DOCUMENT RESUME

ED 295 950 TM 011 421

AUTHOR De Ayala, R. J.; And Others
TITLE A Comparison of the Nominal and Graded Response Models in Computerized Testing.
PUB DATE Apr 88
NOTE 33p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability Identification; *Achievement Tests; Adaptive Testing; *Aptitude Tests; Comparative Analysis; Computer Assisted Instruction; *Computer Assisted Testing; *Diagnostic Tests; *Latent Trait Theory; Postsecondary Education; Test Validity
IDENTIFIERS *Graded Response Model; *Nominal Response Model

ABSTRACT

To date, the majority of computerized adaptive testing (CAT) systems for achievement and aptitude testing have been based on the dichotomous item response models. However, current research with polychotomous model-based CATs is yielding promising results. This study extends previous work on nominal response model-based CAT (NR CAT) and compares its ability estimation as well as its overall performance to graded response model-based CAT (GR CAT). A data set of 275 examinees was used, derived from five administrations of the College Board's Achievement Test in Mathematics, Level I, at the University of Texas, Austin. Results show that both CATs had high convergence rates despite using a small item pool and had average test lengths slightly below 16 items. The NR CAT's ability estimates were highly correlated with and not significantly different from an external criterion and showed no systematic bias in estimating ability throughout the trait continuum. In contrast, the GR CAT had a tendency to underestimate high ability examinees, although its ability estimates were highly associated with the external criterion. Educational implications of the findings include the possibility of merging computer-aided instruction and diagnostic testing with CAT. Six tables and nine graphs are included. (Author/TJH)

ABSTRACT

A comparison of the nominal and graded response models in computerized testing

R.J. De Ayala
University of Maryland
Barbara Dodd & William R. Koch
University of Texas

To date, the majority of computerized adaptive testing (CAT) systems for achievement and aptitude testing have been based on the dichotomous item response models. However, current research with polychotomous model-based CATs is yielding promising results. This study extends previous work on a nominal response model-based CAT (NR CAT) and compares its ability estimation as well as its overall performance to a graded response model-based CAT (GR CAT). Results showed that both CATs had high convergence rates despite using a small item pool and had average test lengths slightly below 16 items. The NR CAT's ability estimates were highly correlated with and not significantly different from an external criterion and showed no systematic bias in estimating ability throughout the trait continuum. In contrast, the GR CAT had a tendency to underestimate high ability examinees, although its ability estimates were highly associated with the external criterion. Implications of NR and GR model-based CATs are discussed.

Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 1988.

A comparison of the nominal and graded response models in computerized testing

R.J. De Ayala
University of Maryland
Barbara Dodd & William R. Koch
University of Texas

Research on the use of polychotomous models in computerized adaptive testing (CAT) has shown promising results (e.g., Koch & Dodd, 1985; De Ayala & Koch, 1987; Sympson, 1986). There are a number of benefits which may be accrued through the use of polychotomous models in CAT. For instance, a new domain of test items which do not require transforming the examinee's actual responses into dichotomous responses (e.g., correct and incorrect) may be administered in an adaptive fashion; the term scoring will refer to this transformation. These items may be either attitude items or test items specifically developed for administration by a computer (i.e., "computerized" items; currently "paper-and-pencil" items are used in CAT).

Two polychotomous models appropriate for attitude as well as aptitude and achievement testing are Samejima's (1969) graded response (GR) model and Bock's (1972) nominal response (NR) model. Both models share an indirect relationship. Specifically, the GR model is a direct extension of the two-parameter model, whereas the NR model reduces to the two-parameter model when an item only has two categories (i.e., correct and incorrect).

Samejima (1969) extended the dichotomously scored two-parameter model to the case of polychotomously scored items with ordered responses. For the GR model the examinee responses to item i are categorized into $m_i + 1$ categories; where higher categories

indicate more of θ. Associated with each category of item $\underline{i}$ is a category score, x_i, with values $0..m_i$. The GR model may be expressed as :

$$P_x(\theta) = \frac{e^{Da_i(\theta-b_x)}}{1 + e^{Da_i(\theta-b_x)}} \qquad (1)$$

where D is a scaling constant, θ is the latent trait, a_i is the discrimination parameter for item $\underline{i}$, $\underline{b}_x$ is the difficulty parameter for category score x for item $\underline{i}$. P_x is the probability, p_x , of the examinee responding in category score x_i or higher for a given item; the p_x of responding in the lowest category (i.e., $P_0(\theta)$) or higher is 1.0. For instance, for an item with four response categories $P_2(\theta)$ is the probability of responding in categories 2 or 3 rather than in categories 0 or 1. Because P_x is the probability of responding in x_i or higher the p_x of responding in a particular category equals the difference between P_x s for adjacent categories.

For the GR model item responses must be ordered a priori in order to fit this model. Therefore, Likert scale items or aptitude/achievement test items whose alternatives have been ordered according to knowledge of the correct answer are appropriate.

Samejima (1977) showed that using the GR model in a CAT resulted in the administration of approximately 25% fewer items than a CAT using dichotomous model. Further, Samejima found that the GR CAT "presenting" items with lower $\underline{a}$s was as efficient as the

binary CAT which administered items with larger $\underline{a}$s (e.g., $\underline{a}$ values of 1.0 and 2.0, respectively).

In contrast to the GR model, Bock's NR model assumes that item alternatives represent responses measured at a nominal level of measurement. Bock's NR model provides a description of the probability of a response to each category as a function of two parameters characteristic of the particular category as well as an ability parameter. These category probabilities for an item are conditional on θ and are constrained to sum to 1.0. The NR model is expressed as :

$$p(u_i=x|\theta;\underline{a};\underline{c}) = \frac{e^{a_x(\theta) + c_x}}{\sum_{k=1,m_i} e^{a_{ik}(\theta) + c_{ik}}} \qquad (2)$$

where $p(\theta)$ is the probability that a subject with ability level θ will provide an item response, $u_i=x$, to item $\underline{i}$ with m_i categories. The item parameters, $\underline{a}_{ik}$ and $\underline{c}_{ik}$, are associated with the k^{th} category of item $\underline{i}$. Specifically, $\underline{a}_{ik}$ is the slope parameter and $\underline{c}_{ik}$ is the intercept parameter of the nonlinear response function associated with the k^{th} category of an $\underline{m}$ category item.

Bock (1972), Thissen (1976), and De Ayala and Koch (1987) have all shown that the NR model provides more information than a dichotomous model, particularly in the lower half of the ability distribution. This information can be used by a NR model-based CAT to provide more precise $\hat{\theta}$s than a three-parameter logistic (3PL) model-based CAT in the lower half of ability range (De Ayala, Dodd &

Koch, in press). Furthermore, the NR CAT was found to have a higher convergence rate than the 3PL CAT while administering approximately the same number of items. Both the 3PL and NR CATs' ability estimates were highly correlated with and not significantly different from an external criterion.

Both the GR and NR models have demonstrated the ability to utilize the information in examinees' incorrect responses for ability estimation. However, the models differ with respect to their implicit assumptions about the level of measurement inherent in the item's alternatives and in the number of parameters required to describe an item. The NR model requires more parameters to describe the examinee-item interaction than does the GR model, but does not have the GR requirement that the item's alternatives be ordered. Therefore, there is a trade-off of advantages and disadvantages between the two models. This study investigates which model may be preferable for CAT.

