Operational Characteristics of Adaptive Testing Procedures Using the Graded Response Model

Barbara G. Dodd and William R. Koch
University of Texas

Ralph J. De Ayala
University of Maryland

The purpose of the present research was to develop general guidelines to assist practitioners in setting up operational computerized adaptive testing (CAT) systems based on the graded response model. Simulated data were used to investigate the effects of systematic manipulation of various aspects of the CAT procedures for the model. The effects of three major variables were examined: item pool size, the stepsize used along the trait continuum until maximum likelihood estimation could be calculated, and the stopping rule employed. The findings suggest three guidelines for graded response CAT procedures: (1) item pools with as few as 30 items may be adequate for CAT; (2) the variable-stepsize method is more useful than the fixed-stepsize methods; and (3) the minimum-standard-error stopping rule will yield fewer cases of nonconvergence, administer fewer items, and produce higher correlations of CAT θ estimates with full-scale estimates and the known θs than the minimum-information stopping rule. The implications of these findings for psychological assessment are discussed. *Index terms: computerized adaptive testing, graded response model, item response theory, polychotomous scoring.*

Computerized adaptive testing (CAT) has emerged as one of the important innovations in measurement applications fostered by recent developments in item response theory (IRT). The major advantage of CAT is that persons may be measured very efficiently when the items used to measure them are matched to each individual's ability level. CAT methods have been researched thoroughly for multiple-choice aptitude and achievement testing, and some procedural guidelines and general recommendations for their implementation have been established (Reckase, 1981; Weiss, 1981, 1983, 1985). Also, several major commercial test publishers are currently marketing CAT versions of their standardized tests, all of which use the multiple-choice item format.

One major limitation of these CAT systems, however, is their reliance on dichotomous item response data (this is also a limitation of many IRT applications to date). The use of the multiple-choice format as a standard for test items dictates to some extent what can be measured on tests, and the binary scoring of such items results in a loss of diagnostic information that might otherwise be obtained from incorrect answers.

In contrast to the dichotomous item response data obtained from scoring multiple-choice tests, numerous measurement applications naturally produce polychotomous item response data. For example, in a mathematics problem worth 5 points, 1 point might be awarded for the successful completion of each step in the problem-solving sequence, so that item scores may range from 0 to 5. With items such as these, partial-credit scoring may be used to represent the steps completed by an examinee in solving the problem. Also, responses to attitude instruments whose items use the Likert format are scored into multiple ordered categories. Again, ordered-response scoring is appropriate because the integers assigned in scoring the

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 2, June 1989, pp. 129–143

0146-6216/89/020129-15$2.00

item may be thought of as representing locations along the attitude continuum from negative to positive.

Fortunately, several IRT models have been developed for the analysis of polychotomous item response data. For example, the graded response model for polychotomous item responses proposed by Samejima (1969) offers great potential for a wide variety of measurement applications, including cognitive, personality, and attitude assessment. In the realm of cognitive assessment, Samejima (1969, 1976) successfully applied the graded response model to items where partial credit was awarded for partially correct solutions to the problems. Also, Koch (1983) and Dodd (1985) have demonstrated that the graded response model may be used for measuring attitude trait levels with Likert-type attitude items. It might be feasible to use graded response CAT procedures effectively for these two applications, as well as for personality assessment.

The results of some initial efforts to apply other polychotomous item response models for CAT applications have been quite encouraging. Specifically, Koch and Dodd (1985, in press) used the partial credit model (Masters, 1982) in CAT procedures for attitude measurement using Likert-type items and for simulated achievement test data. Very high correlations were consistently found between the trait estimates yielded by the short adaptive procedures and the corresponding trait estimates obtained from the full-scale administrations of the items. Additionally, Dodd (1987) found the rating scale model (Andrich, 1978a, 1978b) to perform very well for adaptive attitude measurement. De Ayala and Koch (1987) employed the nominal response model (Bock, 1972) in CAT procedures for mathematics achievement testing and again obtained good results.

The reason for attempting to develop procedures for the computerized adaptive administration of polychotomously scored items is the same as the basis for CAT with multiple-choice aptitude or achievement items. Namely, a person's trait level can be measured quite efficiently and accurately with relatively few items because the items are chosen very carefully to be appropriate (individually tailored) for the person. With CAT, the particular set of items administered to an individual depends on the specific responses he/she makes to the items. In theory, each person's trait level may be measured with a different set of items, yet all estimates of persons' trait levels are on the same measurement scale.

Before any general recommendations can be made for CAT using the graded response model, however, substantial basic research still must be conducted. Therefore, the purposes of the present study were to manipulate systematically certain aspects of a graded response CAT procedure and to determine the effects of these manipulations on the operational characteristics of the CAT. Examined in the research were the effects of three major variables: item pool size, the stepsize used along the trait continuum until maximum likelihood estimates could be calculated, and the stopping rule employed. The basic objective in conducting the study was to attempt to develop general guidelines that might assist practitioners in setting up operational CAT systems based on the graded response model.

The Graded Response Model

The graded response model developed by Samejima (1969) is an extension of the two-parameter logistic model for dichotomously scored items to the polychotomous case. For each item, a discrimination parameter and a set of category boundaries are estimated. Samejima developed a two-stage process to obtain the probability that an individual will receive a given category score on item i. In the first stage, the probability that an individual with a certain trait level will receive a given category score or higher on item i is expressed by

$$P^*_{x_i}(\theta) = \frac{\exp[Da_i(\theta - b_{x_i})]}{1 + \exp[Da_i(\theta - b_{x_i})]}, \quad (1)$$

where D is the scaling constant 1.7 which maximizes the similarity of the logistic function to the cumulative normal ogive function,

a_i is the discrimination parameter of item i,

θ is the trait level, and
b_{x_i} is the category boundary associated with a particular category score x_i ($x_i = 1, \ldots, m_i$).

