.

Insert Tables 5 and 6 about here

DISCUSSION

In general, the distribution of information was affected by the distance between the item's *b*s, whether the *b*s were in order, and the number of item alternatives. Study 1 showed that for two-, three-, and four-category items, items with an I_{max} value of at least 0.16 produced reasonably accurate ability estimation for test lengths of 15 or more items. Shorter length tests required more informative items to maintain reasonable ability estimation.

Results from Study 2 seemed to indicate that the NR CAT was able to produce ability estimates comparable to those of the 3PL CAT. To achieve the same level of accuracy (e.g., SEE = 0.20) the NR CAT administered fewer items, on average, than did the 3PL CAT (e.g., 12.393 versus 16.191,respectively). Although this latter result was nonsignificant, some practitioners may still consider it meaningful because in an implementation the adaptive test administered under the NR model would be shorter than it is under the 3PL model. However, a plot of the difference in average NIA between the NR and 3PL CATs versus θ showed that the NR CAT administered substantially fewer items, on average, primarily for examinees with $\theta_T \leq -1.0$ (see Figure 6). A relative efficiency comparison of the information content of the item pools of the NR and 3PL CATs showed that although the NR model provided slightly more information than did the 3PL model throughout the ability range, the NR model began to provide substantially more information than the 3PL model below $\theta = -1.0$. Past experience with dichotomous models has shown that item pools which are more informative for the ability range below -1.0 than existed in the present study can be constructed. Therefore, practitioners should not consider the NR CAT's shorter average test lengths to necessarily be meaningful. This interpretation is also appropriate for the significant NIA results when category information was used for selecting items for the NR CAT.

Insert Figure 6 about here

It appears that an NR model-based CAT can provide ability estimation comparable to a dichotomous model-based CAT. The NR CAT did not provide more accurate $\hat{\theta}$ for examinees with $\theta < 0.0$, relative to the 3PL CAT, because a variable test length was used. That is, the additional information provided by the NR model over a dichotomous model for the lower half

of the ability distribution resulted in the adaptive test terminating sooner than it would with the dichotomous model. For a given (reasonable) *fixed* length test, one would expect that the NR CAT would provide more accurate $\hat{\theta}$ for examinees with $\theta < 0.0$ than would a dichotomous model.

For those situations presented above (testlets, administration of items which do not contain a correct response, such as, demographic items, innovative computerized item formats or items which contain educational diagnostic information) it appears that the NR CAT may be an viable CAT option. Given the exploratory results, the use of category information for item selection needs to be more systematically investigated. The use of category information for item selection may prove useful in certain situations.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Brown, J.S., & Burton, R.R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*, 155-192.

Brown, J.S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science, 4*, 379-426.

Coombs, C.H., Milholland, J.E., & Womer, F.B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement, 16*, 13-37.

De Ayala, R.J., Dodd, B.G. & Koch, W. R. (1990). A computerized simulation of a flexilevel test and its comparison with a Bayesian computerized adaptive test. *Journal of Educational Measurement, 27*, 227-239.

De Ayala, R.J., Dodd, B.G. & Koch, W.R. (In Press). A comparison of the partial credit and graded response models in computerized adaptive testing. *Applied Measurement in Education.*

Dodd, B.G. (1984). *Attitude scaling: A comparison of the graded response and partial credit latent trait models.* Doctoral Dissertation, The University of Texas at Austin.

Dodd, B.G., Koch, W.R., & De Ayala, R.J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13*, 129-144.

Frary, R.B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2*, 79-96.

Haladyna, T., & Sympson, J.B. (1988, April). *Empirically based polychotomous scoring of multiple-choice items: Historical overview.* Paper presented at the annual meeting of American Educational Research Association, New Orleans.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff Publishing

Kingsbury, G.G., & Houser, R.L. (1988, April). *A comparison of achievement level estimates from computerized adaptive testing and paper-and-pencil testing.* Paper presented at the annual meeting of American Educational Research Association, New Orleans.

Koch, W.R. (1981). *Attitude scaling using latent trait theory.* Doctoral Dissertation, The University of Missouri at Columbia.

Koch, W.R., & Dodd, B.G. (1989). An investigation of procedures for computerized adaptive testing using the partial credit scoring. *Applied Measurement in Education, 2*, 335-357.

Lane, S., Stone, C.A., & Hsu, H. (1990). *Diagnosing students' errors in solving algebra word problems.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Levine, M., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43*, 675-685.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McBride, J.R., & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (ed.), *New Horizons in Testing* (pp 223-237). New York: Academic.

Nedelsky, L. (1954). Ability to avoid gross ever as a measure of achievement. *Educational and Psychological Measurement, 14*, 459-472.

Patience, W.M., & Reckase, M.D. (1980). *Effects of program parameters and item pool characteristics on the bias of a three-parameter tailored testing procedure.* Paper presented at the meeting of the National Council of Measurement in Education , Boston.

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Sympson, J.B. (1986, August). *Extracting information from wrong answers in computerized adaptive testing.* Paper presented at the American Psychological Association, Washington, D.C.

Tatsuoka K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Thissen, D.J. (1976). Information in wrong responses to Raven's Progressive Matrices. *Journal of Educational Measurement, 13*, 201-214.

Thissen, D.J. (1988). *MULTILOG-User's Guide* (Version 5.1). Scientific Software, Inc. Mooresville, IN.

Thissen, D.J., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26*, 247-260.

Urry, V.W. (1977). Tailored testing: a successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

Vale, C.D., & Weiss, D.J. (1977). *A comparison of information functions of multiple-choice and free-response vocabulary items.* (Research Report 77-2). Minneapolis, MN: University of Minnesota.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale: Lawrence Erlbaum Associates.