METHOD

Data : A data set of 1093 examinees was created from five administrations of The College Board's[1] Achievement Test in Mathematics, Level I, at the University of Texas at Austin. This data set contained only individuals who answered at least 80% of the 50 item test and the last question. Each CAT program simulated an adaptive test for each examinee in the data base. For both CATs the unscored responses were used.

To improve item parameter estimation and to work within the constraints of the calibration program, MULTILOG 5 (Thissen, 1986),

[1]The authors wish to thank The College Board for granting permission to use the Mathematics Level I Achievement test data.

the 5-choice items of the Math Level I test were collapsed into 4-choice items (a.k.a, the Collapsed data set). For the GR model's item calibration it was necessary to order the item alternatives. The ordinal relationship among the alternatives for an item was obtained by arranging the options according to the mean number correct score of examinees selecting each option. That is, the alternative with the largest mean number correct score was considered to reflect more of the trait (i.e., the optimal response) than the alternative with the second largest mean number correct score, etc. In general, the optimal response for an item was also its correct response. MULTILOG 5 was used for the NR and GR models' item calibration of the Collapsed data.

Because the CAT simulations required complete response strings, a new data set containing only those Collapsed data set examinees with no non-responses was created. Of 1093 examinees used for item calibration, 275 examinees were found to have answered all items; this data set is called the Complete data set. An ordered alternative version of Complete data was used with the GR CAT.

Programs : A CAT program based on the GR model (called GR CAT) and another program based on the NR model (called the NR CAT) were written. Both programs used maximum likelihood estimation (MLE) of ability and maximum item information for item selection.

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 attained. A previous study (Koch, De Ayala, & Dodd, 1988) demonstrated that

the use of maximum SEE as a termination criterion yielded results which were preferable to those using minimum item information as a termination criterion. The maximum SEEs were determined empirically from simulation results using small samples. Specifically, the SEE value which minimized the average number of items administered while simultaneously resulting in a large linear association between the ability estimates and an external criterion was selected for use in this study. A value of 0.45 was used as the maximum SEE value for the NR CAT and the GR CAT's maximum SEE was set to 0.44. Further, for both CATs the initial ability estimate for an examinee was assumed to be equal to the population mean (i.e., $\theta = 0.0$).

In the GR CAT a variable stepsize was added (after a response of 3 or more) or subtracted (after a response less than 3) from the previous $\hat{\theta}$ in order to provide a new $\hat{\theta}$ when MLE could not be performed (Koch, De Ayala, & Dodd, 1988). This variable stepsize procedure was used until MLE could be performed.

Because there are no correct or incorrect responses with the NR CAT, the above technique was modified based on the fact that the probability of responding in a given category varies as a function of θ. Specifically, until MLE could be performed directional information for the addition or subtraction of the fixed stepsize was obtained by : (a) calculating the probability of responding in the category the examinee chose for the range ± 0.5 θ units about the previous $\hat{\theta}$; and (b) the sign of the $\theta - \hat{\theta}$ associated with the largest probability in this θ range. If the sign of θ from (b) was positive then a stepsize of 0.20 was added

to the previous $\hat{\theta}$, otherwise the stepsize was subtracted from the previous $\hat{\theta}$. This stepsize came from analysis of small sample simulation runs with the NR CAT presented above.

The CAT simulations were analyzed with respect to their nonconvergent and convergent cases as well as those cases which were convergent for both CATs (referred to as jointly convergent cases). In addition to descriptive statistics on $\hat{\theta}$s, SEE, and number of items administered (NIA), the appropriate test characteristic curve for the item pool was used to convert the $\hat{\theta}$s from each program to a number correct score (NC estimate). This NC estimate was compared with the actual number correct score (NC empirical) the examinee received on the exam. It was felt that NC empirical was an unbiased criterion for evaluating performance of the two different models. Unless otherwise stated all correlation coefficients are Pearson Product Moment correlation coefficients.

RESULTS

Tables 1 and 2 present descriptive statistics and factor analyses of the two data sets, Collapsed and Complete. For each data set, Table 1 shows the mean NC empirical and its standard deviation (S.D.) as well as coefficient alpha (alpha). In order to get an indication of the dimensionality of the data a linear factor analysis with phi coefficients was performed. As can be seen from these tables, the elimination of examinees with incomplete response strings from the Collapsed data set did not meaningfully distort the examinee information in the Complete data set. Further, although neither data set had a first factor which accounted for a large percentage of the total variance, each data set's first

factor did account for a large proportion of the common variance. It was concluded that the data sets did not seriously violate the unidimensionality assumption.

Insert Table 1 about here

Insert Table 2 about here

Table 3 presents the results of the GR and NR models' calibration of the Collapsed data set. As can be seen from Table 3, the $\hat{\theta}$s from each model were, as would be expected, highly correlated with one another ($r = 0.99$; $r_{Spearman} = 0.99$) and 98% of the proportion of variance in one model's $\hat{\theta}$s was accounted for by the other model's $\hat{\theta}$s. Figure 1 presents a scatterplot of the calibration $\hat{\theta}$s from each model depicting the strong linear relationship between the two variables and the agreement between the models on their $\hat{\theta}$s for each examinee.

Insert Table 3 about here

Insert Figure 1 about here

Because the quality of the items' parameter estimates (at least with respect to θ) and an item's usefulness can be better summarized by information than by summary statistics for each item parameter, the items' information (Samejima, 1979) for each model's parameter estimates was calculated. These item information functions are summarized in the test information function (Birnbaum, 1968). Each model's respective test information function is pre-

sented in Figure 2. As can be seen from this figure, the NR and GR models' information functions are essentially identical and both models' functions peak approximately at $\theta = -2.0$. The GR model provides slightly more information below $\theta = -2.25$ than does the NR model, whereas the NR model provides more information than the GR model in the range -2.25 to 1.0.

Insert Figure 2 about here

The results of the GR CAT simulation's convergent cases are presented in Table 4. As can be seen, the GR CAT converged on all 275 cases. The average test length was 15.8 items (median = 16); no simulated test reached the termination criterion of 30 items. The average CAT NC estimate and the average NC empirical differed by 2.8 (25.0 and 27.8, respectively) and by 2.4 with respect to their median values (NC estimate median = 24.6, NC empirical median = 27.0). With the large sample size it was not unexpected that the matched t-test showed that the mean difference between NC empirical and NC estimate was statistically significant ($t = -10.48$, $df = 274$, $p = 0.0001$); this significant finding resulted primarily from the GR CAT's difficulty in estimating high ability examinees. In fact, the NC estimates were not significantly different from NC empiricals for NC empiricals below or equal to 25 ($t = -1.83$, $df = 111$, $p = 0.07$).

Insert Table 4 about here

A scattergram of NC estimate and NC empirical (Figure 3) shows the GR CAT had a tendency to underestimate high ability

examinees, although the correlation coefficient between NC estimate and NC empirical was 0.85. The coefficient of determination for predicting NC empirical from NC estimate was 0.73; 73% of the variability in NC empirical was accounted for by NC estimate. The average GR CAT $\hat{\theta}$ of -0.43 (median = -0.52) was significantly different from the calibration's average $\hat{\theta}$ of 0.02 (median -0.14); $t = -12.38$, $df = 274$, $p = 0.0001$.

Insert Figure 3 about here

Figure 4 depicts the relationship of the difference between the GR CAT's NC estimates and their empirical values as plotted against NC empirical. As can be seen, for examinees with NC empiricals greater than 25 the GR CAT had a strong tendency to underestimate their ability. However, the GR CAT did not show this bias towards examinees with NC empiricals below or equal to 25. Given the GR model's total test information function (Figure 2), which indicated that the model did not provide very much information for this upper range of ability, this underestimation of high ability examinees was not unexpected.

