ED 320 920 TM 014 986

AUTHOR	De Ayala, R. J.; And Others
TITLE	A Comparison of the Partial Credit and Graded Response Models in Computerized Adaptive Testing.
PUB DATE	Apr 90
NOTE	17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Ability Identification; *Adaptive Testing; Comparative Analysis; *Computer Assisted Testing; *Estimation (Mathematics); Mathematical Models; Maximum Likelihood Statistics; *Scoring; *Simulation; Test Items
IDENTIFIERS	Ability Estimates; *Graded Response Model; *Partial Credit Model

ABSTRACT

Computerized adaptive testing procedures (CATPs) based on the graded response method (GRM) of F. Samejima (1969) and the partial credit model (PCM) of G. Masters (1982) were developed and compared. Both programs used maximum likelihood estimation of ability, and item selection was conducted on the basis of information. Two simulated data sets, one with 1,000 simulated examinees and one with 500 simulated examinees, were generated according to a linear analytic model. Both contained responses to 180 five-alternative items, of which 55 were retained for 997 simulated examinees with infit statistics between -3.0 and 3.0. The MULTILOG calibration program of D. J. Thissen (1988) was used to obtain item parameter estimates for both models from the data set. It was expected that using a data set fitted to the PCM model would result in no differences between PCM and GRM CATPs. However, the GRM CATPs provided more accurate information than did the PCM CATPs and the estimation was considered adequate. Reasons for difficulties with the PCM model are discussed. Two tables and five graphs present study data. (SLD)

A Comparison of the Partial Credit and Graded Response Models in Computerized Adaptive Testing

R.J. De Ayala,
University of Maryland
Barbara G. Dodd and William R. Koch,
University of Texas at Austin

OBJECTIVES

The objectives of this research were (a) to develop and implement a computerized adaptive testing (CAT) procedure based on Samejima's (1969) graded response (GR) model and Masters' (1982) partial credit (PC) model, and (b) to compare the GR-based CAT performance in ability estimation with that of the PC-based CAT.

MODEL DESCRIPTIONS

The two polychotomous models, GR and PC, are appropriate for items with ordered responses, such as aptitude and achievement test items whose alternatives are inherently ordered or have been ordered according to degree of correctness (e.g., through partial credit scoring). In addition, attitude questionnaires and ratings data may also be fitted by either model.

The GR model is a direct extension of the two-parameter model. As a result, the GR model contains a parameter which allows an assessment of an item's capacity to discriminate among examinees. In the GR model the examinee responses to item i are categorized into $m_i + 1$ categories, where higher categories indicate more of θ and m_i is the number of category boundaries. Associated with each category of item i is a category score, x_i, with values $0..m_i$. The GR model may be expressed as :

$$P_{x_i}(\theta) = \frac{e^{D a_i(\theta - b_{x_i})}}{1 + e^{D a_i(\theta - b_{x_i})}} \qquad (1),$$

where θ is the latent trait, a_i is the discrimination parameter for item i, b_{x_i} is the difficulty parameter for category score x for item i, and the scaling constant D equals 1.702. P_{x_i} is the probability, p_{x_i}, of the examinee responding in category score x_i or higher for a given item; the probability of responding in the lowest category (i.e., $P_0(\theta)$) or higher is defined as 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_i} is the probability of responding in x_i or higher, the probability of responding in a particular category equals the difference between cumulative probabilities for

Paper presented at the AERA Annual Meeting, April 1990, Boston adjacent categories (e.g., $p_2(\theta) = P_2(\theta) - P_3(\theta)$). When an item consists of two categories (correct and incorrect), the GR model reduces to the two-parameter model.

In contrast to the GR model, the PC model provides a direct expression of the probability of an examinee with ability θ responding in a particular category. In the PC model the examinee-item interaction is modeled as :

$$P_{x_i}(\theta) = \frac{e^{\sum_{j=0}^{x_i}(\theta - b_{x_i})}}{\sum_{k=0}^{m_i} e^{\sum_{j=0}^{k}(\theta - b_{x_i})}} \qquad (2),$$

where θ is the latent trait, b_{x_i} is the difficulty parameter of the step associated with category score x_i of item i with m_i categories, where $x_i=1..m_i$. A category score reflects the number of successfully completed steps. A "step" is simply a stage required to complete an item. For instance, the problem $((6/3)+2)^2$ is considered to contain three steps because there are three separate stages which must be completed (in a specific order) to correctly answer the problem (i.e., step 1 : 6/3, step 2 : the addition of 2 to the quotient, and step 3 : the squaring of the quantity) For notational convenience $\Sigma(\theta - b_{x_i})$ where j=0 is defined as being equal to zero.