Wainer, H. & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wang, M.W., & Stanley, J.C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40*, 663-706.

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

Weiss, D.J. (1983). *New Horizons in Testing* . New York: Academic.

Wherry, R.J., Sr., Naylor, J.C., Wherry, R.J., Jr., & Fallis, R.F. (1965). Generating multiple samples of multivariate data with arbitrary population parameters. *Psychometrika, 30*, 303-314.

Table 1: Mean $\hat{\theta}$, standard deviation of $\hat{\theta}$ (SD), and $r_{\hat{\theta}\theta_T}$[a].

m	I_{max}		Test Length 10	15	20	25	30
2	0.25	mean	0.021	0.010	0.002	-0.002	-0.003
		SD	1.936	1.921	1.906	1.898	1.900
		r	0.935	0.956	0.967	0.973	0.977
	0.16	mean	0.027	0.007	0.001	-0.009	-0.009
		SD	1.949	1.927	1.923	1.918	1.914
		r	0.910	0.938	0.954	0.962	0.968
	0.09	mean	0.052	0.006	-0.003	-0.140	-0.003
		SD	1.952	1.948	1.948	1.950	1.937
		r	0.863	0.905	0.926	0.939	0.949
	0.04	mean	0.068	0.061	0.020	0.014	0.009
		SD	1.875	1.908	1.932	1.945	1.956
		r	0.759	0.818	0.855	0.880	0.900
3	0.25	mean	-0.003	0.003	0.000	0.004	0.001
		SD	1.951	1.936	1.929	1.924	3.670
		r	0.936	0.958	0.968	0.974	0.978
	0.16	mean	-0.003	-0.014	-0.004	0.010	0.009
		SD	1.959	1.951	1.952	1.942	1.938
		r	0.918	0.939	0.956	0.964	0.971
	0.09	mean	-0.004	-0.006	-0.009	-0.008	0.000
		SD	1.965	1.963	1.958	1.951	1.950
		r	0.863	0.903	0.929	0.941	0.950
	0.04	mean	-0.0200	0.000	0.009	0.015	0.003
		SD	1.881	1.922	1.939	1.950	1.954
		r	0.763	0.831	0.868	0.890	0.907
4	0.25	mean	-0.007	-0.008	-0.013	-0.014	-0.016
		SD	1.969	1.951	1.941	1.943	1.934
		r	0.942	0.960	0.969	0.974	0.977
	0.16	mean	-0.034	-0.025	-0.028	-0.031	-0.035
		SD	1.979	1.974	1.973	1.960	1.961
		r	0.912	0.939	0.951	0.959	0.964
	0.09	mean	-0.015	-0.007	-0.006	-0.016	-0.017
		SD	1.978	1.979	1.975	1.986	1.985
		r	0.855	0.902	0.925	0.938	0.945
	0.04	mean	-0.034	-0.008	-0.001	-0.002	-0.009
		SD	1.902	1.941	1.963	1.976	1.982
		r	0.752	0.816	0.847	0.876	0.892

[a]Pearson product-moment correlation coefficients between θ_T and $\hat{\theta}$.

Mean $\theta_T = 0.000$ and $s_{\theta T} = 1.872$.

Table 2a: Accuracy analysis for for NR CAT: two category condition.

RMSE

Source	SS	df	MS	F	p
Between Subjects					
I_{max}	10.591	3	3.530	167.500*	0.000
Subjects w/i Groups	1.012	48	0.021		
Within Subjects					
Test Length	4.237	4	1.059	553.451*	0.000
I_{max} X Test Length	0.121	12	0.010	5.288	0.000
Test Length X Subjects w/i Groups	0.367	192	0.002		

Post Hoc Comparison Fs for I_{max}:

Comparison	Test Length				
	10	15	20	25	30
$\mu_{0.25}$ vs $\mu_{0.16}$	17.749*	13.005	8.998	7.270	6.561
$\mu_{0.16}$ vs $\mu_{0.09}$	41.007*	30.926*	28.357*	24.213*	20.397*
$\mu_{0.09}$ vs $\mu_{0.04}$	101.553*	104.288*	92.265*	79.169*	66.524*

Post Hoc Comparison Fs for test length:

Comparison	I_{max}			
	0.04	0.09	0.16	0.25
μ_{30} vs μ_{25}	19.269*	9.831	6.009	4.943
μ_{25} vs μ_{20}	24.599*	14.157*	9.477	6.581
μ_{20} vs μ_{15}	45.253*	32.500*	28.109*	18.281*
μ_{15} vs μ_{10}	85.293*	89.557*	64.613*	49.169*

Bias

Source	SS	df	MS	F	p
Between Subjects					
I_{max}	0.041	3	0.014	0.110	0.954
Subjects w/i Groups	5.964	48	0.124		
Within Subjects					
Test Length	0.074	4	0.018	2.237	0.067
I_{max} X Test Length	0.017	12	0.001	0.176	0.999
Test Length X Subjects w/i Groups	1.580	192	0.008		

*significant at overall α = 0.05, critical F = 13.260 (α = 0.008 per test).

Table 2a: Accuracy analysis for for NR CAT: two category condition (continued).

Average RMSE: (2 categories)

I_{max}	Test Length				
	10	15	20	25	30
0.04	1.298	1.136	1.018	0.931	0.854
0.09	0.999	0.833	0.733	0.667	0.612
0.16	0.809	0.668	0.575	0.521	0.478
0.25	0.684	0.561	0.486	0.441	0.402

Table 2b: Accuracy analysis for for NR CAT: three category condition.