Insert Figure 4 about here

The NR CAT's convergent cases results are summarized in Table 5. The NR CAT converged on 97.8% of the total cases (269/275). The NR CAT's average $\hat{\theta}$ was -0.07 (median = -0.20) with an average test length of approximately 15.7 items (median = 14); 46 cases received adaptive tests of the maximum test length. The average NC estimate from the NR CAT was 27.3 (median = 27.5) with

an average NC empirical of 27.4 (median = 27.0); the difference between the mean NC estimate and mean NC empirical was not statistically significant ($t = -0.47$, $df = 268$, $p = 0.64$). Further, despite the large sample size the NR CAT's mean $\hat{\theta}$ was not significantly different from the NR model's calibration average $\hat{\theta}$ for these examinees ($t = -0.63$, $df = 268$, $p = 0.53$).

Insert Table 5 about here

Figure 5 presents the scatterplot of NC estimate versus NC empirical. As can be the scattering of points about a straight line is substantially less than found with the GR CAT's NC estimate/NC empirical plot and there does not appear to be any meaningful bias throughout the ability scale. The regression of NC empirical upon NC estimate showed that over 85% of the variability in NC empirical was "explained" by NC estimate; the correlation coefficient between NC estimate and NC empirical was 0.93. Both these values were substantially greater than those of the GR CAT.

Insert Figure 5 about here

A plot of the difference between NC estimate and NC empirical versus NC empirical is presented in Figure 6. Except for some cases of high ability where the NR CAT underestimated the examinee's ability, there does not appear to be a systematic tendency to over- or underestimate ability across the ability continuum. Further, the sparcity of points in the upper ability range, relative to the GR CAT, is an indicator that the NR CAT's slight convergency problem exists in the upper ability region. All

6 nonconvergent cases were examinees whose NC empirical were in the range 46 to 49, where the NR model did not provide very much information. As was the case for the GR CAT, there were a few cases with NC empiricals between 20 and 30 which were underestimated by about 10 points. Given the information available to the NR CAT in the upper ability range it was surprising that the NR CAT performed so well.

Insert Figure 6 about here

To compare the two CATs directly the 269 jointly convergent cases were analyzed. The summary statistics on the jointly convergent cases for the GR CAT are presented in Table 6. Because the GR CAT converged on 100% of the cases only its results will be altered by examination of the jointly convergent cases. Therefore, Table 5 contains the NR CAT's results to be used for comparison with the GR CAT.

As can be seen from Tables 5 and 6, the average and median NIAs for the two CATs were almost equal, although the NR CAT had substantially more variability in the number of items administered. However, a comparison of the two CATs with respect to their average NC estimate shows that the NR CAT performed substantially better in estimating NC empirical; as stated above, the average NC empirical and the NR CAT's average NC estimate were not significantly different from one another. The strong positive correlation coefficient of 0.93 between NR CAT's NC estimate and NC empirical and Figure 5 indicate that this nonsignificant result was not due to large positive differences offsetting large negative differences.

Insert Table 6 about here

Figure 7 depicts the relationship between the difference between the GR CAT's NC estimate and NC empirical versus NC empirical. As would be expected, the elimination of the NR CAT's nonconvergent six cases did not have a meaningful impact on the GR CAT's underestimation of high ability examinees. Furthermore, matched t-tests of the difference between NC estimate and NC empirical as well as the difference between GR CAT's $\hat{\theta}$s and the GR model's calibration $\hat{\theta}$s were statistically significant, $t = -10.05$ and $t = 12.19$ (for both tests : $df = 268$, $p = 0.0001$), respectively.

Insert Figure 7 about here

DISCUSSION

In this simulation study both CATs performed well despite a small item pool (50 items). For instance, the two CATs had very high convergence rates and provided ability estimates which were highly correlated with an external criterion while administering, on the average, less than 16 items. However, given the results presented above (e.g., the matched t-test results, the plots of the difference between the NC estimate - NC empirical versus NC empirical) the NR CAT's performance was superior to the GR CAT's. The NR CAT's ability estimate did not show the systematic underestimation of high ability examinees which was prevalent in the GR CAT's ability estimates. In addition, the NR CAT's NC estimates were highly correlated with and predictive of NC empirical. In contrast to the

GR CAT, the NR CAT did have 17% of its convergent cases terminated by the maximum test length criterion. However, the NR CAT had a sufficiently large number of cases with small test lengths so that its average as well as its median NIA was approximately equal to that of the GR CAT.

Of particular interest was the similarity of the models' total test information functions. The GR and NR models' describe the test similarly with the two models showing differences from one another only below $\theta = -1.0$. The similarity of the information functions in the the upper range of the θ scale was not unexpected. In this range both models are face with a reduction in the variation of responses from the higher ability examinees. As ability increases, examinees make fewer and fewer incorrect responses and the majority of the incorrect responses which do occur are most likely accounted for by one or two incorrect alternatives which are especially attractive to high ability examinees. Therefore, both models are extracting information from what is progressively becoming dichotomous-like data.

For θs below -1.0 there is a great deal of variability in the examinees' responses and the differences between the two models become apparent. It can be seen from Figure 2 that the GR model abstracts more information than the NR model for θs below approximately -2.25. In contrast, the NR model provides more information in the range -2.25 to 1.0 (approximately). Assuming a N(0,1) distribution of ability, the NR model provides more information than does the GR model for approximately 83% of the examinees.

This relationship between the two information functions may result from the information provided to the GR model by the ordinal relationship of the item's alternatives (i.e., the large response values indicate higher ability than lower response values). The GR model can use this additional information for estimating the item and examinee parameters whereas this information is not available for the NR model's estimation. This ordered alternative information may, to a certain extent, offset the influence of guessing on the item parameter estimates (e.g., by reducing the effect of guessing on the estimation of the discrimination parameter) and thereby increase the information available for low ability examinees. If this is true, then a NR-like model with a pseudo-guessing parameter (e.g., Thissen & Steinberg's multiple-choice model (1984), or its equivalent, Sympson's Model III (1983)) should subsume the GR model's information function in the lower range of θ. Alternatively, the GR/NR difference below $\theta = -2.25$ may be an artifact reflecting the inaccuracy of estimating the information function at these extreme levels of ability where there are comparatively fewer observations than around $\theta = 0.0$.

Given the similarity of the information functions in the upper half of the ability scale it is surprising that the two CATs would perform differently in this range of the θ continuum. The GR CAT consistently underestimated examinees with NC empiricals greater than 30. Recall that the longest GR CAT simulated test was 21 items, therefore, the maximum SEE value was not set to low for these high ability examinees. An additional simulation (N=275) of the GR CAT with the maximum SEE set to 0.34 was performed. Analy-

sis of these data showed an increase test length of, on the average, 10 items and did not alleviate the GR CAT's problem with underestimation of high ability examinees.

One potential explanation for the GR CAT's bias in estimating high ability examinees may lie in the distribution of an item's information and its interaction with the item selection strategy. Samejima (1969) has shown that the closer the category difficulty parameters are to one another the greater the information for a narrow θ range (i.e., a leptokurtic item information distribution). Conversely, the greater the distance between $\underline{b}_x$ and $\underline{b}_{x+1}$ the broader the distribution of item information (i.e., the information is distributed over a larger θ range at the expense of becoming more platykurtic).

