The Effect of Item Selection Procedure and Stepsize on Computerized Adaptive Attitude Measurement Using the Rating Scale Model

Barbara G. Dodd
University of Texas at Austin

Real and simulated datasets were used to investigate the effects of the systematic variation of two major variables on the operating characteristics of computerized adaptive testing (CAT) applied to instruments consisting of polychotomously scored rating scale items. The two variables studied were the item selection procedure and the stepsize method used until maximum likelihood trait estimates could be calculated. The findings suggested that (1) item pools that consist of as few as 25 items may be adequate for CAT; (2) the variable stepsize method of preliminary trait estimation produced fewer cases of nonconvergence than the use of a fixed stepsize procedure; and (3) the scale value item selection procedure used in conjunction with a minimum standard error stopping rule outperformed the information item selection technique used in conjunction with a minimum information stopping rule in terms of the frequencies of nonconvergent cases, the number of items administered, and the correlations of CAT θ estimates with full scale estimates and known θ values. The implications of these findings for implementing CAT with rating scale items are discussed. *Index terms: adaptive testing, attitude measurement, computerized adaptive testing, item response theory, rating scale model.*

Methods for computerized adaptive testing (CAT) situations that involve the dichotomous scoring of item responses (correct or incorrect) have been studied fairly extensively (Reckase, 1981; Weiss, 1981, 1983, 1985). In fact, several test publishers now offer CAT versions of some of their tests. These applications, however, have typically been restricted to either aptitude or achievement tests in which responses to each test item are scored in a dichotomous fashion. There exist, however, several measurement situations in which responses to items are scored using more than two categories. For instance, partial credit is typically awarded for a partially correct solution to a problem worth several points, and responses to attitude and personality items are usually scored using more than two categories.

Fortunately, several item response theory (IRT) models have been developed specifically for the case where item responses are scored using more than two categories, and a few studies have investigated CAT procedures for some of these models. Two models, the graded response model (Samejima, 1969) and the partial credit model (Masters, 1982), are appropriate when responses to an item can be scored using more than two ordered categories to represent varying degrees of the trait measured by the item. The operating characteristics of CAT procedures for the graded response model were studied by Dodd, Koch, and DeAyala (1989), and procedural guidelines for CAT for the partial credit model were investigated by Koch and Dodd (1989). Both of these studies produced promising results because a high degree of correspondence was found between the trait estimates yielded by the CAT procedures and the full scale calibration trait estimates, as well as between the trait estimates yielded by the CAT procedures and the known trait levels used to generate the data.

CAT versions of Likert-type attitude scales could be implemented with the procedures that have been recommended for the graded response or partial credit models. However, a simpler IRT

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 14, No. 4, December 1990, pp. 355-366

0146-6216/90/040355-12$1.85

model—the rating scale model—was developed specifically for Likert-type attitude scales. This model, which was proposed by Andrich (1978a, 1978b), has properties that can yield simplified CAT procedures, compared to those required for the graded response or partial credit models. Therefore, the purpose of the present investigation was to study the effects of various item selection procedures and stepsize methods that could prove useful for CAT based on the rating scale model.

The Rating Scale Model

Andrich (1978a, 1978b) extended the Rasch model for dichotomously scored items to the polychotomous case of rating scale items in which responses to an item are scored using ordered categories to represent varying degrees of the attitude level. In the rating scale model, a scale value is estimated for each item to reflect the location of the item on the attitude continuum. In addition, a single set of response thresholds is estimated for the entire set of items included in the rating scale. The response threshold values are assumed to be constant across items on a given rating scale because the same response scale is used to respond to all items on the rating scale. The rating scale model also has been shown to be a special case of the partial credit model (Wright & Masters, 1982). The probability of responding in a given category is defined as

$$P_{x_i}(\theta) = \frac{\exp\left\{\sum_{j=0}^{x_i}[\theta - (b_i + t_j)]\right\}}{\sum_{k=0}^{m_i}\exp\left\{\sum_{j=0}^{k}[\theta - (b_i + t_j)]\right\}} . \qquad (1)$$