RMSE

Source	SS	df	MS	F	p
Between Subjects					
I_{max}	9.492	3	3.164	135.396*	0.000
Subjects w/i Groups	1.122	48	0.023		
Within Subjects					
Test Length	4.326	4	1.081	495.580*	0.000
I_{max} X Test Length	0.168	12	0.014	6.407	0.000
Test Length X Subjects w/i Groups	0.419	192	0.002		

Post Hoc Comparison Fs for I_{max}:

Comparison	Test Length				
	10	15	20	25	30
$\mu_{0.25}$ vs $\mu_{0.16}$	7.487	13.622*	7.400	5.694	3.891
$\mu_{0.16}$ vs $\mu_{0.09}$	52.625*	29.602*	24.008*	21.579*	21.284*
$\mu_{0.09}$ vs $\mu_{0.04}$	82.226*	64.287*	62.766*	54.019*	46.795*

Post Hoc Comparison Fs for test length:

Comparison	I_{max}			
	0.04	0.09	0.16	0.25
μ_{30} vs μ_{25}	14.224*	8.163	7.840	4.232
μ_{25} vs μ_{20}	22.495*	13.796*	10.609	6.322
μ_{20} vs μ_{15}	48.601*	46.240*	33.972*	17.881*
μ_{15} vs μ_{10}	119.122*	81.515*	33.309*	56.036*

Bias

Source	SS	df	MS	F	p
Between Subjects					
I_{max}	0.002	3	0.001	0.007	0.999
Subjects w/i Groups	5.114	48	0.107		
Within Subjects					
Test Length	0.005	4	0.0013	0.157	0.960
I_{max} X Test Length	0.010	12	0.0008	0.091	1.000
Test Length X Subjects w/i Groups	1.677	192	0.0087		

*significant at overall $\alpha = 0.05$, critical $F = 13.260$ ($\alpha = 0.008$ per test).

Table 2b: Accuracy analysis for for NR CAT: three category condition (continued).

Average RMSE: (3 categories)

I_{max}	Test Length				
	10	15	20	25	30
0.04	1.286	1.095	0.973	0.890	0.824
0.09	1.001	0.843	0.724	0.659	0.609
0.16	0.773	0.672	0.570	0.513	0.464
0.25	0.687	0.556	0.482	0.438	0.402

Table 2c: Accuracy analysis for for NR CAT: four category condition.

RMSE

Source	SS	df	MS	F	p
Between Subjects					
I_{max}	11.713	3	3.904	135.736*	0.000
Subjects w/i Groups	1.381	48	0.029		
Within Subjects					
Test Length	3.731	4	0.933	556.861*	0.000
I_{max} X Test Length	0.177	12	0.015	8.826	0.000
Test Length X Subjects w/i Groups	0.322	192	0.002		

Post Hoc Comparison Fs for I_{max}:

Comparison	Test Length				
	10	15	20	25	30
$\mu_{0.25}$ vs $\mu_{0.16}$	20.337*	15.961*	15.008*	11.905	10.488
$\mu_{0.16}$ vs $\mu_{0.09}$	45.553*	29.024*	18.212*	15.720*	15.481*
$\mu_{0.09}$ vs $\mu_{0.04}$	75.979*	79.718*	86.898*	67.274*	54.091*

Post Hoc Comparison Fs for test length:

Comparison	I_{max}			
	0.04	0.09	0.16	0.25
μ_{30} vs μ_{25}	13.375*	4.232	4.000	2.560
μ_{25} vs μ_{20}	33.309*	13.375*	9.522	5.224
μ_{20} vs μ_{15}	28.242*	36.689*	15.546*	13.796*
μ_{15} vs μ_{10}	94.357*	102.299*	56.895*	43.184*

Bias

Source	SS	df	MS	F	p
Between Subjects					
I_{max}	0.018	3	0.006	0.059	0.981
Subjects w/i Groups	4.888	48	0.102		
Within Subjects					
Test Length	0.004	4	0.0010	0.134	0.970
I_{max} X Test Length	0.008	12	0.0007	0.078	1.000
Test Length X Subjects w/i Groups	1.599	192	0.0083		

*significant at overall $\alpha = 0.05$, critical $F = 13.260$ ($\alpha = 0.008$ per test).

Table 2c: Accuracy analysis for for NR CAT: four category condition (continued).

Average RMSE (4 categories):

I_{max}	Test Length				
	10	15	20	25	30
0.04	1.321	1.151	1.058	0.957	0.893
0.09	1.033	0.856	0.750	0.686	0.650
0.16	0.810	0.678	0.609	0.555	0.520
0.25	0.661	0.546	0.481	0.441	0.413

Table 3: Descriptive statistics for NR and 3PL CATs. Item selection on the basis of item information for both NR and 3PL CATs.

CAT	SEE	Mean $\hat{\theta}$	SD $\hat{\theta}$	Mean NIA[a]	Median NIA[a]	SD NIA[a]	r[b]
3PL	0.30	0.168	1.193	12.759	10.000	5.927	0.902
	0.25	0.152	1.165	15.073	13.000	6.335	0.925
	0.20	0.171	1.164	16.191	13.000	6.879	0.928
NR	0.30	0.275	1.200	9.682	8.000	5.871	0.926
	0.25	0.267	1.190	10.763	9.000	6.472	0.926
	0.20	0.269	1.186	12.393	10.000	6.532	0.929

[a]Number of items administered

[b]Spearman rank-order correlation coefficients between $\hat{\theta}$ and θ_T.

Note: $\bar{\theta}_T = 0.000$, $s_{\theta T} = 1.292$.

Table 4: Accuracy analysis for NR and 3PL CATs. Item selection on the basis of item information for both NR and 3PL CATs.