Inspection of the item information functions for items with some positive $\underline{b}$s (no item had all positive $\underline{b}$s) showed that the majority of these functions were relatively platykurtic. For instance, Figure 8 presents item information functions for such a set of items, specifically items 44 - 50. The items with relatively flat information functions (items 44, 46 and 49) had large differences between $\underline{b}_1$ and $\underline{b}_3$ (9.09, 6.41, and 7.90, respectively) and, as would be expected, these items also had small $\underline{a}$s (0.23, 0.42, and 0.42, respectively). The comparatively peaked information function (item 48) had a $\underline{b}_1/\underline{b}_3$ difference of 2.39 and an $\underline{a}$ of 1.13.

Insert Figure 8 about here

Figure 9 presents the NR model's information functions for

these same items (items 44 - 50). Except for items 44 and 49, these information functions tend to be more peaked than the corresponding functions based on the GR model. This increase in information over a narrower range probably results from the NR model's use of a parameter which indicates the alternative's, not the item's, capability to discriminate among different abilities. In contrast, the GR model uses an "average-like" (i.e., over all alternatives) discrimination parameter for assessing the item's capacity to differentiate among abilities.

Insert Figure 9 about here

The implications of this approach become more apparent in the upper ability range where a fewer number of alternatives are accounting for more of the examinee responses. Due to the decrease in responses to the less attractive alternatives, these distractors will not discriminate well among the examinees and their parameters will not be well estimated. These poorly discriminating alternatives will, in effect, attenuate the discrimination of the more attractive alternatives. The item's a will reflect both the alternatives which discriminate well and those that do not. The attenuation effect of poorly discriminating alternatives will not be as great in the NR model because each alternative's a will assess the option's capacity to discriminate among examinees. The alternative(s) which discriminate well will have large as, whereas the poorly discriminating alternative(s) will have low a values.

It is believed that, for the GR CAT, the proclivity of

items which provide the most information in the upper ability range to have broad and/or platykurtic item information functions (i.e., low discriminatory power) results in the underestimation of high ability examinees. That is, the initial administration of highly discriminating items provides sufficient information to indicate that the examinee's ability is substantially different from 0.0. As a result, after the administration of a few items there have been comparatively large changes in $\hat{\theta}$. However, the subsequent administration of poorly discriminating items only slightly increases $\hat{\theta}$ beyond the $\hat{\theta}$ prior to their administration. In contrast to the GR CAT, for the upper half of the θ scale the NR CAT has available items which are comparatively more discriminating at certain θs than at other θs and therefore, are more likely to be selected for the $\hat{\theta}$ at which their information is most localized. The administration by the NR CAT of these more discriminating items results in the appropriate adjustment(s) of $\hat{\theta}$.

It is interesting to note that the similarity of the total test information functions for the GR and NR models concealed the differences between the two models in the way the item information functions are distributed in the upper ability range.

The educational implications of polychotomous model-based computerized adaptive testing include the possibility of merging computer-aided instruction and diagnostic testing with CAT. It should be noted that a polychotomous model-based CAT could use both items which have to be scored as well as those that do not need to be scored. However, the lack of an item scoring requirement for a polychotomous model-based CAT would allow for computer-aided item

creation for the item pool, a complete IRT item analysis (Thissen & Steinberg, 1984), and for the development and use of new and innovative item formats. In this latter case, these polychotomous items represent a new domain of items which may be used in adaptive testing environments and which would be develop specifically for polychotomous models.

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.

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

De Ayala, R.J. & Koch, W.R. (1987). Computerized adaptive testing : A comparison of the nominal response model and the three parameter model. Paper presented at the annual meeting of the National Council of Measurement in Education, April, Washington, D.C.

De Ayala, R.J., Dodd, B.G. & Koch, W.R. (in press). A comparison of the nominal response model and the three parameter logistic model in computerized adaptive testing. Submitted for publication.

Koch, W.R., De Ayala, R.J. and Dodd, B.G. (1988). Operational characteristics of adaptive testing procedures using the grade response model. Paper presented at the annual meeting of the American Educational Research Association, April, New Orleans.

Koch, W.R. & Dodd, B.G. (1985). Computerized adaptive attitude measurement. Paper presented at the annual meeting of the American Educational Research Association.

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

Samejima, F. (1977). The applications of graded response models: The promise of the future. In D. Weiss (Ed.) Proceedings of the 1977 Computerized Adaptive Testing Conference, Minneapolis, MN.

Samejima, F. (1979). A new family of models for the multiple-choice item, (Research Report 79-4). Knoxville: University of Tennessee, Department of Psychology, 1979.

Sympson, J.B. (1986). A new item response theory model for calibrating multiple-choice items. Paper presented at the meeting of the Psychometric Society, Los Angeles, CA.

Thissen, D.J. & Steinberg, L. (1984). A response model for multiple choice items. Psychometrika, 49, 501-519.

Thissen, D.J. (1986). MULTILOG-User's Guide (Version 5). Scientific Software, Inc. Mooresville, IN.

Table 1 : Descriptive Statistics of the data sets

Data Set	N	NC Mean	S.D.	Alpha
Collapsed	1093	27.3	7.5	0.85
Complete	275	27.8	8.5	0.88

Table 2 : Principal Axes Analyses : Factors with eigenvalues greater than 1.0.

	Eigenvalues			Variance Accounted for by Factor I	
Data Set	I	II	III	Total	Common
Collapsed	6.902	1.299	1.036	13.8%	74.7%
Complete	6.902	1.299	1.035	13.8%	74.7%

Table 3 : $\hat{\theta}$s from the GR and Nominal Response models calibrations of the Collapsed data set (N=1093).

Model	θ Mean	SD	Min	Max	Median	r^a	r^b
GR	0.12	0.87	-2.05	2.88	0.078		
Nominal	0.06	0.96	-2.17	3.05	0.017	0.99	0.99

[a] Pearson Product Moment Correlation Coefficient between Nominal Response and GR Models' Calibration $\hat{\theta}$s

[b] Spearman Rank-Order Correlation Coefficient between Nominal Response and GR Models' Calibration $\hat{\theta}$s

Table 4 : GR CAT simulation's convergent cases (N = 275).

	Mean	SD	Min	Max	Mdn	r^a
$\hat{\theta}$	-0.43	0.86	-2.83	1.53	-0.52	
SEE	0.43	0.01	0.37	0.44	0.43	
NIA	15.8	1.37	14.0	21.0	16.0	
NC estimate	25.0	6.60	7.6	38.3	24.6	0.85
NC empirical	27.8	8.51	10.0	49.0	27.0	

[a] Pearson Product Moment Correlation Coefficient between NC estimate and NC empirical

Table 5 : Nominal CAT simulation's convergent cases (N = 269).

	Mean	SD	Min	Max	Mdn	r^a
$\hat{\theta}$	-0.07	1.28	-2.80	3.6	-0.20	
SEE	0.46	0.05	0.41	0.70	0.45	
NIA	15.7	7.82	7.0	30.0	14.0	
NC estimate	27.3	8.54	6.7	44.5	27.5	0.93
NC empirical	27.4	8.07	10.0	46.0	27.0	

[a] Pearson Product Moment Correlation Coefficient between NC estimate and NC empirical

Table 6: GR CAT simulation - jointly convergent cases (N = 269).