In essence, each use of Equation 1 reduces the polychotomously scored item to a dichotomously scored item, such that the graded responses are classified into two categories—scores lower than x_i and scores equal to or greater than x_i. Figure 1 depicts a set of category characteristic curves obtained from the use of Equation 1 for a hypothetical item with category scores that range from 0 to 4.

The second stage in obtaining the probability that an individual will respond in a given category involves subtracting adjacent category characteristic curves. Samejima defined the probability that an individual will respond in a given category as

$$P_{x_i}(\theta) = P^*_{x_i}(\theta) - P^*_{x_i+1}(\theta) \quad . \tag{2}$$

Equation 2 is the general form for obtaining the operating characteristic curves of an item for the graded response model. In order to use this equation to obtain the probability of responding in the lowest category, the first category characteristic curve is subtracted from 1.0. The probability of responding in the highest category is obtained with Equation 2 by subtracting 0.0 from the last category characteristic curve. The operating characteristic curves obtained with Equation 2 for the hypothetical item used to demonstrate the category characteristic curves are presented in Figure 2.

Samejima also extended Birnbaum's (1968) formulation of information functions to the case where items are polychotomously scored. Item information is defined as

$$I_i(\theta) = \sum_{x_i=0}^{m} \frac{P'_{x_i}(\theta)^2}{P_{x_i}(\theta)} \quad , \tag{3}$$

where $P_{x_i}(\theta)$ is the probability of receiving category score x_i conditional on θ, and $P'_{x_i}(\theta)$ is the first derivative of $P_{x_i}(\theta)$. The information an item provides is a function of the item discrimination parameter and the category boundary parameters. Items with category boundary parameters that span a wide range provide information across a wider range of the θ scale than items with category boundaries that span a small range. The maximum amount of information an item provides is also a function of the discrimination parameter; items with high discrimination power provide more accurate measurement of trait levels in the range of the category boundary parameters than items with low discrimination power. Because some item information functions are quite peaked, some are relatively flat

Figure 1
Category Characteristic Curves for a Graded Response Item

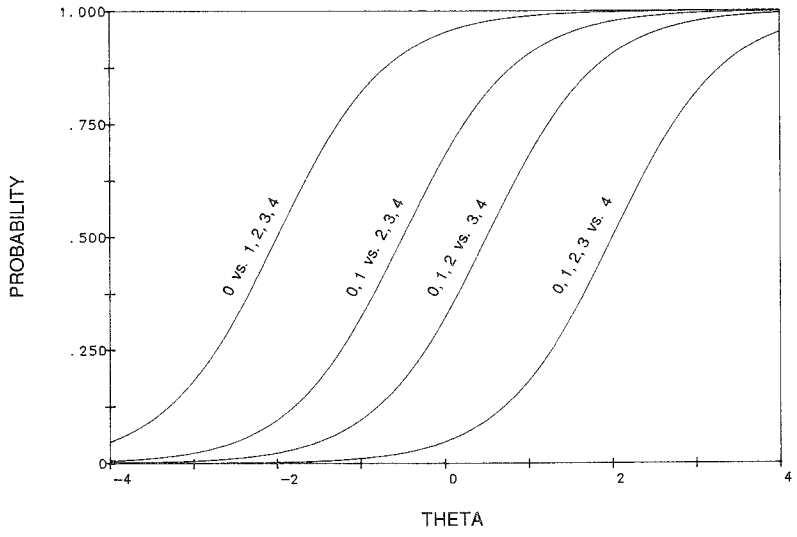
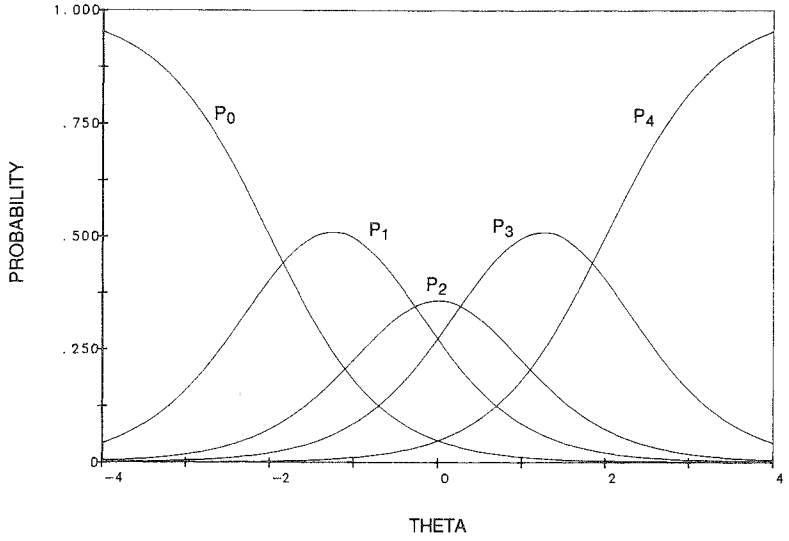


Figure 2
Operating Characteristic Curves for a Graded Response Item



across levels of θ, and some are multimodal, item information functions can be quite useful in CAT systems to decide on the specific items to administer to a particular examinee once an estimate of his/her θ level is available.

Method

Overview of Procedures

The effects of three major variables were studied in the present research: (1) the size of the graded CAT item pool (30 items and 60 items), (2) the stepsize method used prior to obtaining a maximum likelihood θ estimate (fixed—either .4 or .7—and variable), and (3) the stopping rule used to terminate the CAT (prespecified minimum item information or minimum standard error of the θ estimate). The basic procedure followed was to manipulate these variables systematically within the context of a graded response model-based CAT method to evaluate their impact on the operational characteristics of the CAT.