RMSE

Source	SS	df	MS	F	p
Between Subjects					
CAT Type	0.054	1	0.054	0.681	0.421
Subjects w/i Groups	1.267	16	0.079		
Within Subjects					
SEE Term	0.022	2	0.011	12.584*	0.000
CAT Type X SEE Term	0.004	2	0.002	2.527	0.096
SEE Term X Subjects w/i Groups	0.028	32	0.001		

Bias

Source	SS	df	MS	F	p
Between Subjects					
CAT Type	0.154	1	0.154	0.661	0.428
Subjects w/i Groups	3.736	16	0.234		
Within Subjects					
SEE Term	0.001	2	0.0005	1.492	0.240
CAT Type X SEE Term	0.001	2	0.0005	0.763	0.475
SEE Term X Subjects w/i Groups	0.014	32	0.0004		

NIA

Source	SS	df	MS	F	p
Between Subjects					
CAT Type	187.638	1	187.638	8.068	0.012
Subjects w/i Groups	372.095	16	23.256		
Within Subjects					
SEE Term	85.231	2	42.615	76.371*	0.000
CAT Type X SEE Term	3.455	2	1.728	3.096	0.059
SEE Term X Subjects w/i Groups	17.856	32	0.558		

*significant at overall $\alpha = 0.05$, critical $F = 10.223$ ($\alpha = 0.0056$ per test).

Table 5: Accuracy analysis for NR and 3PL CATs. Item selection on the basis of category information for NR CAT and via item information for 3PL CAT.

RMSE

Source	SS	df	MS	F	p
Between Subjects					
CAT Type	0.018	1	0.018	0.203	0.658
Subjects w/i Groups	1.450	16	0.091		
Within Subjects					
SEE Term	0.035	2	0.017	9.023	0.001
CAT Type X SEE Term	0.008	2	0.004	2.026	0.148
SEE Term X Subjects w/i Groups	0.062	32	0.002		

Bias

Source	SS	df	MS	F	p
Between Subjects					
CAT Type	0.196	1	0.196	0.767	0.394
Subjects w/i Groups	4.085	16	0.255		
Within Subjects					
SEE Term	0.004	2	0.002	1.206	0.313
CAT Type X SEE Term	0.007	2	0.004	2.416	0.105
SEE Term X Subjects w/i Groups	0.048	32	0.001		

NIA

Source	SS	df	MS	F	p
Between Subjects					
CAT Type	335.653	1	335.653	13.883*	0.002
Subjects w/i Groups	386.833	16	24.177		
Within Subjects					
SEE Term	102.072	2	51.036	74.531*	0.000
CAT Type X SEE Term	2.123	2	1.062	1.550	0.228
SEE Term X Subjects w/i Groups	21.912	32	0.685		

Table 6: Descriptive statistics for NR and 3PL CATs. Item selection on the basis of category information for the NR CAT and item information for the 3PL CAT.

CAT	SEE	Mean $\hat{\theta}$	SD $\hat{\theta}$	Mean NIA[a]	Median NIA[a]	SD NIA[a]	r[b]
3PL	0.30	0.168	1.193	12.759	10.000	5.927	0.902
	0.25	0.152	1.165	15.073	13.000	6.335	0.925
	0.20	0.171	1.164	16.191	13.000	6.879	0.928
NR	0.30	0.302	1.157	8.121	6.000	4.956	0.916
	0.25	0.292	1.170	9.532	8.000	5.116	0.918
	0.20	0.259	1.180	11.411	10.000	6.195	0.924

[a]Number of items administered

[b]Spearman rank-order correlation coefficients between $\hat{\theta}$ and θ_T.

Note: $\bar{\theta}_T = 0.000$, $s_{\theta_T} = 1.292$.

Figure Captions

Figure 1. Multivariate logit plot for a three category item, $\mathbf{a}$ = (-0.75, -0.25, 1.0) and $\mathbf{c}$ = (-1.5, -0.25, 1.75), in the category selected and logit spaces.

Figure 2. Example OCCs for a three category item, $\mathbf{a}$ = (-0.75, -0.25, 1.0) and $\mathbf{c}$ = (-1.5, -0.25, 1.75).