	Mean	SD	Min	Max	Mdn	r[a]
$\hat{\theta}$	-0.46	0.84	-2.83	1.53	-0.53	
SEE	0.43	0.01	0.37	0.44	0.43	
NIA	15.8	1.34	14.0	21.0	15.0	
NC estimate	24.8	6.51	7.6	38.3	24.4	0.85
NC empirical	27.4	8.07	10.0	46.0	27.0	

[a] Pearson Product Moment Correlation Coefficient between NC estimate and NC empirical

Figure 1

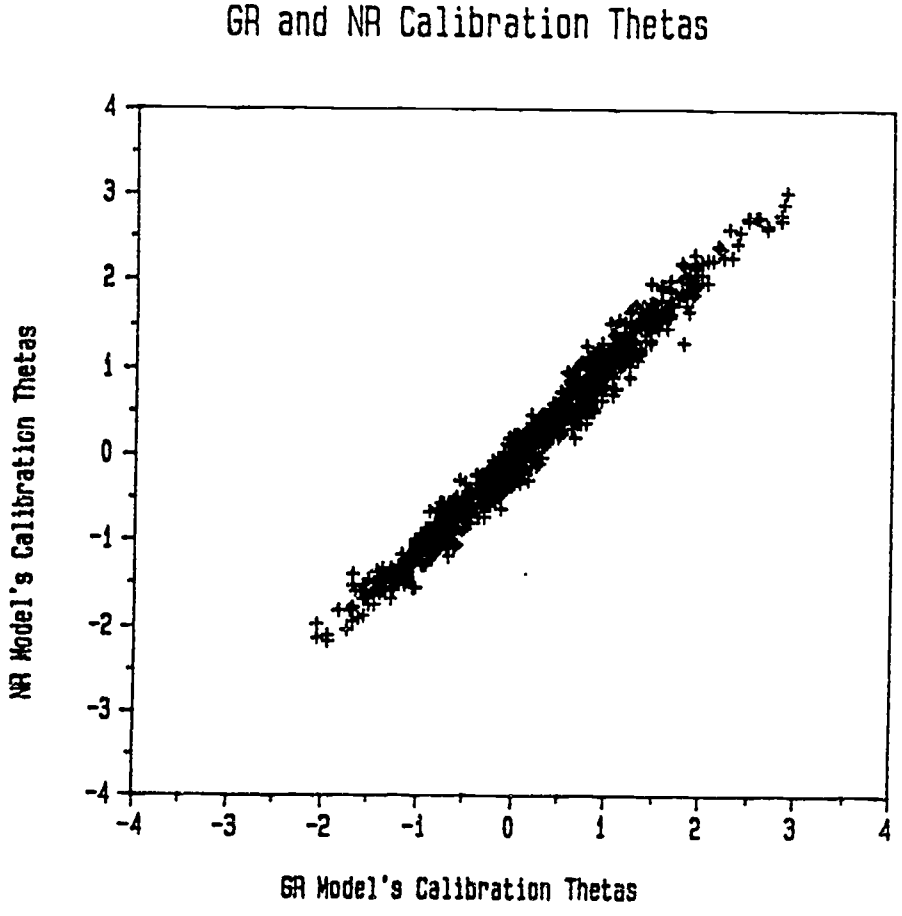

GR and NR Calibration Thetas
NR Model's Calibration Thetas
GR Model's Calibration Thetas
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4

Figure 2

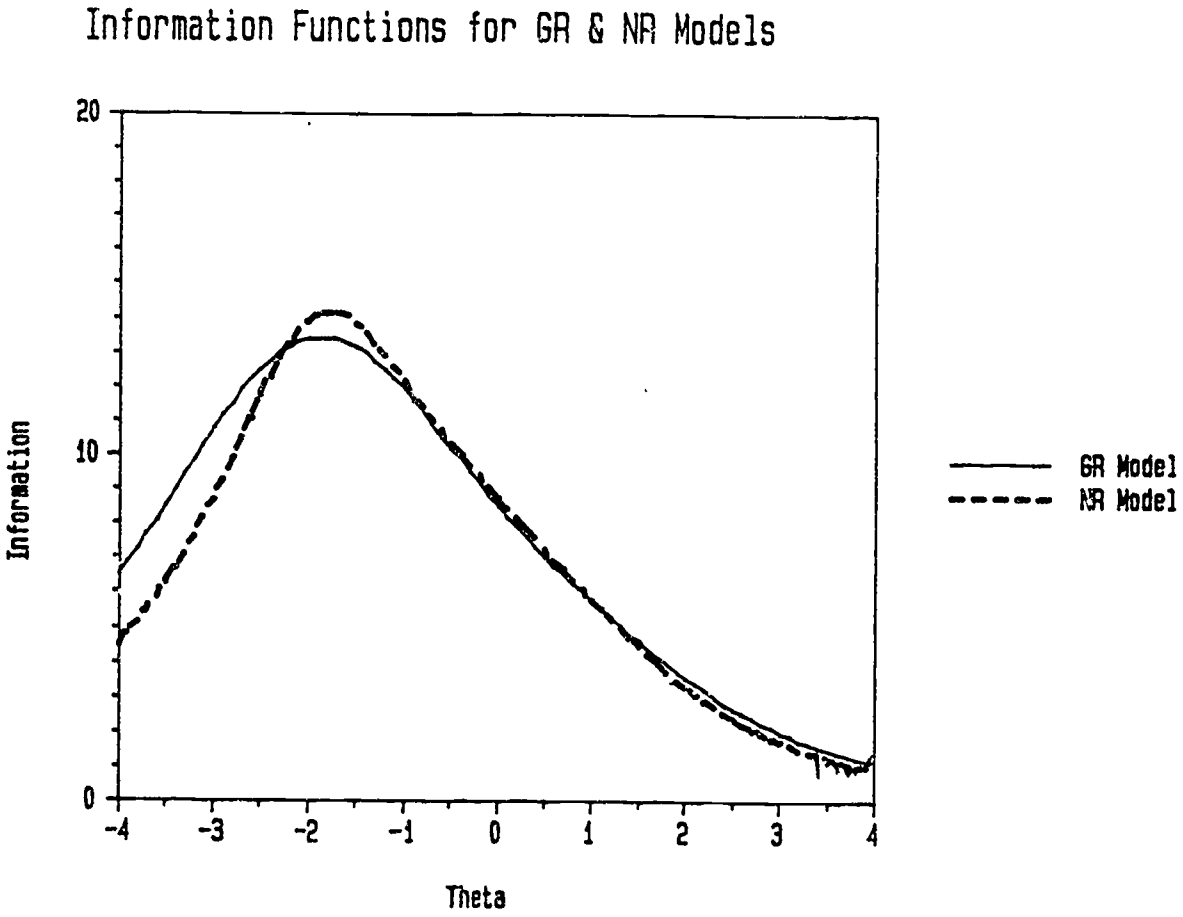

Information Functions for GR & NR Models
20
10
0
Information
-4
-3
-2
-1
0
1
2
3
4
Theta
GR Model
NR Model