Because the PC model is an extension of the Rasch model it assumes that all items are equally good at discriminating among examinees. In addition, as a member of the Rasch family, the PC model's item and person parameters may be estimated on the basis of the existence of sufficient statistics. Specifically, an examinee's test score contains all the information for estimating his or her ability and the items' difficulties may be estimated from a simple count of the number of persons completing each "step" of an item. Unlike the GR model, the PC model requires that the steps within an item be completed in sequence, although the steps need not be equally difficult nor be ordered in terms of difficulty. If an item consists of only two categories, then the PC model reduces to the Rasch model.

Except for a few researchers (e.g., Dodd, Koch, & De Ayala, 1989; De Ayala, 1989; Sympson, 1986) CAT research has been primarily concerned with dichotomous item response theory (IRT) models. However, a number of exams are scored in a graded fashion. For example, statistics, mathematics, chemistry, and physics exams are typically graded by given partial credit for some incorrect answers. It is reasonable and desirable (i.e., for the acceptance of CAT application to these area) to expect that CAT implementations in these subjects to incorporate a graded scoring system. In addition, relative to a dichotomous model-based CAT the use of a polychotomous model permits the use of the examinee's partial knowledge of the correct response for their ability estimation and should result in decreased test length; the existence of information in incorrect responses has been demonstrated in several studies (Levine & Drasgow, 1983; Thissen, 1976).

METHOD

Programs : Two CAT programs were written, one program was based on the PC model (called the PC CAT), whereas the other was based on the GR model (GR CAT). Both programs used maximum likelihood estimation of ability and item selection was on the basis of information. The adaptive testing simulation was terminated when either of two criteria were met : a maximum of twenty items was reached oı when a predetermined standard error of estimate (SEE) was obtained (SEE termination criteria of 0.10, 0.25, 0.30 were used). Previous work with polychotomous model-based CATs has shown that SEE results in better CAT performance than does the minimum item information criterion (e.g., Dodd, Koch, & De Ayala, 1989). The initial ability estimate for an examinee was the population's mean.

Data : Two simulation data sets were generated according to a linear factor analytic model (Wherry, Naylor, Wherry, & Fallis, 1965). Both data set were unidimensional and contained responses to 180 5-alternative items. One data set contained 1000 simulees (randomly selected from a N(0,1) distribution) and was used for obtaining item parameter estimates; this data was called the calibration data set. The second data set (called the CAT data set) consisted of responses from 500 simulees (randomly selected from a N(0,1)) to the same 180 items as the calibration data set; the z-values used for generating responses were considered to be the simulees' true ability (θ_T). The CAT data set was used for the simulated CATs. The use of a linear factor analytic approach for generating the data sets minimized any bias in favor of one IRT model or the other. All factor loadings were uniformly high and ranged from 0.62 to 0.85. Further, the use of separate data sets for calibration and CAT simulations minimize capitalizing on chance by using the same data set for both the calibration of the item pool as well as in the CAT simulations.

MULTILOG (Thissen, 1988) was used to obtain item parameter estimates for both the PC and GR models from the calibration data set. The use of a single calibration program for both models controlled for differences in the implementation of estimation algorithms when different calibration programs are used. Although the item parameter estimates used for the CAT simulations were obtained from MULTILOG, MSTEPS (Wright, Congdon, & Schultz, 1989) was used to obtain fit statistics for the PC model.

Analysis : The simulation 1000 examinee by 180-item data set was fitted to the PC model. Items which were found to fit the PC model were used to form an item pool for the PC CAT. PC and GR item parameter estimates were obtained for this fitted set of items. In addition, GR item parameters were estimated for the original 180-item set. The CAT simulations were analyzed by comparing each CAT's estimated ability ($\hat{\theta}$) with θ_T. These comparisons involved correlational analysis (Pearson product-moment and Spearman rank-order correlation coefficients), standardized root mean squared differences (SRMSD), standardized differences between means (SDM), and descriptive statistics. The differences between $\hat{\theta}$ and θ_T were graphically examined. Further, descriptive statistics on the number of items administered by each CAT were calculated and the relationship of SEE to θ_T was also inspected.