Initially, item pools having prespecified properties were constructed and simulated item response data were generated according to the graded response model. The item response data were then calibrated using the MULTILOG computer program (Thissen, 1986) to obtain parameter estimates for the items and θ estimates for the simulated examinees (simulees). Next, simulated graded response CATs were administered to the simulees. Finally, comparisons were made among the estimated θ levels from the CAT method, the θ estimates from the initial calibrations of the response data, and the known θ levels used to generate the data. The test lengths and the standard errors of the θ estimates obtained under the various CAT conditions were also studied.

Construction of Item Pools

Previous studies by the authors concerning CAT procedures for other polychotomous models have shown that smaller item banks can be used than for CAT procedures using dichotomous items. Specifically, pools with 60 items have been found to be more than adequate for successful CAT procedures with polychotomous models. Thus an item pool size of 60 was used in the present study. In addition, a 30-item pool size was selected for com-

parison purposes as a probable lower limit. The objective was to determine what, if any, deterioration occurred when the 60-item pool was reduced by half.

The first step in developing the datasets was to construct a 30-item pool which was intended to simulate, for example, a set of mathematics word problems worth 4 points each (scored from 0 to 4). In deciding on the values to set for the known item parameters, a deliberate attempt was made to include a wide variety of items whose category boundaries and discrimination values reflected those typically obtained from calibrations of real graded response ability test data. That is, some items had category boundaries spaced fairly tightly together while others had widely spaced category boundaries. Also, some items had all negative category boundaries, some were all positive, and some were roughly symmetric around the center of the θ continuum.

The item discrimination values were specified to range from .90 to 2.15 (in .05 increments) and were randomly assigned to the 30 items. Table 1 shows the complete set of known item parameters for the 30 items. The 60-item pool was created simply by duplicating the 30-item pool. Each of the items was specified to have four category boundaries which formed five response categories; therefore, the possible item scores ranged from 0 to 4.

The objective in setting the category boundary and discrimination values was to make them as realistic as possible based on experience with the values obtained for item parameter estimates from real data. A conscious effort was made, however, to spread the items uniformly across the θ continuum as expressed by their category boundaries. The motivation was to construct item pools with a roughly uniform distribution of item category boundary estimates, which has been found to be ideal in prior research with CAT using multiple-choice items for cognitive testing.

Simulated Data Generation

The data generation procedure began by selecting a z score from a normal (0,1) distribution to represent the θ of the simulee along the θ continuum. Next, the program calculated the probability of the simulee responding in a particular score category or higher for each item based on the known item parameters. Then, using a random number generator for a uniform distribution, a value from 0 to 1 was sampled for each simulee for each item. If the randomly generated value was greater than the probability of responding in the second category or higher according to the model, then the response to that item was designated as being in category 1 (reflecting an item score equal to 0). If the random value was less than or equal to the model probability, it was compared with the model probability for category 3 or higher, and so forth. If the random value was less than the probability for category 5, then the response was designated as being in category 5 (an item score equal to 4).

This data generation procedure was repeated for each of the remaining items in the 60-item pool for that simulee, and began again for the next randomly selected simulee. The resulting response strings to the 60 items for 1,000 simulees were to be used later as input to the MULTILOG parameter estimation program and for the graded response CAT procedure. The simulees' θ values were to serve as known parameters against which the parameter estimates from the calibration and CAT runs could be compared. The purpose in generating the data specifically to fit the graded response model was to study the CAT procedures under ideal conditions. If the CAT methodology performed poorly in such circumstances, there would be little reason to be optimistic about its working with real data. Estimated item parameters were used rather than their known values because item parameters are never known in practice.

Parameter Estimation

The category boundary and discrimination parameter estimates for the items and θs for the persons were obtained from MULTILOG (Thissen, 1986), which was designed to estimate parameters for a variety of polychotomous item response models including the graded response model. A minimum of two computer runs is required to obtain the item

Table 1
Graded Response Model Known Item Parameters Used for Simulated Data Generation

Item Number	b_1	b_2	b_3	b_4	a
1	.50	1.00	1.50	2.00	.90
2	.00	.50	1.00	1.50	1.15
3	−.50	.00	.50	1.00	.40
4	−.75	−.25	.25	.75	1.55
5	−1.00	−.50	.00	.50	1.70
6	−1.50	−1.00	−.50	.00	1.95
7	−2.00	−1.50	−1.00	−.50	.95
8	.00	.50	1.00	2.00	1.20
9	−2.00	−1.00	−.50	.00	1.45
10	.19	.38	.75	1.50	1.55
11	−1.50	−.75	−.38	−.19	1.75
12	−1.00	−.50	.50	1.00	2.00
13	−.70	−.20	.20	.70	1.00
14	−2.00	−1.00	.00	1.00	1.25
15	−1.00	.00	1.00	2.00	1.50
16	−1.50	−.50	.50	1.50	1.60
17	−.50	.00	1.00	1.50	1.80
18	−1.50	−1.00	.00	.25	2.05
19	−1.80	−.90	.90	1.80	1.05
20	−.50	.00	1.00	2.00	1.30
21	−2.00	−1.00	.00	.50	1.50
22	−1.75	−1.25	−.75	−.25	1.60
23	−1.25	−.75	−.25	.25	1.85
24	−.25	.25	.75	1.25	2.10
25	.25	.75	1.25	1.75	1.10
26	−.50	.00	.75	1.50	1.35
27	−1.50	−.75	.00	.50	1.55
28	−2.00	−1.00	−.50	.50	1.65
29	−.50	.50	1.00	2.00	1.90
30	−2.00	−.60	.60	2.00	2.15

and person parameter estimates. In the first run, the item parameters are estimated using the marginal maximum likelihood method (Bock & Aitkin, 1981). If the estimation of some of the item parameters is unsatisfactory, the unacceptable items are deleted and the first run is repeated. Once satisfactory item parameters are obtained, the person parameters are estimated in a separate MULTILOG run using standard maximum likelihood estimation.