Figure 3

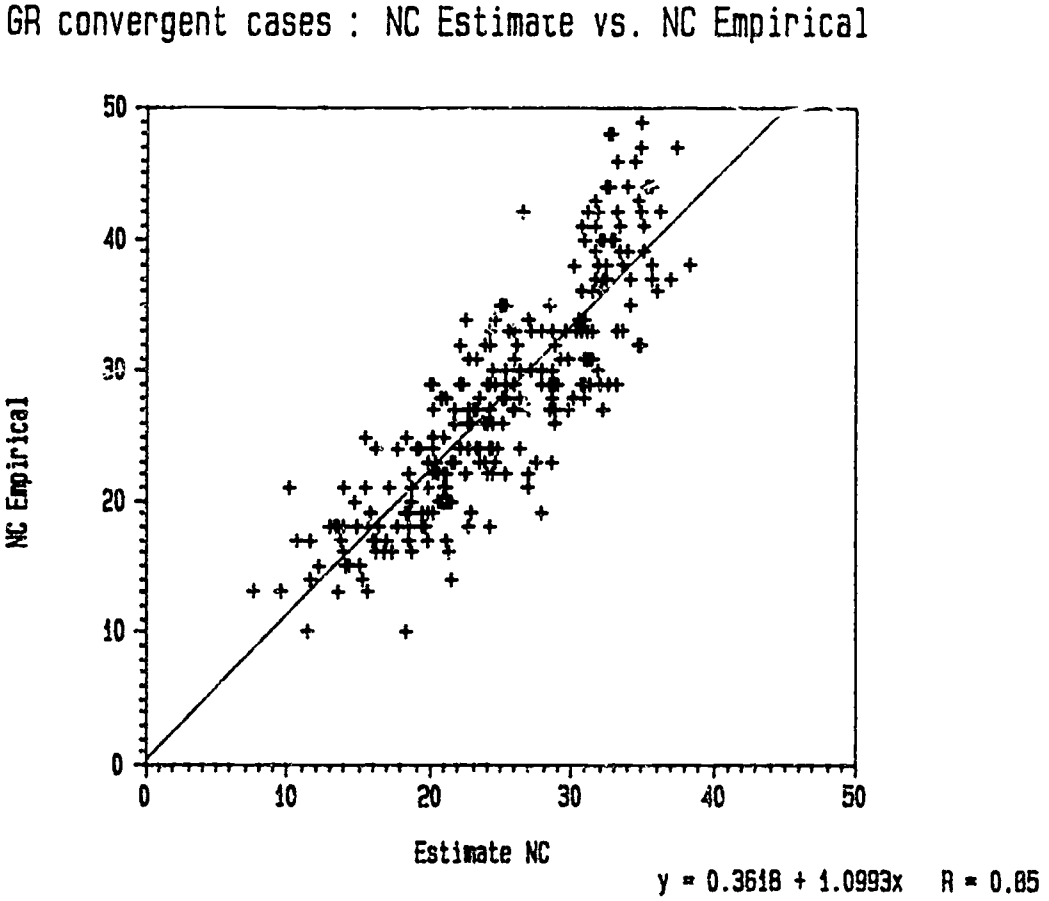

GR convergent cases : NC Estimate vs. NC Empirical
NC Empirical
Estimate NC
0
10
20
30
40
50
y = 0.3618 + 1.0993x R = 0.85

Figure 4

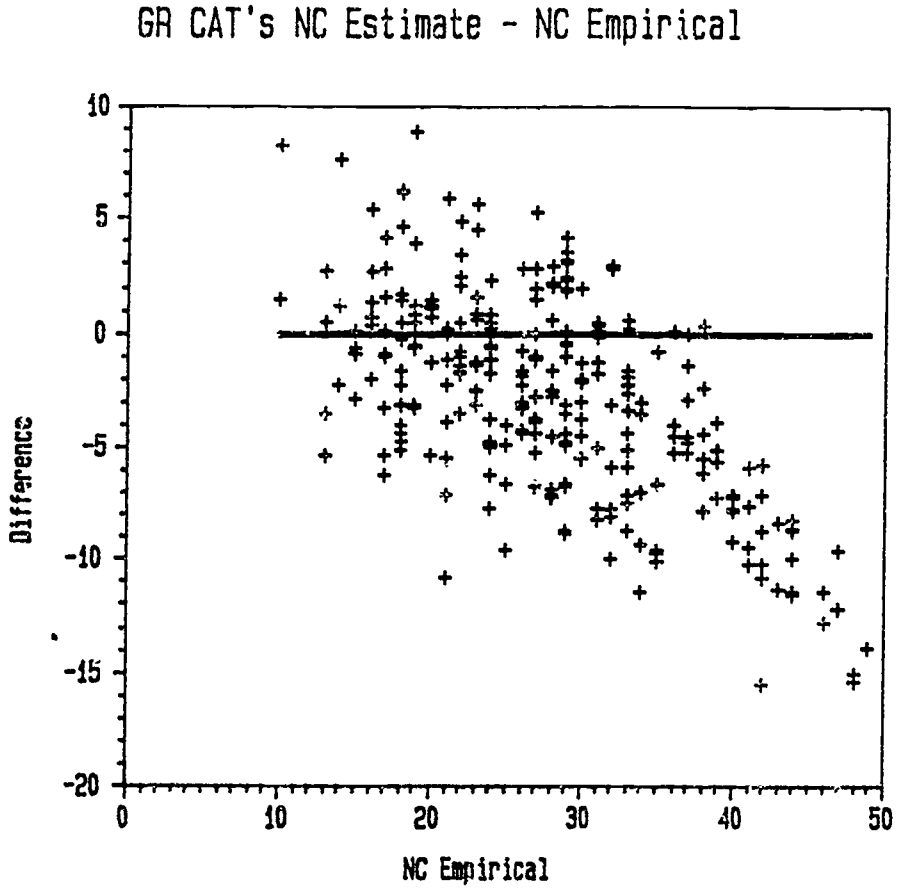

GR CAT's NC Estimate - NC Empirical
Difference
10
5
0
-5
-10
-15
-20
0
10
20
30
40
50
NC Empirical