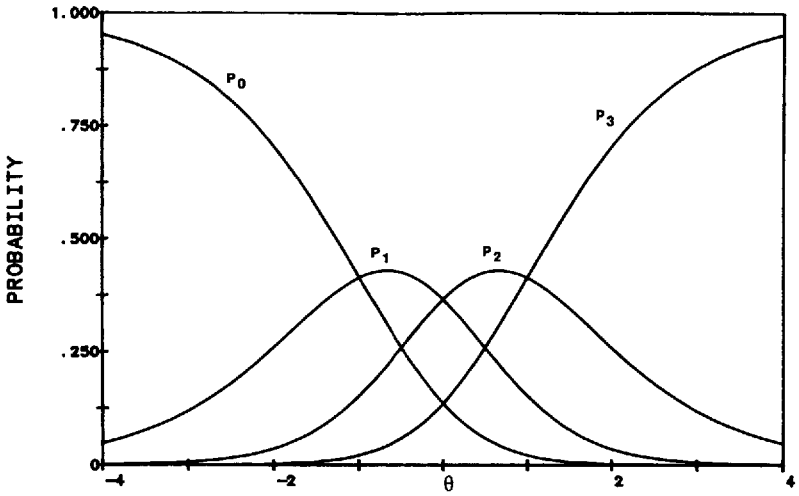
Equation 1 is the general form for obtaining the operating characteristic curves (OCCs) for an item based on the rating scale model. θ is the attitude level, b_i is the scale value or location parameter for item i, and the t_j terms are the response threshold parameters for the set of items. For notational convenience, $\sum[\theta - (b_i + t_j)]$ for $j = 0$ to 0 is defined as being equal to 0. Figure 1 depicts the OCCs for a four-category hypothetical item with a scale value of 0 and response threshold values of -1, 0, and 1.

Dodd (1987) applied Samejima's (1969) formulation of information functions for polychotomously scored items to the rating scale model. The results showed that the distribution of item information for a set of items with the same response threshold values was a function of the scale values for the items. Each item information function peaked near the scale value for the item. Rating scales with threshold values that spanned a small range along the attitude continuum also yielded more peaked item information functions than rating scales with threshold values that spanned a large range. Thus, the distribution of item information was a function of both the scale value for the item and the set of response threshold values for the rating scale. These properties of item information could prove useful in item selection procedures during CAT administrations using the rating scale model. Items could be selected for administration on the basis of either maximum information or closest scale value to the latest θ estimate. The two item selection procedures might select different items to administer because the item information function does not always peak at exactly the scale value of the item.

Method

This study investigated the impact of the systematic variation of two major variables on the operating characteristics of the rating scale CAT: (1) the item selection procedure (maximum information or closest scale value), and (2) the stepsize method (variable or fixed, .4 or .7). Initially, the item responses for the three datasets were calibrated separately according to the rating scale model. Next, simulated CATs that were based on the rating scale model were administered to the calibration samples. The analyses included comparisons of the θ estimates and the standard errors yielded by the six adaptive testing conditions with the full scale calibrations. The number of items administered under the six CAT conditions was also investigated.

Figure 1
OCCs for a Four-Category Rating Scale Item



Datasets

Three datasets were used. Two datasets consisted of real response data for two different attitude scales; the third dataset consisted of simulated response data generated specifically to fit the rating scale model.

Audit of Administrator Communication. Responses made by 491 teachers to the Audit of Administrator Communication (ADCOM; Valentine, 1978) were available for use in the present study. ADCOM is a 40-item Likert-type attitude scale designed to measure attitudes of teachers toward the communication skills of their school administrators. All items are scored on a five-point scale on which 0 represents an unfavorable response toward the communication skills of the administrator, and a score of 4 represents a favorable response. Factor analysis of the ADCOM scale (Koch, 1983) indicated that the scale is approximately unidimensional; the first factor accounted for about 85% of the common variance.

Attitude Toward Women Scale. The Attitude Toward Women Scale (AWS; Spence, Helmreich, & Stapp, 1973) was designed to measure attitudes toward the rights and roles of women in contemporary society. Each of the 25 items has four response alternatives ranging from "AGREE STRONGLY" to "DISAGREE STRONGLY." Responses are scored so that profeminist attitudes receive a score of 3, whereas very traditional attitudes receive a score of 0. Response data were available for 533 women. Previous factor analytic studies (Dodd, 1985) demonstrated that the AWS has one dominant factor that accounts for about 83% of the common variance.

Artificial data. The third dataset consisted of simulated responses to 32 items from 500 simulees. These data were generated according to the rating scale model using standard procedures. The items were constructed to have four response alternatives. Consequently, three response threshold values were specified for the set of 32 items, and a scale value was specified for each of the items. The item parameters used to generate the data were those estimates reported by Andrich (1978b) based on real data. More specifically, the item parameter estimates for 16 items that Andrich found to fit the rating scale model were treated as known item parameters and were used as input into the data generation program. Given the fact that 16 items is probably too small an item bank for CAT, the size of the item pool was doubled by duplicating Andrich's item parameter estimates for the 16 items, and simulated item

responses were thus generated for 32 items.