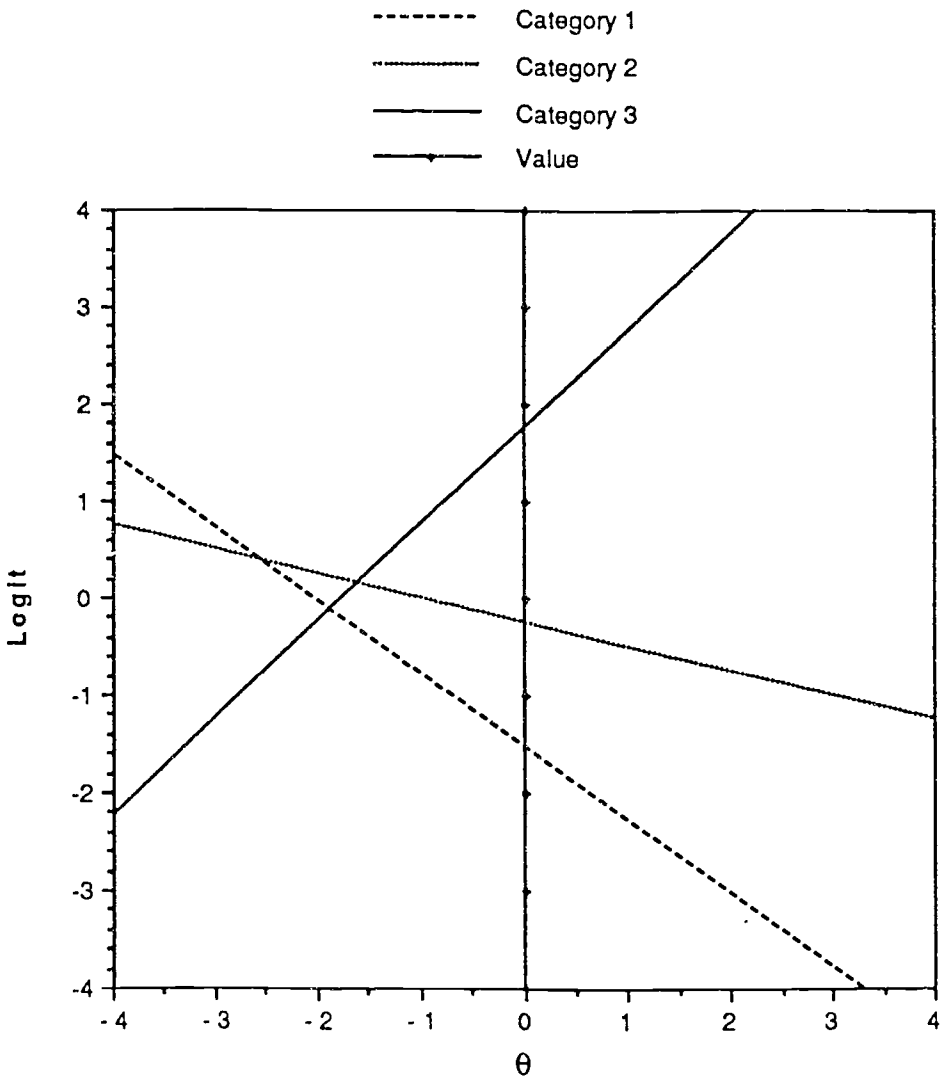
Figure 3. NR model's OCCs ($a_2 = 0.40$, $a_1 = -0.40$, $c_2 = 0.2$, and $c_1 = -0.20$) and the 2PL ICC ($a = 0.80$ and $b = -0.5$).

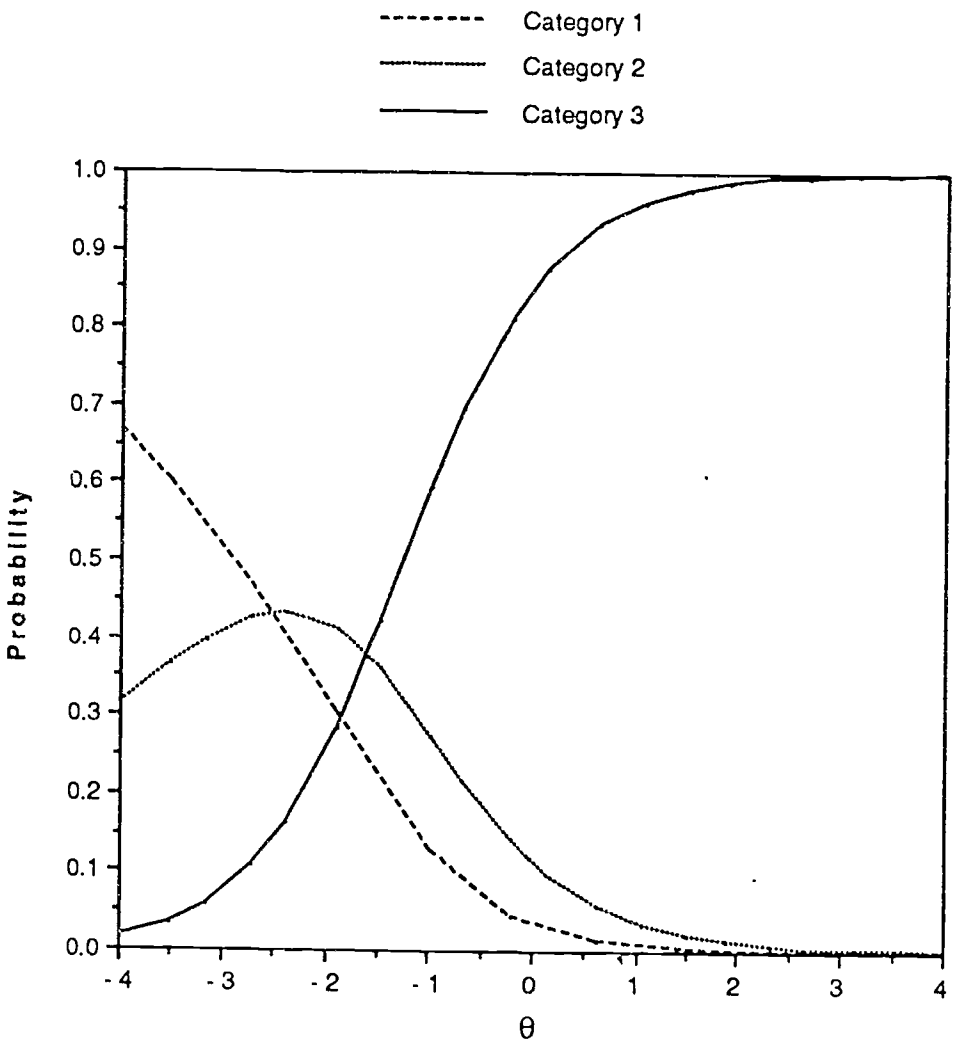
Figure 4. Bimodal information function for an item where $\mathbf{a}$ = (1, 0.1, -0.1, -1) and $\mathbf{c}$ = (0.1, 2.4, -2.6, 0.1)