Figure 5

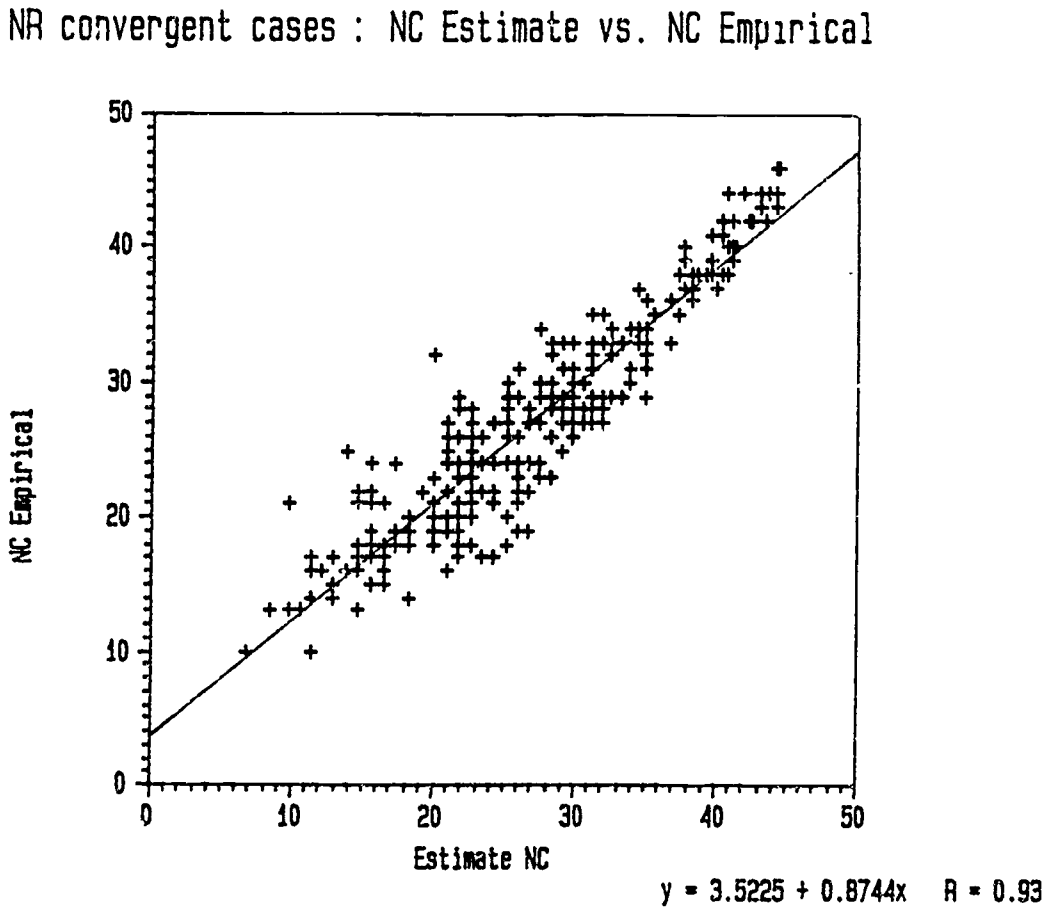

NR convergent cases : NC Estimate vs. NC Empirical
NC Empirical
Estimate NC
0
10
20
30
40
50
y = 3.5225 + 0.8744x R = 0.93

Figure 6

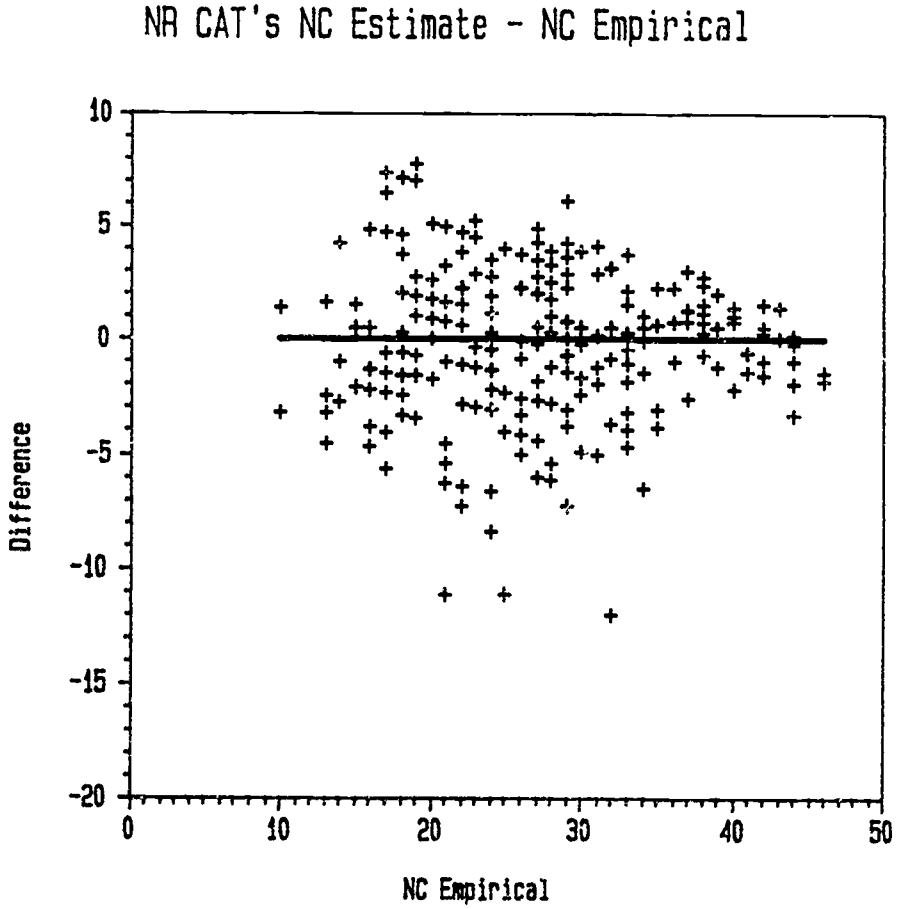

NR CAT's NC Estimate - NC Empirical
Difference
10
5
0
-5
-10
-15
-20
0
10
20
30
40
50
NC Empirical

Figure 7

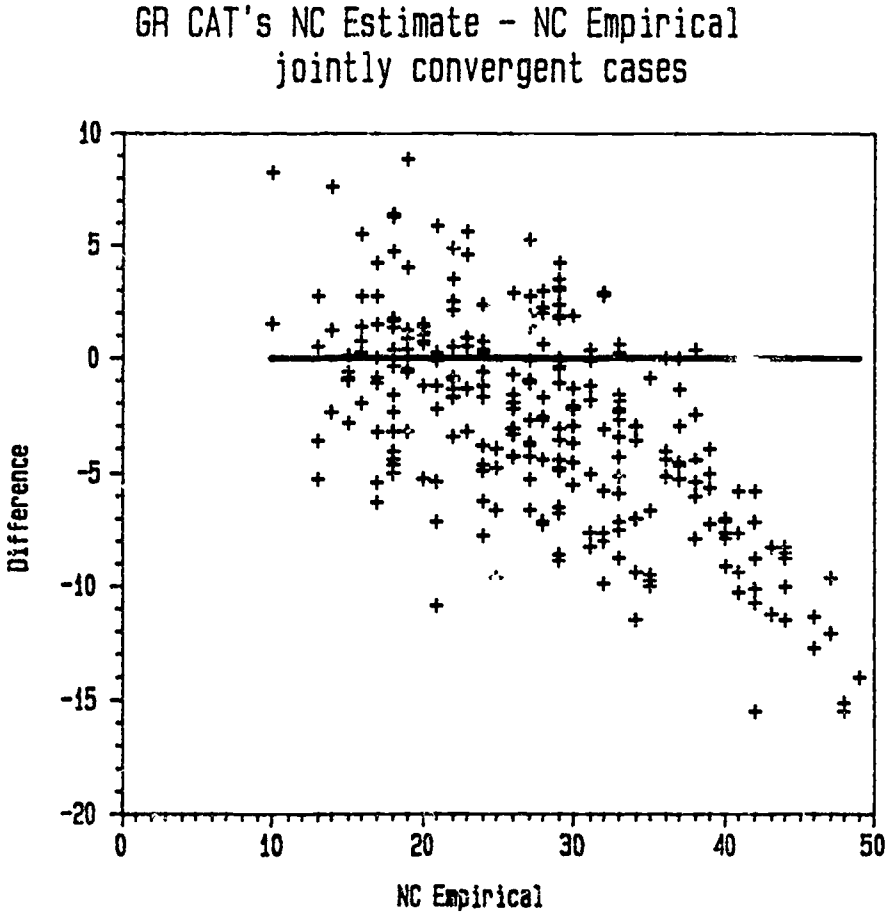

GR CAT's NC Estimate - NC Empirical
jointly convergent cases
Difference
NC Empirical
10
5
0
-5
-10
-15
-20
0
10
20
30
40
50