The data generation procedures began by selecting a z score from a normal distribution (0,1) to represent the simulee's θ level. Next, the program calculated the probability of the simulee responding in each of the score categories for the first item. These probabilities were then summed to obtain a subtotal probability for each category, as well as a total probability for the item. Each subtotal probability served as a boundary between the adjacent categories. For example, the sum of the probabilities of selecting categories 0, 1, and 2 served as the lower boundary for selecting category 3. Once the boundaries were calculated, a value was drawn from a uniform distribution ranging from 0 to 1, and the simulee's response for the item was determined by where the random value fell relative to the calculated boundaries.

The data generation procedure was repeated for each of the remaining items until responses had been generated for all 32 items for the first simulee. The entire procedure was then repeated for the next randomly selected simulee. The resulting response strings to 32 items for 500 simulees were stored for later use in the simulated adaptive measurement procedures. Because these data were generated according to the rating scale model, there was no need to assess the unidimensionality of the data.

Parameter Estimation

A two-stage process outlined by Wright and Masters (1982) was used to obtain estimates of the item parameters according to the rating scale model for each of the three datasets.

The first stage involved the calibration of each dataset according to the partial credit model with the PARTIAL computer program, which implemented the calibration procedures specified by Masters (1982) for the partial credit model. PARTIAL used maximum likelihood estimation (MLE) and a Newton-Raphson iteration procedure to obtain parameter estimates. Iteration cycles were continued until all item parameter estimates had converged, or until a specified maximum number of iterations had been performed. The PARTIAL program set the origin of the θ scale to 0 and the unit of measurement to 1.

The second stage involved obtaining estimates of the response threshold parameters and of the scale value parameters from the step value estimates obtained from the PARTIAL program. The estimate of the scale value for an item was simply the average of the partial credit model's step value estimates for the item. The response threshold estimates were obtained by first transforming each of the partial credit step value estimates for an item into a deviation score from the scale value for that item. The deviation scores for each step were then averaged across the items to obtain the estimate of the response threshold parameter for each step.

Information Analyses

The information function was computed for each item of the three datasets, based on the item parameter estimates obtained with the two-stage process outlined above. The equation specified by Samejima (1969) was used to calculate item information for θ values (attitude levels) ranging from −4 to 4 in .1 increments. Total information was also determined by summing the item information functions for each of the item banks used in the adaptive procedures.

CAT Procedures

There are three basic components of CAT systems: a procedure to estimate θ, an item selection method, and a stopping rule. When items are scored dichotomously, the choice of an estimation procedure usually involves the selection of a Bayesian or maximum likelihood procedure. Such a choice does not currently exist, however, for polychotomous IRT models because Bayesian procedures for such models have not yet been developed. Thus, a standard maximum likelihood method of estimation was employed in the present research. Although it is possible to use MLE to estimate θ after the administration of a single item when the response to the first item

is not in either of the extreme categories, previous research has indicated that the θ estimate based on one item is very unstable (Dodd et al., 1989; Koch & Dodd, 1989). Therefore, MLE was not used until at least two different category responses had been obtained.

Prior to MLE, it is common procedure to use some sort of stepping rule to change the θ estimate. This can be based on the last item response to obtain a new θ estimate, which then is used to select the next item to be administered. A variable stepsize and two fixed stepsizes (.4 and .7) were studied here. The variable stepsize changed the θ estimate by half the distance to the appropriate extreme scale value in the item pool. More specifically, if an individual response to the last item administered was in the upper half of the range of possible category values, the θ estimate was incremented by half the distance to the highest scale value for the item pool. If the response was in one of the lower categories, the θ estimate was decreased by half the distance to the lowest scale value in the item pool. For the fixed stepsizes, the θ estimate was similarly increased or decreased by the specified stepsize, depending on the item response. The values selected for the fixed stepsize conditions were values that have been shown to be useful for CAT systems with the 1PL and 3PL models (Patience & Reckase, 1979; Patience & Reckase, 1980).

Two different item selection methods were investigated. One procedure, which has been recommended for CAT systems using either the partial credit or the graded response model, involves selecting the item that provides the most information for the last θ estimate (Dodd et al., 1989; Koch & Dodd, 1989). The second procedure involves selecting the item with the scale value estimate that is closest to the last θ estimate. Given the fact that the item information function does not peak at exactly the scale value for the item, it was not clear whether these two item selection procedures would result in the same item being selected for the current θ estimate. Even if the two procedures were found to select the same item, the scale value selection method would be preferred because the computations would be easier and faster than for the item information method.