MULTILOG was run for the entire matrix of simulated item response data for 1,000 simulees and 60 items. Based on the authors' previous experience with the program and polychotomous models, it was felt that this number of simulees would be sufficiently large to obtain stable estimates of the item parameters. As a check on MULTILOG's accuracy in estimation of known parameters, the estimated θ values were correlated with the known θ values used to generate the data. The resulting correlation coefficient was extremely high (r = .99). Although the correlation between the θ estimates and the known θ values is not a direct assessment of MULTILOG's ability to recover known

item parameters, a high correlation coefficient between them could not have been obtained if the item parameter estimation had been poor.

CAT Simulations

There are three basic components of CAT systems: (1) a procedure to estimate θ, (2) an item selection method, and (3) a stopping rule. The present research used the maximum likelihood method for θ level estimation in conjunction with the maximum information procedure to select appropriate items for administration. Although Bayesian methods for θ estimation and item selection in CAT are well established for the dichotomous case, no research to date has been reported which applies the Bayesian approach to the polychotomous case. Therefore, maximum likelihood estimation was selected for use in the present study.

Two different stopping rules were studied in the CAT simulations. The test was terminated either when no item remained in the pool that had at least a prespecified minimum level of item information given the current estimate of the simulee's θ level, or when the standard error associated with the θ estimate fell below a prespecified value. If neither condition was met after 20 items had been administered, the CAT was terminated.

Depending on the shape of the information function for the item pool, the minimum-information and the minimum-standard-error stopping rules may lead to different results in terms of (1) the number of items administered in the CAT and (2) the accuracy of θ estimation with dichotomously scored items. If the information function for the item pool has a uniform distribution, either rule should yield approximately the same results. However, when the information function is peaked, different results may be obtained (Weiss, 1982). For example, the use of the minimum-standard-error stopping rule may result in the administration of items that are inappropriate for individuals with extreme θ levels in an attempt to reduce the standard error to the specified minimum cutoff. On the other hand, the use of the minimum-information stopping rule is likely to administer a much shorter CAT to individuals with extreme θ levels because relatively few informative items are available to measure such individuals. Thus both stopping rules were included in the present research to investigate their properties in CAT with polychotomously scored items.

The computer program GRCAT was written to simulate a procedure for CAT. First, all of the precalibrated item parameter estimates for the items within a specific pool were stored in a computer file. Next, 200 simulees were randomly selected from the original 1,000 who were used for the full-scale calibrations, because it was thought that a sample size of 200 would be adequate for evaluation purposes. These same 200 simulees were run through each of the 12 CAT conditions (explained in more detail below). For each simulee, the initial θ estimate was set equal to .10, which was about in the middle of the difficulty range of both the 30-item and 60-item pools and at which point the true information of the pools reached a maximum.

Given this initial θ estimate, item information was computed for each item in the pool, and the item with the highest information was selected for presentation. Thus, everyone was administered the same first item (of course, this would be neither necessary nor desirable in practice). Then, the simulee's original response string was checked to determine the actual response that had been made to the item. With maximum likelihood θ estimation procedures in the context of the dichotomous item response models, no estimate of a person's θ level is possible until at least one item is answered correctly and one is answered incorrectly. Similarly, with the polychotomous graded response model, no maximum likelihood θ estimate is possible after the first item if the person receives a score in either the lowest or the highest category. However, a maximum likelihood estimate may be obtained after only one item if the response is scored in one of the middle categories, although the estimate will be unstable and will have a high standard error.

Also, a person might receive scores of 0 or 4 on both the first and second items, again precluding a maximum likelihood θ estimate. In such cases, a systematic procedure must be used to obtain some preliminary θ estimate which, in turn, will lead to

the selection of the next item to be administered. As soon as an item score other than 0 or 4 occurs, maximum likelihood estimation may be performed. In light of the above, the decision was made not to attempt maximum likelihood θ estimation until a simulee received scores in two different categories.

Given this constraint, three different methods were used to obtain a new estimate of the θ level before computing a maximum likelihood θ estimate. These three methods were then compared to determine their impact on the operational characteristics of the CAT. In two of the methods, a fixed stepsize along the θ scale (either .40 or .70) was used after the first item to obtain the next estimate of θ for the simulee. Based on previous experimentation, the CAT algorithm used a rule in which the initial θ estimate was decreased by the fixed stepsize for category scores of 0 or 1 and was increased by the stepsize for category scores of 2, 3, or 4. The other method used a variable-stepsize technique in which the estimate of θ, after the first item, was set halfway on the θ scale between the initial θ estimate and the highest category boundary value for any item in the pool if the score was 2, 3, or 4, or the lowest category boundary value if the score was 0 or 1. The highest category boundary estimate was +3.65 and the lowest was −3.41.

For this new θ estimate, the pool was again searched for the item with maximum information; the most informative item was then presented. As before, the original response string for the simulee was checked for the actual category score on the item. If the scores on the first two items were still the same, the stepsize methods were used again: The fixed-stepsize techniques changed the estimate of θ by plus or minus the stepsize depending on the score, while the variable-stepsize technique again set the new θ estimate at a value halfway between the previous θ estimate and the highest or lowest category boundary value in the item pool.