Figure 8

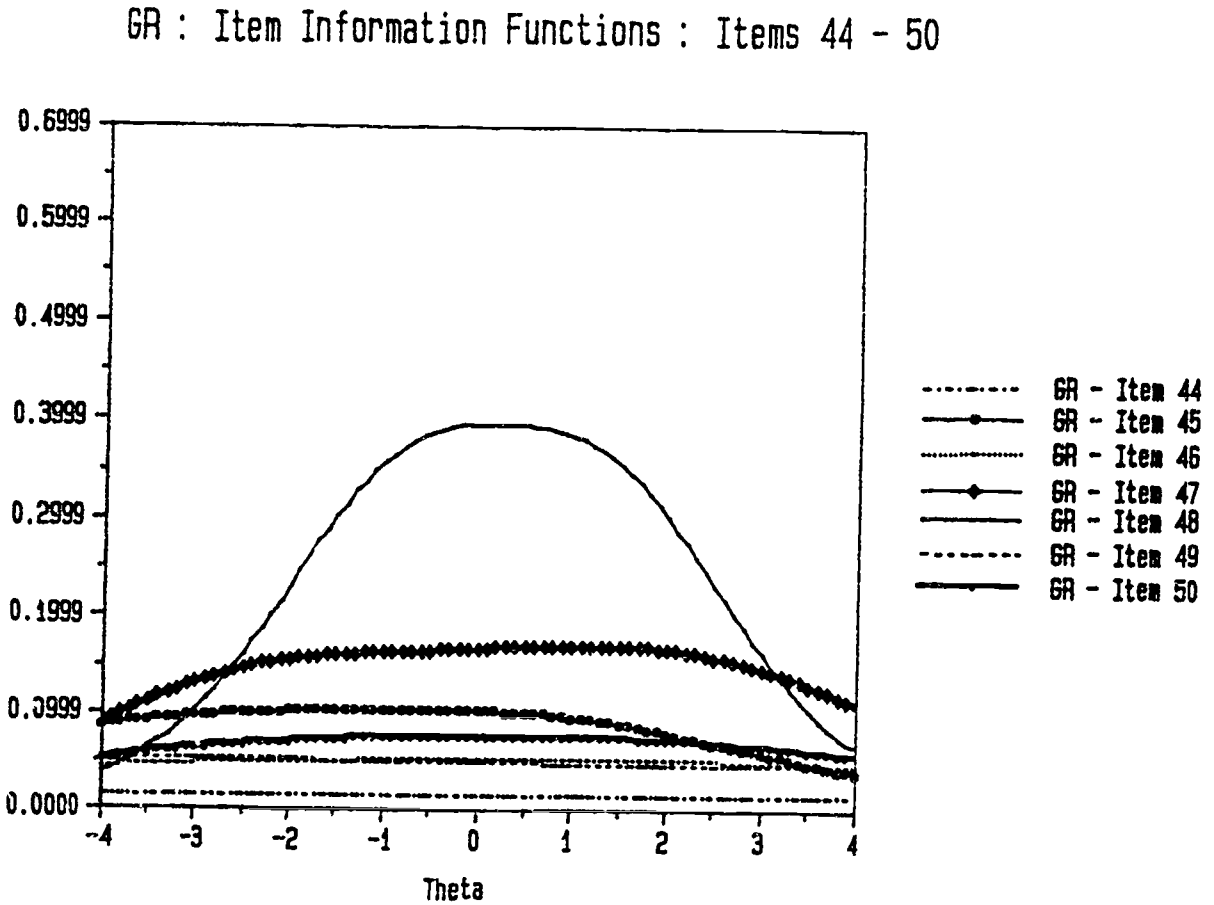
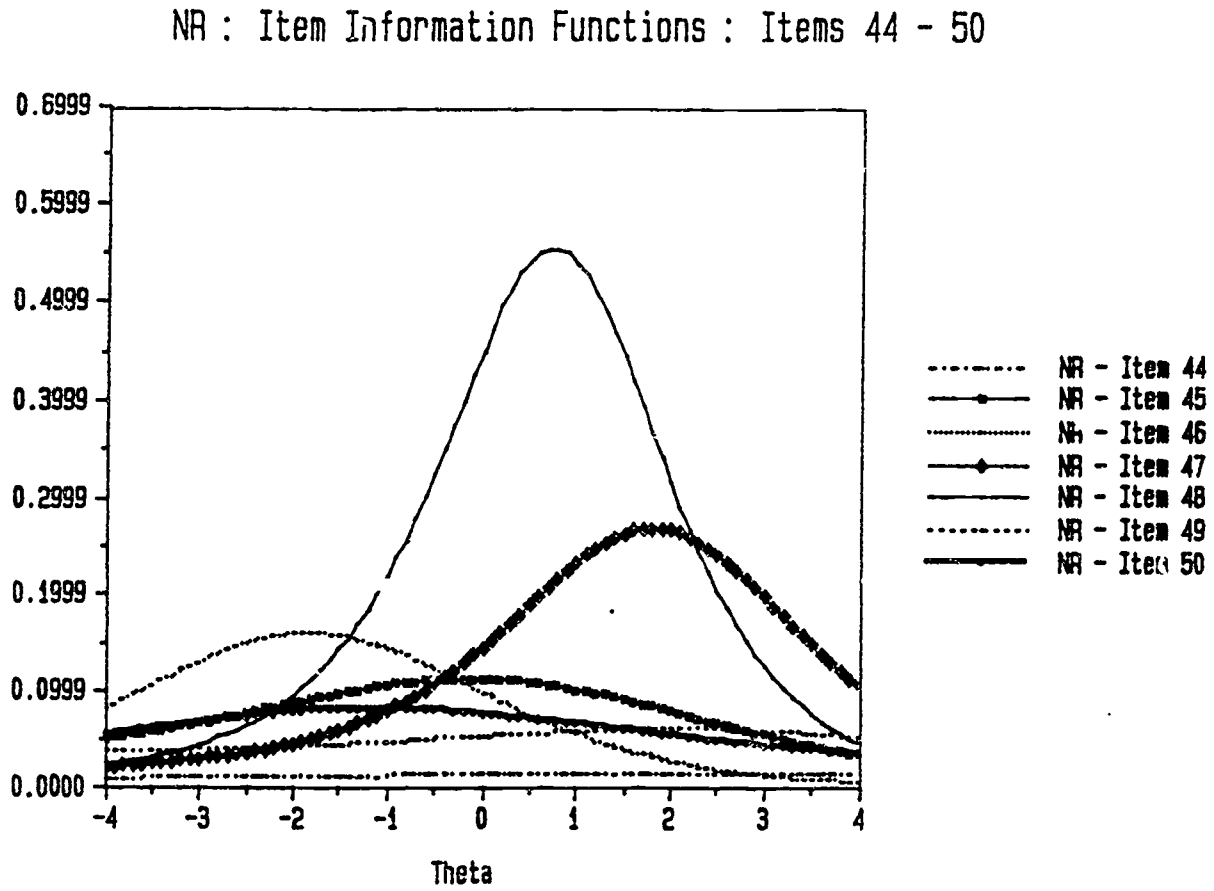

GR : Item Information Functions : Items 44 - 50
0.6999
0.5999
0.4999
0.3999
0.2999
0.1999
0.0999
0.0000
-4
-3
-2
-1
0
1
2
3
4
Theta
GR - Item 44
GR - Item 45
GR - Item 46
GR - Item 47
GR - Item 48
GR - Item 49
GR - Item 50

Figure 9


NR : Item Information Functions : Items 44 - 50
0.6999
0.5999
0.4999
0.3999
0.2999
0.1999
0.0999
0.0000
-4
-3
-2
-1
0
1
2
3
4
Theta
NR - Item 44
NR - Item 45
NR - Item 46
NR - Item 47
NR - Item 48
NR - Item 49
NR - Item 50