RESULT

Calibrations

Fifty-five items with weighted total fit statistics between -3.0 and 3.0 were retained for use with the PC CAT. Further, 997 simulees were found to have infit statistics between -3.0 and 3.0. Therefore, the PC calibration was performed on 55-item pool (a.k.a., the PC calibration data set) with 997 examinees. Item parameter estimates for the GR model were obtained for both the 55-item pool and the original 180-item pool; the three examinees identified as not fitting the PC model were retained for the GR calibrations. In the following the GR CAT using the 55-item pool will be referred to as the GR-55 CAT, whereas GR-180 CAT will indicate the GR CATs with the 180-item pool. Dodd, Koch, and De Ayala (1989) and Koch and Dodd (1989) have been successful in using item pools of about this size in GR and PC CAT simulations, respectively. The 500 examinee/55-item data set used for the CAT simulations will be referred to as the CAT data set.

Item Pools

The PC 55-item pool had step difficulty estimates which ranged from -2.365 to 3.124, with a positively skewed distribution of difficulties for the first step difficulty, a negatively skewed distribution of difficulties for the last step difficulty, and more or less unimodal difficulty distributions for the second and third step difficulties. The GR 55- and 180-item pools had average discrimination estimates of 1.320 (median=1.300, standard deviation=0.102) and 1.467 (median=1.453, standard deviation=0.255), respectively. The difficulty estimates for the GR 55-item pool ranged from 4.093 to 4.189 and from -4.527 to 4.924 for GR 180-item pool. For all category scores in the GR 180-item pool the difficulty estimates tended to be normal-like in distribution, whereas for the GR 55-item pool the distributions for the first and third difficulty estimates were positively skewed and these distributions were rectangular-like for the second and fourth difficulties. Given Urry's (1977) guidelines, the item pool for the GR CATs consisted of desirable items. It would have been desirable to have items with step difficulties below -2.365, however, the absence of these items was not problematic for the PC CAT. Figure 1 shows the total item pool information for both the GR-55 and the PC item pools; the estimate of the information function for the GR-180 item pool was similar to and about twice that of the GR-55 item pool. As can be seen the PC 55-item pool provides greater information than the GR-55 item pool for the approximate range -2.25 to 2.5. Because the simulees abilities were generated from a normal distribution the majority of the examinees had abilities within ±2.0 standard deviations about 0.0. The observed percent of examinees with abilities greater than 2.0 and less than -2.0 was 14% and only 1.4% of the simulees had abilities outside the range -3.0 to 3.0.

Insert Figure 1 about here

CAT Simulations

For the PC CAT simulations the correlation coefficients between $\hat{\theta}$ and θ_T decreased with increases in the SEE termination criterion. As can be seen from Table 1 all correlation coefficients are equal to or above 0.93 and the corresponding scatterplots showed strong linear associations. The correlation coefficients for the GR-180 CAT simulations followed the same pattern as for the PC CAT simulations, albeit with slightly higher values. In contrast, for the GR-55 CAT increases in the SEE termination criterion had no effect on the correlation coefficients between $\hat{\theta}$ and θ_T. The linear relation between $\hat{\theta}$ and θ_T as assessed by the Pearson product-moment correlation coefficient was slightly higher for the GR-55 CAT than for the PC CAT, although the Spearman rank-order coefficients were lower for the GR-55 CAT than those of the PC CAT for all SEE termination criteria, except for the SEE termination criterion of 0.30. On the average, the GR CATs administered slightly longer tests than did the PC CAT

Insert Table 1 about here

SRMSD provides an assessment of the accuracy of estimation across examinees, while SDM assesses the overall bias between the $\hat{\theta}$s and θ_Ts. The SRMSD and SDM for the CATs are presented in Table 2. As can be seen, regardless of whether the 55-or 180-item pool was used the SRMSDs for the GR CATs were approximately one-third that of the PC CATs. This indicated that the GR CATs were providing ability estimates which were comparatively more accurate that those of the PC CATs. On average, the GR-180 CATs $\hat{\theta}$

were very similar to $\bar{\theta}_T$ ($\bar{\theta}_T$ = 0.076 for the 494/497 convergent cases). Similarly, the GR-55 CATs $\hat{\theta}$ were close to the average θ_T ($\bar{\theta}_T$ = 0.076 for the 489/491 convergent cases). Further, the SDMs for the GR CATs revealed a slight overall underestimation of θ_T. The bias for the GR-55 and GR-180 CATs (SEE = 0.25) is graphically depicted in Figures 2 and 3, respectively. These figures are typical of the pattern exhibited by the other GR CATs. As can be seen from these figures, the GR CATs had a tendency to overestimate $\theta_T > 1.0$ and to underestimate $\theta_T < -1.0$.