As soon as two different responses were made, the log likelihood function of the response string was calculated. The point on the θ scale at which the likelihood reached a maximum became the new θ estimate. The procedure was then repeated after each item until one of the stopping rules was encountered. There were two basic conditions for termination of the CAT: (1) to stop when no item remained in the pool (not yet administered) that had an item information value of at least .50 for the current θ estimate, or when a maximum of 20 items had been administered; or (2) to stop when the standard error associated with the current θ estimate dropped below .25, or when a maximum of 20 items had been administered.

Data Analyses

The data analyses consisted primarily of descriptive statistics, scatterplots, correlations, and repeated-measures analyses of variance (ANOVAs). Also, information functions were computed to compare the total information for the 30-item and 60-item pools. Descriptive statistics were calculated to obtain means and standard deviations of the various θ estimates, standard errors of estimate, and test lengths (number of items administered) produced by the various conditions of the CAT procedures. Scatterplots and correlations provided information on the degree to which the various θ estimates were linearly related to each other and to the known θs. Finally, the variables manipulated to study their effects on the CAT procedures comprised a 2 × 3 × 2 design (two item pool sizes, three stepsize techniques, and two stopping rule conditions). Therefore, ANOVAs evaluated the effects of these factors on the resulting CAT θ estimates, associated standard errors of estimate, and test lengths.

Results

Parameter Estimation

MULTILOG was run on the simulated item responses generated from 1,000 simulees to the 60 items in the pool. After 25 cycles of the program, parameter estimates had converged for all of the item category boundaries and item discrimination values. In addition, θ estimates were obtained for all 1,000 simulees. It was not surprising that no cases of nonconvergence of parameter estimation

occurred, because the data were generated deliberately to fit the graded response model. Table 2 contains a listing of the item parameter estimates for all 60 items.

Item Pool Information

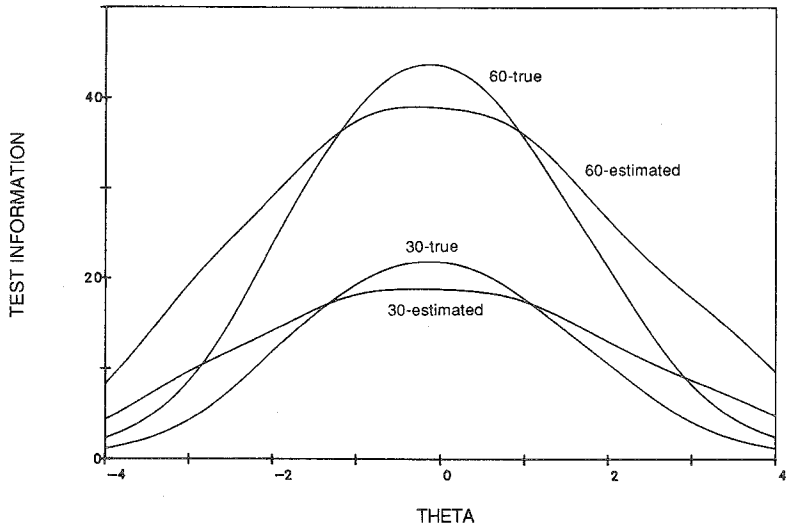
Figure 3 illustrates the total information obtained for the 30-item and 60-item pools. Both the true and the estimated information functions are shown for each pool. It is interesting to note that the estimated item parameters obtained from MULTILOG resulted in flatter information functions than the information plot based on the known item parameters. This result was not due to systematic underestimation of the item discrimination parameters, however. Rather, the negative category boundaries tended to be underestimated whereas the positive category boundaries tended to be overestimated. However, the curves are generally quite similar in shape; this provides some evidence of MULTILOG's capability to recover known item parameters.

Figure 3 shows that the item pool information functions, regardless of size, had roughly the same shape, as would be expected. Despite the attempt to develop item pools having somewhat uniform distributions of information, both pools had fairly peaked information functions. The information was roughly constant in the range from -1 to $+1$ on the θ scale, but it dropped off rapidly outside these limits; relatively little information was available below $\theta = -3$ or above $\theta = +3$. The information functions were symmetric around 0 and reached their maximum at approximately $\theta = -.10$. As was expected from the summative property of in-