Given these two item selection procedures, the stopping rule was combined with the item selection method. It is not uncommon to use a minimum information stopping rule when items are selected on the basis of maximum information—that is, the CAT is terminated when there are no more items left in the item pool for the latest θ estimate that provide at least a prespecified, minimum amount of information (Patience & Reckase, 1979, 1980). The minimum information that was specified for each dataset was a function of the average item information across the majority of the θ scale for the item bank. For all datasets, the average item information for θ values in the range of −3 to 3 was found to be .44. Thus, the CAT was terminated when no item was left in the item pool that provided at least .44 information for the latest θ estimate.

Specifying a minimum standard error stopping rule was used for the scale value item selection procedure because there would then be no need to calculate information during the adaptive test. The well-known relationship between the total information and the standard error of the θ estimate was used to determine the specific standard error value to prevent the results of the CATs from being biased in favor of either procedure. The CAT session was terminated, therefore, when the standard error of the latest θ estimate reached a level of .3. If the CAT was not terminated by either stopping rule after 20 items had been administered, the CAT was terminated because research with other polychotomous IRT CAT systems has indicated that 20 items is usually a sufficient number of items to obtain an accurate θ estimate (Dodd et al., 1989; Koch & Dodd, 1989).

A computer program was written to simulate the six different CAT procedures (two item selection techniques × three stepsize methods). For each dataset, the precalibrated item parameter estimates were stored in a computer file, and the item response strings used to calibrate the items were stored in another computer file for use

during each of the CAT simulations. For each simulee, the initial θ estimate was set at the θ level where the item pool information reached a maximum. This approach to selecting the initial θ estimate is equivalent to setting the first θ value near the middle scale value of the items in the pool.

Depending on the item selection technique employed, the first item selected for presentation was either the item with the highest information for the initial θ estimate, or the item with the closest scale value to that θ estimate. Once the item had been selected, the computer file that contained the actual person's responses to the paper-and-pencil version of the instrument was examined to determine which category the person had selected in response to that item. If the individual responded in one of the highest categories, the θ estimate was increased by the designated stepsize value. If the person had responded in one of the lower categories, the θ estimate was decreased by the specified stepsize.

Given the new θ estimate, the item pool was then searched for the next item to administer that had not been given previously. If the response to the selected item was in the same category as the response to the first item, the stepsize method was used again to estimate a new θ. If the responses to the first two items were different, MLE was used to estimate the current θ based on the responses to the items that had already been administered during the CAT. MLE was then used until the CAT was terminated by the specified stopping rule for that CAT simulation or until a maximum of 20 items had been administered.

Data Analyses

Within each dataset, values were available for each person's θ estimate from the full scale calibration, as well as from the six adaptive testing procedures. For the artificial dataset, the z scores from the data generation procedure were also available. Scattergrams and correlations were obtained to examine the relationship among these variables. In addition, various descriptive statistics were calculated for the standard errors associated with the θ estimates and the number of items that were administered under each of the six CAT procedures.

Results

Item Pool Calibration

ADCOM data. One person obtained a total score of 160 by responding in the highest category to all items, so that case was deleted from the dataset prior to calibrating the data. The PARTIAL program was run on the item responses from the remaining 490 individuals to 40 items. Initial results revealed that an MLE of the lowest step value for one item was unobtainable, because no person responded in the lowest category. In effect, the item was a three-category item rather than a four-category item, and thus did not have the same functional response scale as the other 39 items. Consequently, the item was deleted from the potential item pool, and the remaining items were recalibrated. After eight stages of the program, parameter estimates of the step values had converged for the remaining 39 items. Table 1 shows descriptive statistics for the scale values and response threshold parameter estimates that were obtained with the procedure described by Wright and Masters for the three datasets.

Table 1
Descriptive Statistics, Scale Value Estimates, and Threshold Estimates for Three Datasets

	ADCOM	AWS	Artificial
Scale Value			
Mean	-.855	-.475	.022
SD	.709	.838	.604
Minimum	-1.985	-1.864	-.838
Maximum	.829	.903	1.611
Number of Items	39	24	32
Threshold			
1	-1.347	-.728	-.969
2	-.536	-.091	-.028
3	.024	.819	.997
4	1.859		

AWS data. Two individuals responded in the highest category to all 25 items and were deleted from the subsequent analyses. The calibration of

the remaining data from 531 persons revealed that the step value for the lowest category of one item could not be estimated, because only one person responded in the lowest category for that item, and thus for all practical purposes this item had only three response alternatives. Consequently, the item was deleted from the item pool. After seven stages, the step value estimates had converged for the remaining 24 items.