Insert Table 2 about here

Figures 2 and 3 about here

The SDMs for the PC CATs showed that there was a strong tendency to overestimate θ_T. This was also apparent from a comparison of the mean $\hat{\theta}$ and the average $\bar{\theta}_T$ of 0.083 for all 500 simulees (for the 470 convergent cases the $\bar{\theta}_T$ = -0.031). The relationship between the ($\hat{\theta}$-θ_T) difference and θ_T (Figure 4) for the PC CAT SEE = 0.25 showed that there was a tendency to overestimate throughout the ability scale; this pattern was typical of the other two PC CATs. It was not surprising given the shape of the information function that, in general, larger SEEs (e.g., SEE >0.35) tended to be associated with high $\hat{\theta}$ (e.g., $\hat{\theta}$ >3.00).

Insert Figure 4 about here

Convergence

The convergence rate for the GR CATs were over 97.8%. For the GR-180 CATs with termination SEEs of 0.10 and 0.25, two of the nonconvergent cases were high ability examinees (θ_T = 2.774 & 2.084), while the third case was a very low ability examinee (θ_T = -3.025); the six nonconvergent cases for the GR-180 CAT (SEE = 0.30) had θ_Ts of 1.665, 1.779, 2.084, 2.774, -1.479, and -1.479. Similarly, the GR-55 CATs nonconvergent cases were distributed throughout the ability range. In contrast, the majority of the nonconvergent cases for the PC CAT were associated with $\theta_T \cong$ 2.0 simulees (convergence rate = 94%); the nonconvergence was nonsymmetric. The three PC CATs were unable to estimate the same 30 simulees and four of the 30 cases were examinees for which the GR-55 CAT was unable to obtain an ability estimate (θ_T =1.784, θ_T =2.576, θ_T =2.774, θ_T =3.162). Figure 5 shows the relationship between the GR-55 CATs and the PC CAT nonconvergent cases. Infit statistics calculated for the CAT data set revealed thirty-six examinees with fit values greater than 2.0, only three of which were nonconvergent cases.

Insert Figure 5 about here

DISCUSSION

Given the similarity in results for the GR-55 and GR-180 CATs, it appears that item pools smaller than are suggested fcr dichotomous model-based CATs can be used with GR model-based CATs. It was expected that using a data set which was fitted to the PC model would result in no differences between the GR and PC CATs. However, despite this characteristic and the fact that the PC model provided more information for 86% of the examinees than did the GR model, the GR-55 CATs provided more accurate estimation than the PC CATs. In the authors' opinions the results of the GR CATs were acceptable.

The fitting of the CAT data set to the PC model identified eight items which no longer fit the model (i.e., infit values greater than 3.0), although they had fit the PC calibration data set. The misfitting CAT data set items and the examinees were retained for the CAT simulations because in an real-life implementation this information would only be available post hoc. That is, after a CAT was operational and the misfit information had been gathered, it would be difficult to justify to an examinee that he/she had to be eliminated because on the basis of his/her performance on the adaptive test he/she was found not to fit the CAT's IRT model. Conceivably, the misfitting items could be eliminated from future use in the CAT, although the items would still have had an effect on the examinees who had already been administered the tailored tests. Therefore, the retention of misfitting items and examinees for the CAT simulations was consistent with a the procedures of a real-life CAT implementation. Further, given that only three of the simulees did not fit the PC model, it does not appear that the PC CAT nonconvergent cases were a result of simulees which did not fit the PC model. The role of the misfitting items on the PC CAT convergence and bias is not known.

It may be speculated that some of the PC CAT's difficulties are a result of MULTILOG's implementation of the PC model. That is, in MULTILOG PC parameter estimation requires imposing triangular contrasts on Bock's (1972) nominal response (NR) model (cf., Thissen & Steinberg, 1986). Imposing these triangular contrasts or the NR model is the logical equivalent of making the a priori order assumption necessary for the PC model (Thissen, 1988; Masters & Wilson, 1988). In this regard, the calibration of the data showed that the a of best fit for PC model was 0.754, not $a = 1.0$ as the Rasch PC model assumes. As would be expected given the differences in estimation techniques between MULTILOG and the Rasch program MSTEPS, as well as the difference in the approach to fixing the scale's origin, the programs' difficulty estimates were not equal. However, there was a very high linear agreement between the two sets of estimates ($r_{b_1} = 0.989$, $r_{b_2} = 0.977$, $r_{b_3} = 0.986$,

r_{b4} = 0.995). Given the similarity in the magnitudes of the item parameter estimates as well as the above correlations it does not appear that the results are due to MULTILOG's implementation of the PC model.