Table 2
Graded Response Model Item Parameter Estimates for the 60-Item Pool

Item Number	b_1	b_2	b_3	b_4	a	Item Number	b_1	b_2	b_3	b_4	a
1	.77	1.48	2.32	2.97	1.04	31	.92	1.80	2.48	3.39	.94
2	.17	1.00	1.76	2.65	1.21	32	.19	1.09	1.98	2.82	1.08
3	−.75	.06	.94	1.83	1.43	33	−.73	.12	.94	1.86	1.40
4	−1.23	−.33	.54	1.35	1.61	34	−1.20	−.32	.57	1.42	1.54
5	−1.65	−.84	−.03	.95	1.51	35	−1.63	−.81	.05	.98	1.71
6	−2.67	−1.62	−.77	.13	1.83	36	−2.42	−1.57	−.78	.16	1.91
7	−3.14	−2.37	−1.61	−.73	.92	37	−3.24	−2.54	−1.74	−.85	.96
8	.01	.97	1.79	3.52	1.14	38	.12	.99	1.83	3.57	1.18
9	−3.17	−1.51	−.70	.19	1.43	39	−3.03	−1.50	−.75	.03	1.57
10	.35	.66	1.39	2.81	1.41	40	.33	.78	1.45	2.69	1.52
11	−2.36	−1.20	−.61	−.30	1.79	41	−2.42	−1.20	−.57	−.25	1.79
12	−1.67	−.82	.95	1.81	1.93	42	−1.59	−.79	1.02	1.70	2.04
13	−1.14	−.20	.58	1.54	.83	43	−1.20	−.34	.32	1.14	1.03
14	−3.30	−1.70	.02	1.76	1.28	44	−3.34	−1.61	.12	1.92	1.21
15	−1.73	.06	1.74	3.65	1.39	45	−1.47	.10	1.79	3.46	1.53
16	−2.47	−.83	.95	2.54	1.52	46	−2.42	−.77	1.01	2.63	1.63
17	−.83	.01	1.80	2.60	1.79	47	−.80	.17	1.83	2.55	1.83
18	−2.43	−1.53	.11	.52	2.04	48	−2.36	−1.65	.04	.47	2.16
19	−2.87	−1.44	1.61	3.06	1.08	49	−2.86	−1.49	1.63	2.99	1.03
20	−.81	.16	1.86	3.62	1.20	50	−.80	.01	1.80	3.46	1.30
21	−3.24	−1.60	.22	1.04	1.56	51	−3.41	−1.67	.06	.78	1.49
22	−2.86	−2.02	−1.26	−.31	1.47	52	−2.87	−1.97	−1.12	−.32	1.50
23	−2.08	−1.25	−.24	.56	1.75	53	−2.01	−1.15	−.33	.54	1.90
24	−.37	.53	1.36	2.26	2.06	54	−.28	.57	1.38	2.19	2.15
25	.50	1.30	2.23	3.14	.99	55	.51	1.37	2.20	2.99	1.17
26	−.66	.11	1.18	2.46	1.37	56	−.74	.09	1.31	2.54	1.37
27	−2.45	−1.18	.08	.98	1.53	57	−2.49	−1.22	.19	.96	1.51
28	−3.08	−1.61	−.82	.85	1.71	58	−3.08	−1.58	−.66	.96	1.71
29	−.85	.99	1.77	3.42	1.74	59	−.77	.91	1.70	3.28	1.90
30	−3.30	−1.02	1.12	3.46	2.04	60	−3.06	−.92	1.10	3.45	2.08

Figure 3
True and Estimated Item Pool Information Functions



formation, the total information of the 60-item pool was about twice that of the 30-item pool.

CAT Simulations

Descriptive statistics. The combinations of the levels of the three factors studied produced a total of 12 CAT experimental conditions. Separately for each of these conditions, the same subsample of 200 simulees selected from the original calibration sample was run through the simulated CAT procedures. The assumption was made that 200 simulees would be a sufficiently large number to compare the 12 CAT conditions rather than using the full calibration sample of 1,000. Across the CAT conditions, 12 θ estimates were obtained for each simulee, as well as 12 standard errors of estimate and 12 test lengths. Table 3 presents descriptive statistics for each of these 12 CAT conditions.

Nonconvergence cases. Under one or more of the CAT experimental conditions, nonconvergence of θ estimation occurred for 66 simulees. However, under any one of the CAT conditions, nonconvergence frequencies ranged from 4 to 28. Therefore, complete data across all 12 conditions were available for only 134 out of the 200 simulees. Table 4 summarizes the nonconvergence rates across the 12 CAT conditions. As can be seen in the table, the use of the minimum-standard-error stopping rule reduced the nonconvergence problem substantially.

The occurrence of nonconvergent cases was particularly a problem when using the minimum-item-information stopping rule. For persons with very high or very low known θ levels, the CAT θ estimates moved toward the extremes of the item pools after relatively few items had been administered. Furthermore, few items were available in these extremes that had sufficient information to meet the .50 criterion to be administered; thus the majority of the simulees with extreme known θs were administered only five or six items. Therefore, the CAT terminated prematurely before stable θ estimates were obtained. This result occurred regardless of the stepsize method or the item pool size used.

The nonconvergent cases with known θ levels in the middle of the θ scale received extreme θ estimates due to one of two reasons. First, some of the nonconvergent cases resulted because the simulees responded in the same category to the first four or five items that were administered. Thus the

Table 3
Descriptive Statistics for Graded Response Model CAT Procedures Under 12 Experimental Conditions Defined By Item Pool Size, Stopping Rule, and Stepsize Method (N = 134)

	Stepsize Method		
Dependent Variable	Fixed .40	Fixed .70	Variable
30 Items, Minimum Information			
Est. θ			
Mean	.359	.350	.343
SD	1.374	1.305	1.391
S.E.			
Mean	.246	.242	.246
SD	.063	.054	.059
No. Items			
Mean	17.418	17.522	17.254
SD	3.293	3.188	3.426
30 Items, Minimum Standard Error			
Est. θ			
Mean	.315	.302	.324
SD	1.207	1.211	1.208
S.E.			
Mean	.246	.246	.246
SD	.006	.006	.006
No. Items			
Mean	15.552	15.500	15.463
SD	1.417	1.353	1.402
60 Items, Minimum Information			
Est. θ			
Mean	.316	.297	.314
SD	1.232	1.214	1.251
S.E.			
Mean	.210	.209	.210
SD	.016	.015	.020
No. Items			
Mean	19.836	19.784	19.724
SD	.806	.984	1.312
60 Items, Minimum Standard Error			
Est. θ			
Mean	.290	.300	.286
SD	1.208	1.201	1.226
S.E.			
Mean	.244	.245	.245
SD	.005	.004	.003
No. Items			
Mean	14.515	14.321	14.284
SD	1.296	1.205	1.128

use of the stepsize methods led to extreme θ estimates for which there were few informative items available to administer. In effect the θ estimates stepped out of the range of the item pool. Second, in some cases the likelihood function was flat due to inconsistent responses to the items administered.