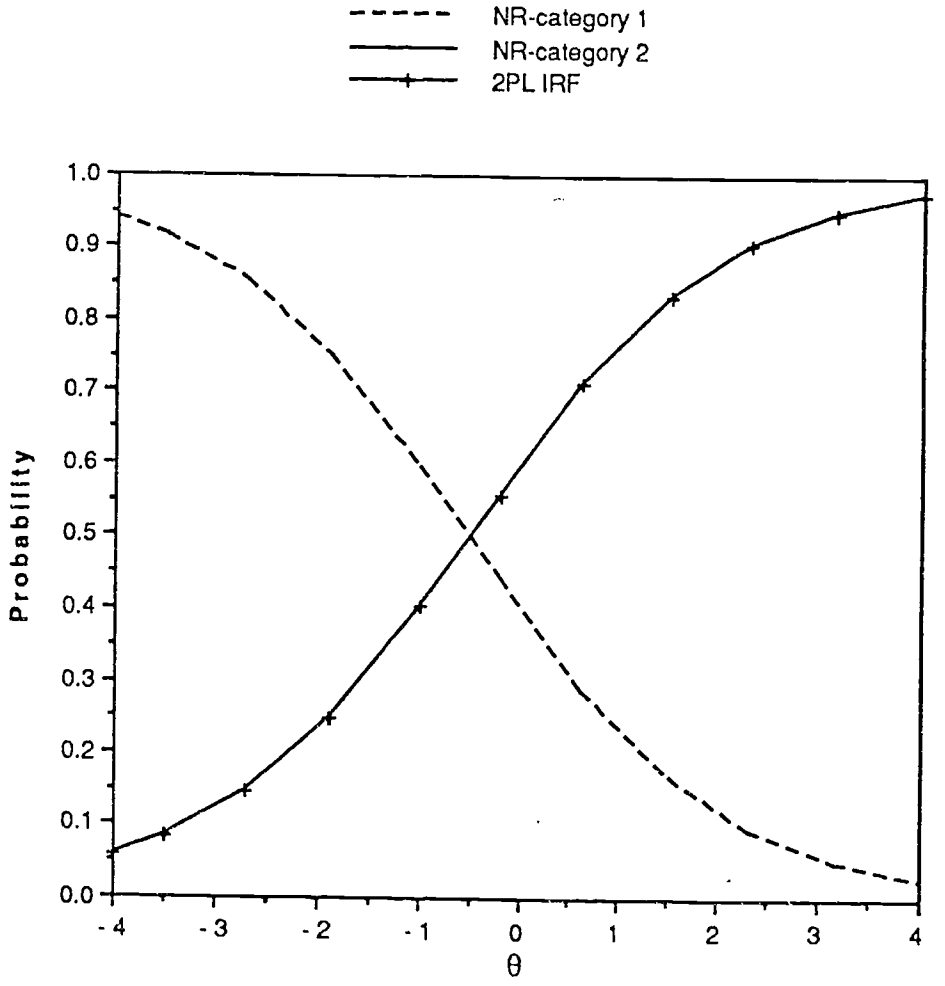
Figure 5a. RMSE plot for NR CAT ($m_i = 3$, NIA = 20).

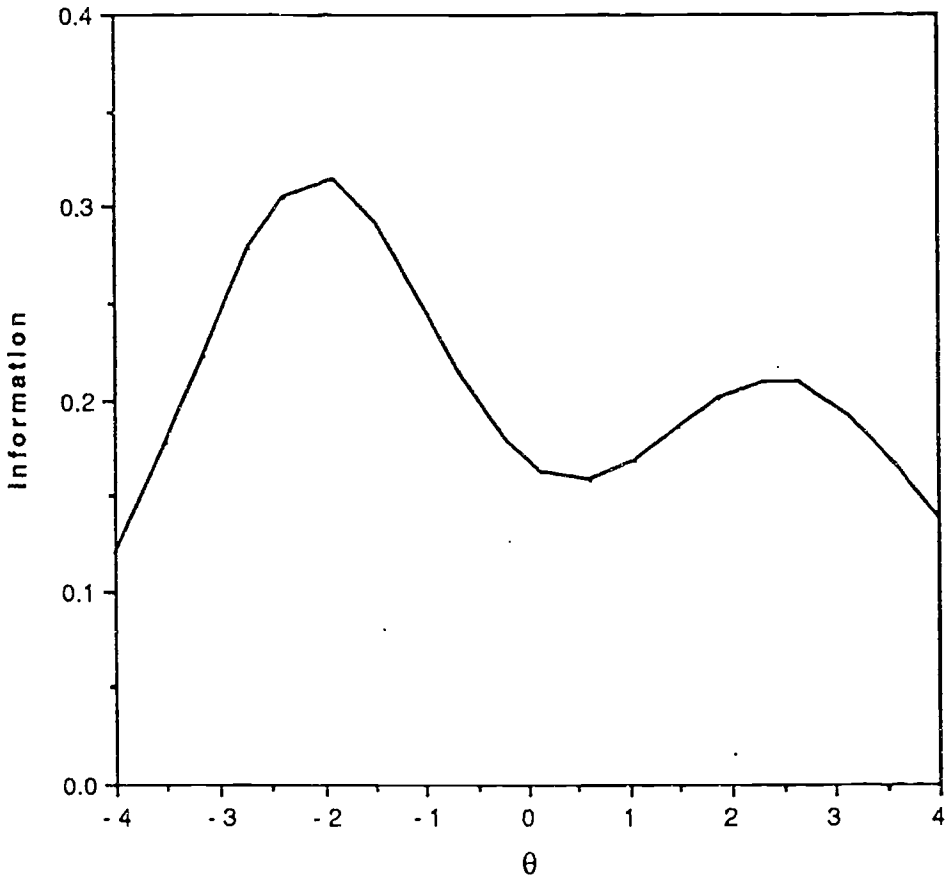
Figure 5b. Bias plot for NR CAT ($m_i = 3$, NIA = 20).