Because the PC and GR CAT programs have been successfully used in previous studies (Koch & Dodd, 1989; Dodd, Koch & De Ayala, 1989) it is not likely that the CAT programs were at fault. A possible explanation for the PC CAT's difficulties may be the use of an infit criterion of ±3.0 for retaining items; a more conservative criterion may be required for the creation of PC item pools. Future research will investigate the relationship between the degree of fit of items to the PC model for inclusion to an item pool and PC CAT ability estimation.

References

De Ayala, R.J. (1989). Computerized adaptive testing: A comparison of the nominal response model and the three-parameter model. *Educational and Psychological Measurement, 49*, 789-805.

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.

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.

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

Masters, G. & Wilson, M. (1988, April). *Understanding and using partial credit analysis : an IRT method for ordered response categories.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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

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.

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

Thissen, D.J. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.

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

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

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.

Wright, B.D., Congdon, R., & Schultz, M. (1989). *A user's guide to MSTEPS* (Version 2.4). Chicago : MESA Psychometric Laboratory.

Table 1 : Correlation coefficients between $\hat{\theta}$ and θ_T ($\bar{\theta}_T$=0.083, σ =1.024) and descriptive statistics on CAT

		Correlation		Average	SD	Mean	SD	Number of con-
CAT	SEE	Pearson	Spearman	$\hat{\theta}$	$\hat{\theta}$	NIA[1]	NIA[1]	vergent cases
GR-180 CAT								
	0.10	0.966	0.964	0.068	1.398	20.0	0.0	497
	0.25	0.961	0.958	0.024	1.379	14.396	1.229	497
	0.30	0.937	0.927	-0.010	1.391	10.215	1.073	494
GR-55 CAT								
	0.10	0.961	0.958	-0.001	1.481	20.0	0.0	489
	0.25	0.961	0.958	-0.001	1.481	20.0	0.0	489
	0.30	0.961	0.958	-0.004	1.472	17.179	1.073	491
PC CAT								
	0.10	0.959	0.973	0.957	1.068	20.0	0.0	470
	0.25	0.948	0.960	0.918	1.100	12.672	3.609	470
	0.30	0.933	0.940	0.925	1.115	9.306	3.983	470

NIA[1] : number of items administered

Table 2 : SRMSD and SDM for PC and GR CATs

CAT	SEE	SRMSD	SDM
GR-180 CAT			
	0.10	0.325	-0.012
	0.25	0.332	-0.049
	0.30	0.381	-0.077
GR-55 CAT			
	0.10	0.373	-0.065
	0.25	0.373	-0.065
	0.30	0.370	-0.064
PC CAT			
	0.10	1.028	0.984
	0.25	0.973	0.929
	0.30	0.982	0.928