Under the minimum-standard-error stopping rule, the few nonconvergent cases that occurred involved simulees who had extreme known θ levels. In each of these cases the length of the CAT was 20 items. Thus all of the nonconvergent cases that occurred under the minimum-standard-error stopping rule were the result of allowing only 20 items to be administered. If the CAT were not arbitrarily stopped after a maximum of 20 items, perhaps the administration of additional items would have resulted in convergent estimates for these few cases. Collectively, these results showed that nonconvergence of maximum likelihood estimation was a substantial problem when using the minimum-information stopping rule, but not when using the minimum-standard-error stopping rule.

Intercorrelations of θs. All of the intercorrelations among the θ estimates from the 12 CAT conditions, the MULTILOG calibrations, and the known θs were very high, and the scatterplots revealed relationships that were essentially linear. None of the correlations was lower than $r = .90$, and the great majority were $r = .95$ or higher. The correlations of the estimated θs from all 12 CAT conditions with the known θs ranged from $r = .91$ to $r = .98$. Of these, the correlations were lowest for the 30-item pool using the minimum-information stopping rule. All correlations were based on the 134 simulees for whom complete data were available.

ANOVAs. Three separate ANOVAs were run to examine any mean differences that resulted among the estimated θs, standard errors of estimate, and number of items administered under the 12 CAT conditions. Each ANOVA was a $2 \times 3 \times 2$ repeated-measures design because the same 134 simulees were tested under each of the 12 CAT conditions.

The ANOVAs showed no statistically significant main effects or interactions for the estimated θs. The means of the θ estimates observed under all 12 CAT conditions were essentially the same.

The analyses for the standard-error variable revealed significant main effects for item pool size

Table 4
Frequencies of Nonconvergence Cases Under 12 CAT Conditions ($N = 200$)

Number of Items and Stopping Rule	Stepsize Method: Fixed .40	Fixed .70	Variable
30 Items, Minimum Information	22	28	16
30 Items, Minimum Standard Error	6	5	5
60 Items, Minimum Information	13	14	26
60 Items, Minimum Standard Error	5	4	4

and for type of stopping rule used ($p < .0001$). Also, there was a significant two-way interaction found between item pool size and stopping rule ($p < .0001$), thus precluding discussion of the main effects. Simple main-effects analysis showed that the mean standard error of estimate was significantly lower ($p < .0001$) for the 60-item pool when the information-cutoff stopping rule was employed; however, there were no differences among the remaining three conditions. The cell means for this interaction are presented in Table 5.

Based on the number-of-items-administered dependent variable, the results revealed significant main effects for item pool size ($p < .0001$), stepsize method ($p < .001$), and type of stopping rule ($p < .0001$). However, these main effects will not be discussed because two-way interactions were found between item pool size and stopping rule ($p < .0001$) and between stepsize method and stopping rule ($p < .05$). Simple main-effects analysis of the item pool size and stopping rule interaction showed that all four cell means differed from each other. As Table 5 shows, the mean number of items administered was highest for the 60-item pool using the minimum-information stopping rule and lowest for the 60-item pool using the minimum-standard-error stopping rule. Regarding the stepsize method × stopping rule interaction, simple main-effects analyses indicated that the mean number of items administered using the minimum-information stopping rule was greater than that for the minimum-standard-error stopping rule in each of the three stepsize conditions, as shown in Table 6. No other mean differences were significant.

Table 5
Mean Standard Errors for CAT Trait Estimates and Mean Number of Items Administered for Minimum-Information and Minimum-Standard-Error Stopping Rule, by Item Pool Size ($N = 134$)

Stopping Rule	Item Pool Size: 30	60
Mean Standard Error		
Minimum Information	.24	.21
Minimum Standard Error	.25	.24
Mean Number of Items		
Minimum Information	17.40	19.78
Minimum Standard Error	15.51	14.37

In considering the significant mean differences described above, it is important to note that the very large number of degrees of freedom for the error terms in the ANOVA F tests (133 or 266, depending on the particular F test) provided a great deal of power. Thus, relatively small mean differences were still found to be statistically significant.

Finally, the mean number of items administered was smaller for the 30-item pool than for the 60-item pool and was smaller for the standard-error stopping rule than for the information-cutoff stopping rule.

Discussion

In general, the graded response CAT procedure performed reasonably well except for the substantial nonconvergence problems encountered under certain conditions. That is, the frequency of nonconvergence of θ estimation was particularly high when the minimum-information stopping rule was used. The major problem was that at the extremes

Table 6
Mean Number of Items Administered During CAT: Stopping Rule by Stepsize Method (N = 134)

	Stepsize Method		
Stopping Rule	Fixed .40	Fixed .70	Variable
Minimum Information	18.63	18.65	18.49
Minimum Standard Error	15.03	14.91	14.87

of the θ scale, relatively few items were available to be administered that had sufficiently high amounts of information to meet the .50 cutoff. Thus the CAT procedures terminated prematurely before stable θ estimates were obtained. The use of the standard-error stopping rule virtually eliminated nonconvergence problems.

Another approach to deal with nonconvergence of θ estimation would be to employ a Bayesian estimation procedure rather than maximum likelihood estimation. However, in CAT research with dichotomous item responses, it is well known that biased θ estimates (toward the mean of the prior distribution) are obtained under Bayesian estimation procedures. Alternatively, some researchers (Weiss, 1982) have suggested a strategy in which Bayesian estimation is used for the first several items of the CAT until maximum likelihood estimation is possible, in which case the procedure switches to maximum likelihood estimation for the remainder of the CAT. Future CAT research should investigate the utility of these approaches in dealing with the nonconvergence problems for the polychotomous case.

Nonconvergence problems aside, however, the graded response CAT procedures performed well based on the simulees for whom θ estimates were obtained. Under all 12 conditions, the correspondence was very high among the θs estimated from the adaptive procedures, the MULTILOG-calibrated θs, and the known θs. These results were particularly impressive in light of the relatively small item pool (only 30 items) used in six of the CAT conditions. The results also did not differ when the item pool size was doubled.