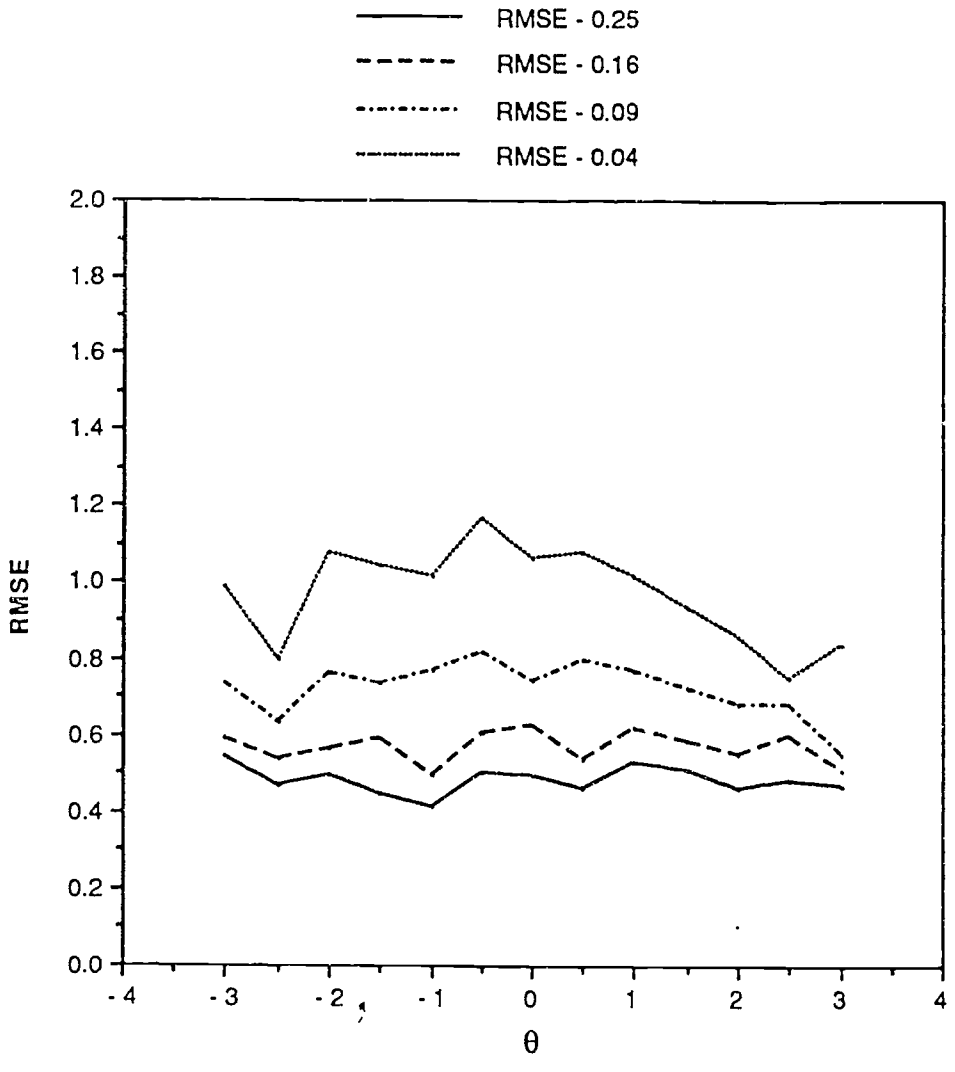
Figure 6. Average NIA for NR CAT minus average NIA for 3PL CAT.