Figure 1

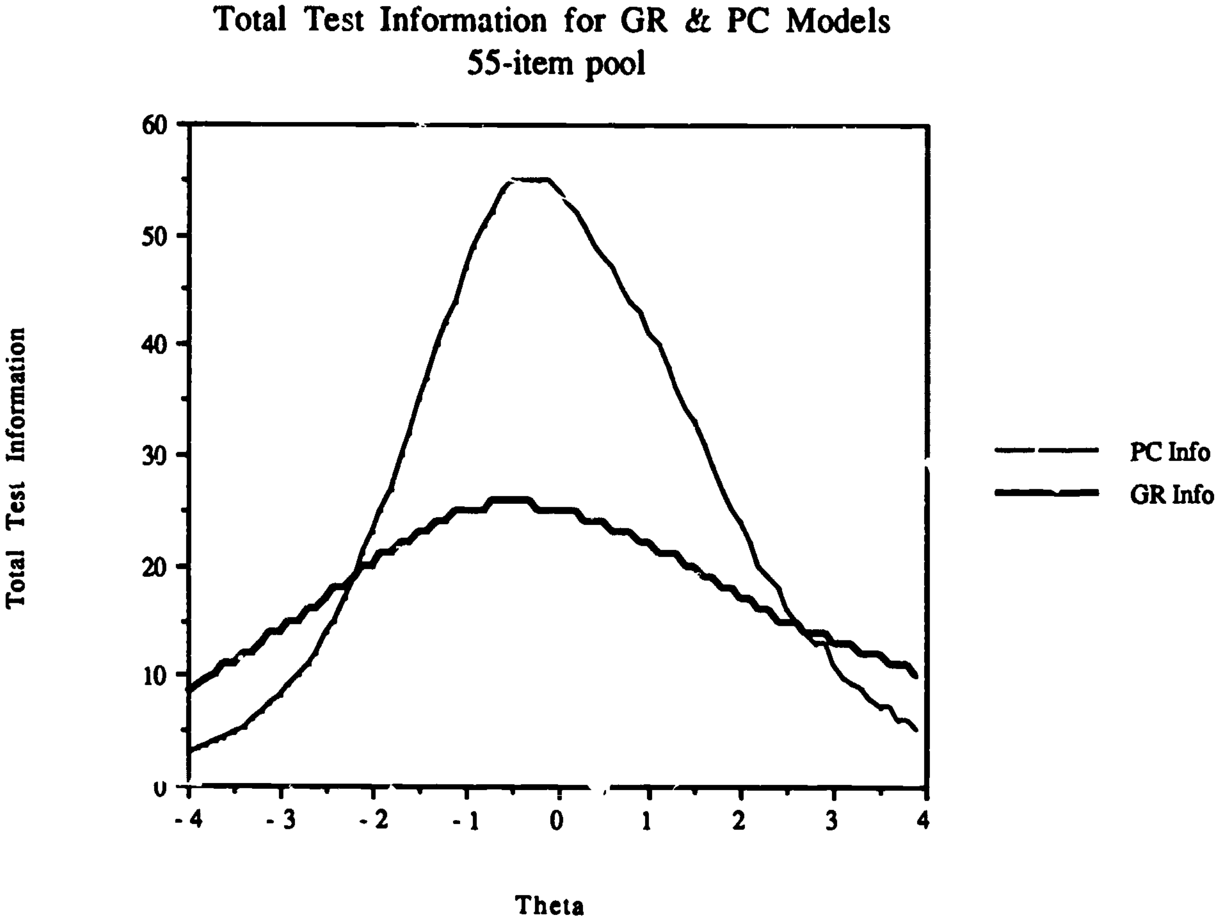

Total Test Information for GR & PC Models
55-item pool
Total Test Information
60
50
40
30
20
10
0
-4
-3
-2
-1
0
1
2
3
4
Theta
PC Info
GR Info