It has usually been recommended that item pools should consist of at least 100 items for CAT procedures using the three-parameter logistic model for dichotomous item responses (Urry, 1977). McKinley and Reckase (1983), however, obtained fairly good results using the simple Rasch model with a pool of only 40 mathematics achievement items. The main issue seems to be whether the item pools have adequate numbers of items whose location parameters are equally distributed across the entire range of the θ scale used. In such conditions, there will be sufficient information available to administer informative items to examinees with relatively high or low θ levels.

The results of the present research with the graded response model suggest that it may be possible to implement CAT successfully using item pools that are substantially smaller than the pools required for dichotomous items. The apparent reason is that polychotomous scoring of items provides more information across the full range of the θ scale, which reduces the possibility that gaps will occur in the pool.

Regarding the effects of item pool size on CAT performance, it was not surprising to find that the 60-item pool produced smaller standard errors than the 30-item pool when using the minimum-information stopping rule. With so many items in the pool, numerous appropriate and informative items were available for administration across the levels of θ. The results also showed, however, that CAT procedures based on the graded response model still performed well with the 30-item pool; very little improvement was achieved through doubling the pool to 60 items. Because a cutoff value of .250 was used under the minimum-standard-error stopping rule, the means of the standard errors of θ estimates were just below the cutoff value for both the 30-item and 60-item pools. On the aver-

age, however, one additional item was administered with the 30-item pool to achieve the same standard error level as the 60-item pool.

The results of the present research indicated that the type of stepsize method, whether fixed or variable, makes little difference in CAT applications with the graded response model. Regardless of the stepsize method used, the standard errors were about the same, as were the number of items administered and the rates of nonconvergence. However, stepsize method might be an issue depending on the specific characteristics of the item pool being used.

The present findings suggest several guidelines for graded response CAT procedures. First, item pools with as few as 30 items will probably be adequate, although item pools with a uniform distribution of information will perform better (and have fewer nonconvergence problems) than peaked-information item pools. Second, the variable-stepsize method, being more flexible and more generally applicable than the fixed-stepsize method, is the method of choice. Third, the minimum-standard-error stopping rule will outperform the minimum-information stopping rule in terms of the mean number of items administered, frequencies of nonconvergence, and correlations of CAT θ estimates with MULTILOG full-scale θ estimates and the known θs.

The graded response model offers considerable potential for CAT applications. Previous research has shown that other polychotomous item response models perform well in CAT procedures under a variety of conditions using both simulated and real data. The present study extends that research to the graded response model.

Perhaps the most interesting and important result of the present research for practical applications is the finding that substantially smaller item pools can be used successfully for polychotomous CAT compared to those required for dichotomous CAT. This result offers the possibility of implementing CAT versions of attitude scales which typically consist of 30 or fewer items. Moreover, the development of innovative multiple-response-category item formats for measuring cognitive abilities may lead to polychotomous CATs which could be implemented with relatively small item pools.

Practitioners as well as researchers have expressed concerns to the authors at professional meetings that adaptive testing procedures based on simple models for dichotomously scored items are already complicated, and that more complex item response models will further complicate the CAT procedures. The evidence from the present study suggests that such pessimism is unjustified. The present results should encourage more researchers to try out item response models for polychotomously scored items in the context of adaptive testing.

A CAT system based on the graded response model could prove particularly useful in academic areas such as mathematics, physics, chemistry, and engineering because tests in these areas typically consist of word problems which are scored in a polychotomous fashion to reflect partially correct solutions to the problems. Attitude scales, personality instruments, and interest inventories consisting of items that are polychotomously scored to represent varying degrees of the trait measured by the instrument could also be administered adaptively with the procedures outlined in the present research. The practitioner using adaptive testing in these areas should be able to assess an individual's trait level quite efficiently and accurately with relatively few items.

References

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581–594.

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

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 abilities when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

De Ayala, R. J., & Koch, W. R. (1987, April). *Computerized adaptive testing: A comparison of the nominal response model and the three-parameter logistic*

model. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington DC.

Dodd, B. G. (1985). Attitude scaling: A comparison of the graded response and partial credit latent trait models (Doctoral dissertation, University of Texas at Austin, 1984). *Dissertation Abstracts International, 45,* 2074A.

Dodd, B. G. (1987, April). *Computerized adaptive testing with the rating scale model*. Paper presented at the Fourth International Objective Measurement Workshop, Chicago.

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7,* 15–32.

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

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

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

McKinley, R. L., & Reckase, M. D. (1983). *An evaluation of one- and three-parameter logistic tailored testing procedures for use with small item pools* (Research Report ONR83-1). Iowa City IA: American College Testing Program.

Reckase, M. D. (1981). *Final report: Procedures for criterion referenced tailored testing*. Columbia MO: University of Missouri.

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

Samejima, F. (1976). Graded response model of the latent trait theory and tailored testing. In C. K. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing*. Washington DC: U.S. Government Printing Office.

Thissen, D. (1986). *MULTILOG Version 5 user's guide*. Mooresville IN: Scientific Software, Inc.

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

Weiss, D. J. (1981). *Final report: Computerized adaptive ability testing*. Minneapolis: University of Minnesota.

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

Weiss, D. J. (1983). *Final report: Computer-based measurement of intellectual capabilities*. Minneapolis: University of Minnesota.

Weiss, D. J. (1985). *Final report: Computerized adaptive measurement of achievement and ability*. Minneapolis: University of Minnesota.

Author's Address

Send requests for reprints or further information to Barbara G. Dodd, Measurement and Evaluation Center, University of Texas, Austin TX 78713, U.S.A.