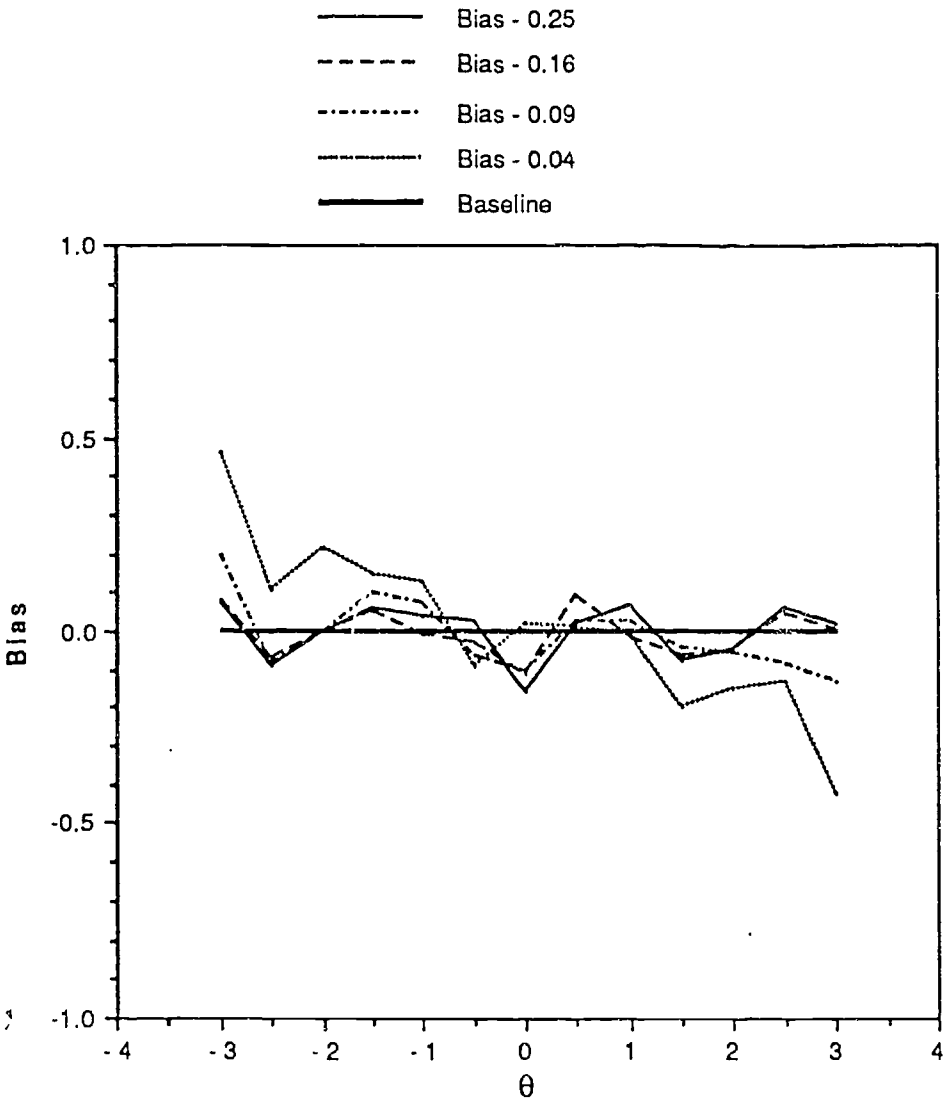

Category 1
Category 2
Category 3
Value
Logit
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4
θ

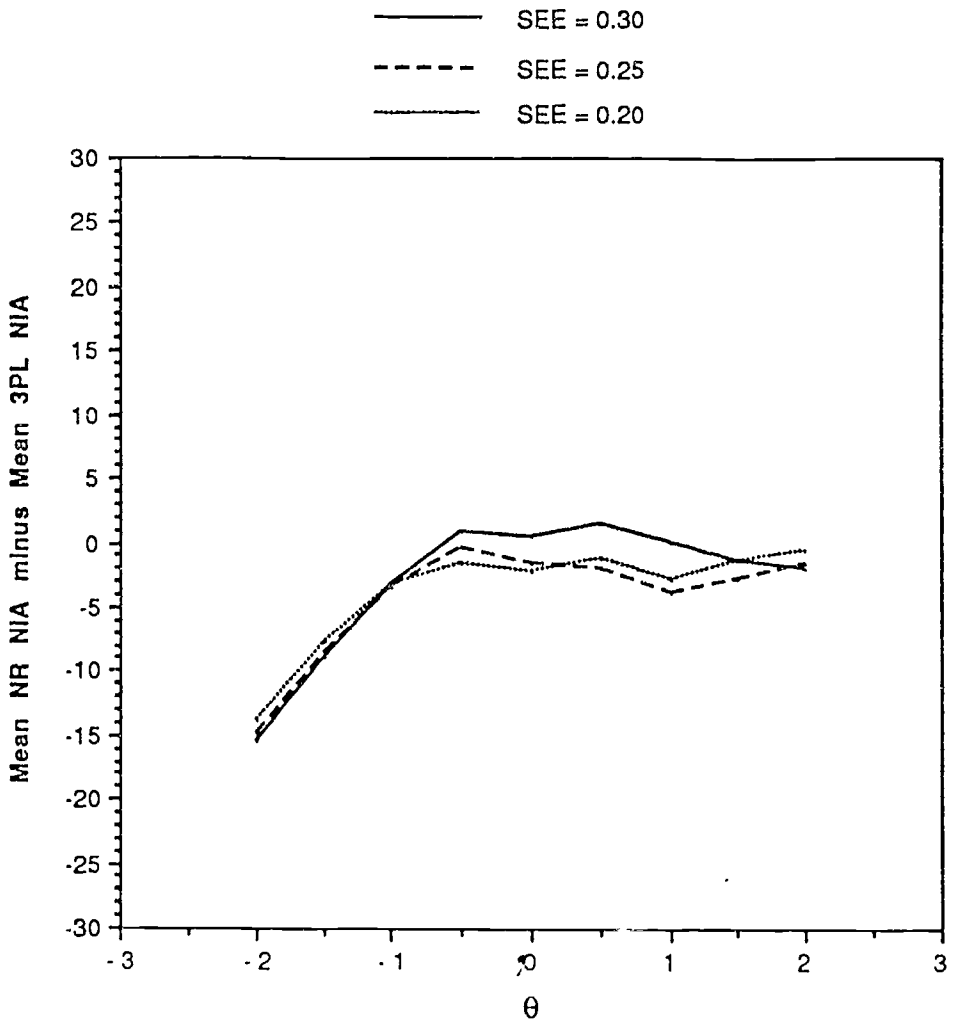

Category 1
Category 2
Category 3
Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4
-3
-2
-1
0
1
2
3
4
θ


NR-category 1
NR-category 2
2PL IRF
Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-4
-3
-2
-1
0
1
2
3
4
θ


0.4
0.3
0.2
0.1
0.0
Information
-4
-3
-2
-1
0
1
2
3
4
θ


RMSE - 0.25
RMSE - 0.16
RMSE - 0.09
RMSE - 0.04
RMSE
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-4
-3
-2
-1
0
1
2
3
4
θ


Bias - 0.25
Bias - 0.16
Bias - 0.09
Bias - 0.04
Baseline
1.0
0.5
0.0
-0.5
-1.0
Bias
-4
-3
-2
-1
0
1
2
3
4
θ


SEE = 0.30
SEE = 0.25
SEE = 0.20
Mean NR NIA minus Mean 3PL NIA
30
25
20
15
10
5
0
-5
-10
-15
-20
-25
-30
-3
-2
-1
0
1
2
3
θ

Acknowledgements

The author would like to thank Dr. William D. Schafer and the editor for constructive and helpful comments.