Figure 2

Difference between estimate & true ability
GR-55 CAT, SEE=0.25

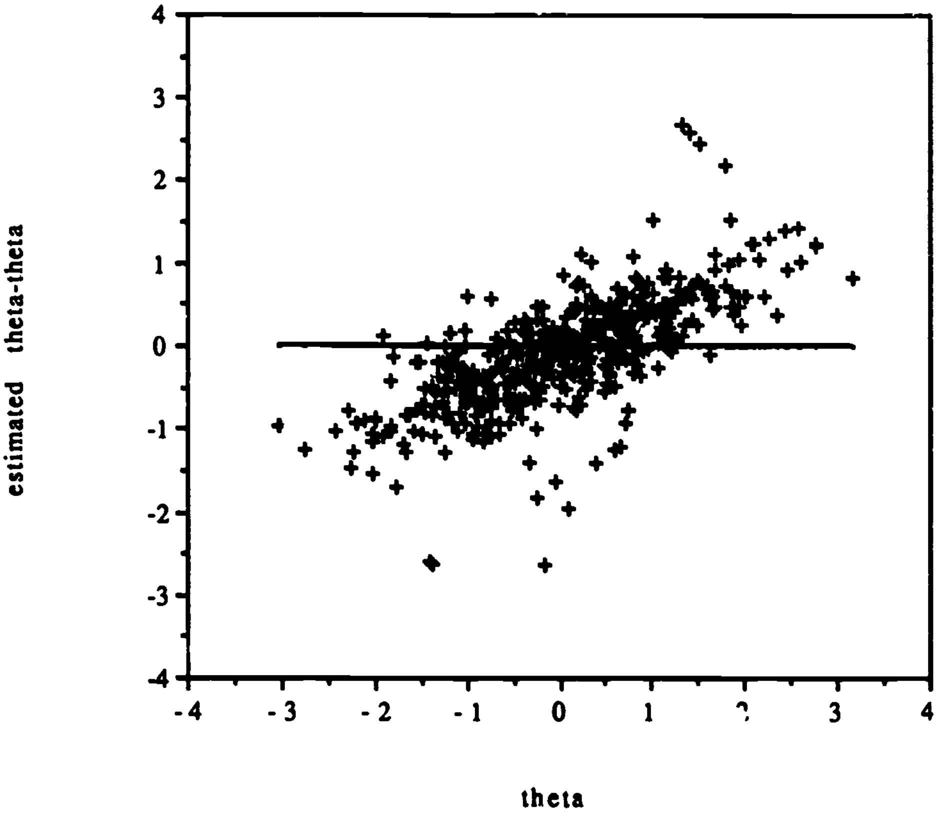


Figure 3

Difference between estimate & true ability
GR-180 CAT, SEE=0.25

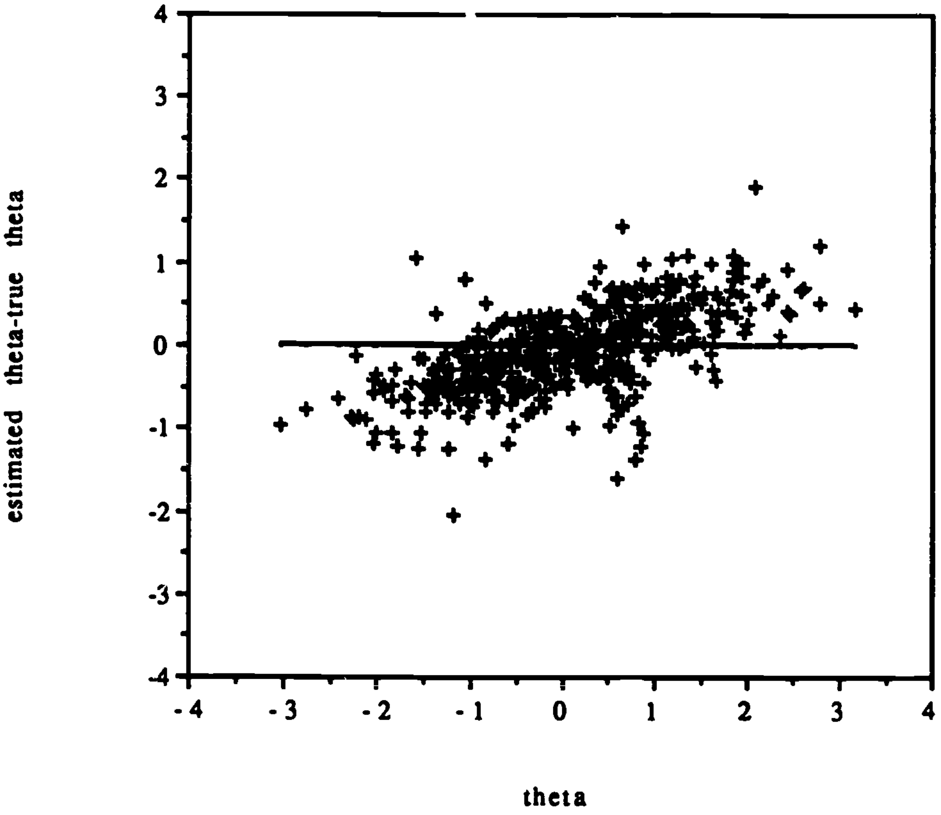


Figure 4

Difference between estimate & true ability
PC CAT, SEE=0.25

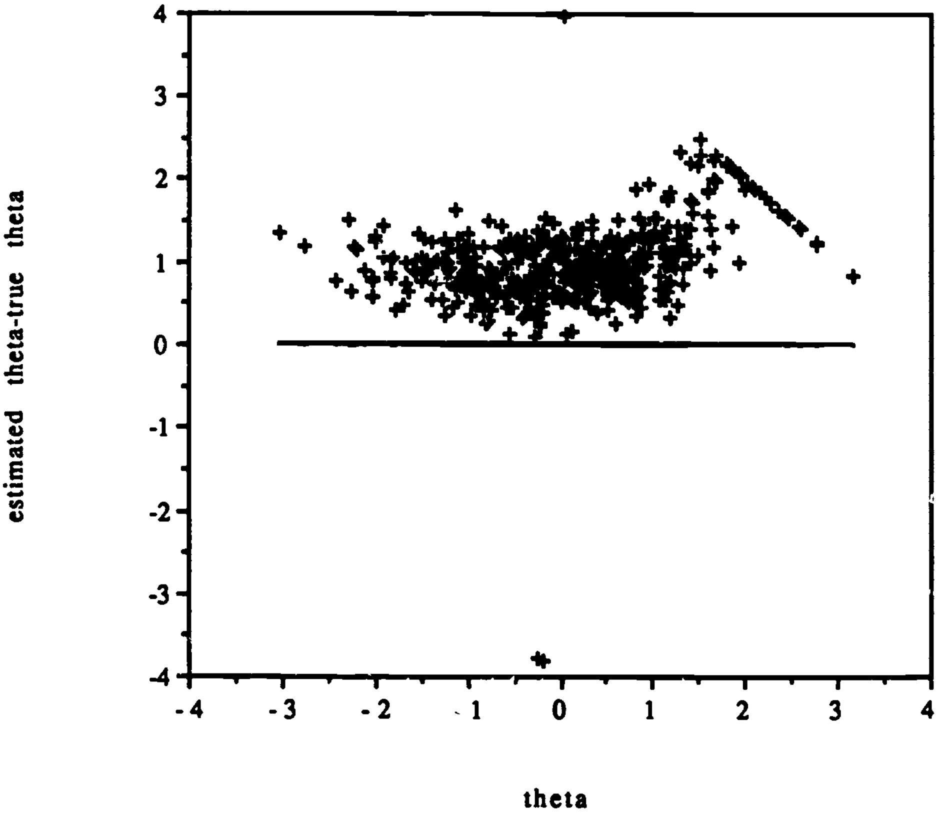
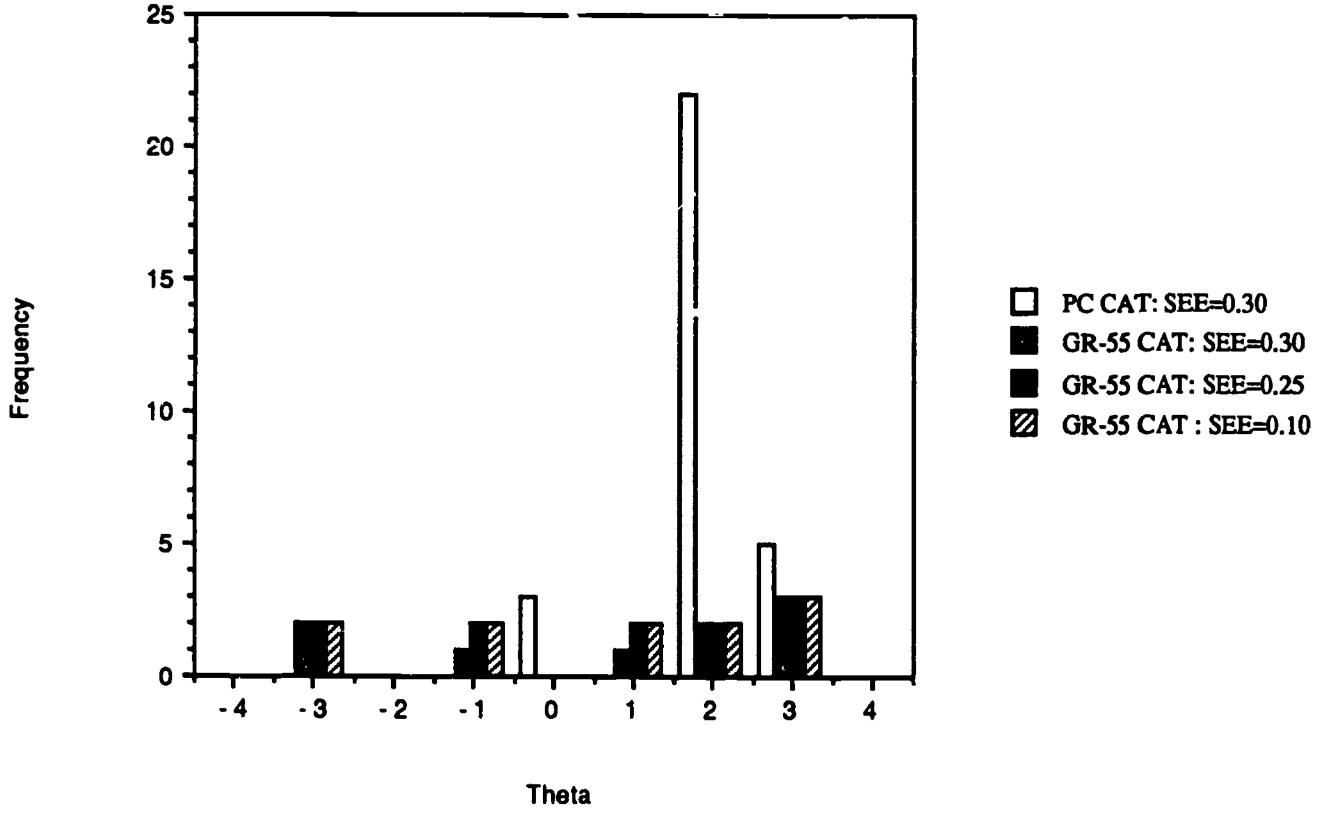


Figure 5

Nonconvergent Cases for PC and GR-55 CATs


Frequency
Theta
□ PC CAT: SEE=0.30
■ GR-55 CAT: SEE=0.30
■ GR-55 CAT: SEE=0.25
▨ GR-55 CAT : SEE=0.10