Artificial data. The PARTIAL program required five stages to reach convergence for the parameter estimates for all 32 items. One simulee responded in the highest category to every item and was deleted.

Item Pool Information

Figure 2 shows the total information function for the three item pools. Although the ADCOM, AWS, and artificial item pools are displayed in the same figure, direct comparisons cannot be made across the item pools because the item pools were not equated. The information function for the 39 ADCOM items that were used for the simulated CATs was quite peaked and slightly positively skewed. The function reached its maximum level at a θ value of –1.3, which was the value used as the initial θ estimate for persons during the adaptive testing procedures. Few items were available in the pool that provided much information for persons with relatively positive attitudes toward the communication skills of their administrators (θ values greater than 2).

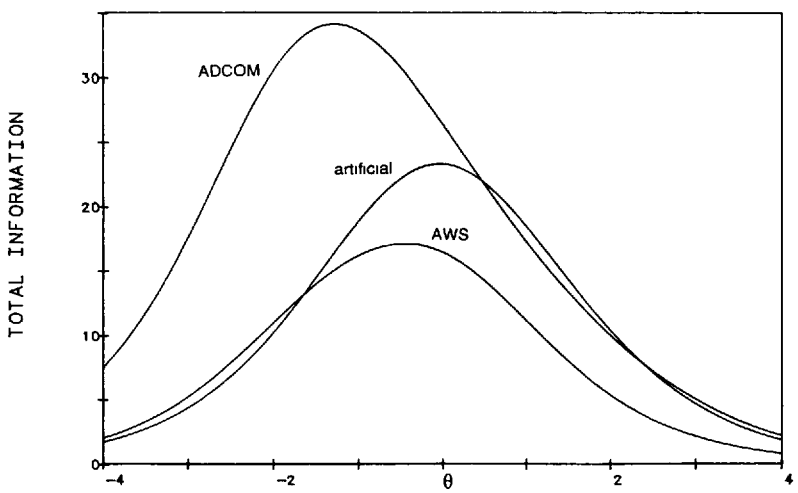
As was the case with the ADCOM item pool, the information function for the 24 AWS items was quite peaked. For this item bank, however, the function was fairly symmetrical and shifted to the lower end of the θ scale. The information reached a maximum level for a θ value of –.5, which was used as the initial θ estimate for the simulated adaptive testing procedures for all individuals. This type of information function is typical of those found for most Likert-type scales, in that it provides accurate measurement near the middle of the attitude continuum and much less accuracy at the extremes.

The artificial scale of 32 items had a symmetric function around θ = 0. Thus, the initial θ estimate for all simulated CATs was set at 0.

CAT Simulations

In order to address the question of whether the two item selection procedures resulted in the same items being administered during the CAT simulations, several audit trails of the CAT procedures were inspected and compared. The two item selection procedures did not select the same items

Figure 2
Total Information for the Three Item Pools



for administration to an individual with the same θ level. A comparison of the amount of information provided by the different items selected for the same θ estimate revealed that, for all practical purposes, the items provided the same accuracy of measurement.

Nonconvergent cases. Several cases of nonconvergence of θ parameter estimation occurred during the CATs. Table 2 summarizes the frequency of nonconvergent cases for each dataset separately and for the combined datasets. For all three datasets, the information selection procedure used in conjunction with a fixed stepsize of .7 yielded the largest number of nonconvergent cases. This was because a fixed stepsize of .7 resulted in θ estimates that were functionally outside the range of the item pool after only a few items had been administered. For the extreme θ estimates, there were no informative items left in the item bank that met the .44 minimum information required for selection, and thus the CAT session was terminated prematurely. Considerably fewer cases of nonconvergence were found for the fixed stepsize of .4 used in conjunction with the item information selection technique. No cases of nonconvergence were found for the variable stepsize, regardless of the item selection procedure employed. The scale value item selection method, in combination with a fixed stepsize of .4, also resulted in no cases of nonconvergence.

Table 2
Number of Nonconvergent Cases for Three Datasets (ADCOM, AWS, Artificial) Under Six Adaptive Conditions, and All Datasets Combined

Dataset and Item Selection Method	Stepsize		
	.4	.7	Variable
ADCOM (N = 490)			
Information	1	5	0
Scale Value	0	0	0
AWS (N = 531)			
Information	5	11	0
Scale Value	0	1	0
Artificial (N = 499)			
Information	0	7	0
Scale Value	0	2	0
Combined Data (N = 1,520)			
Information	6	23	0
Scale Value	0	3	0

Descriptive statistics. The means and standard deviations of the θ estimates, standard errors, and the number of items administered under each of the six CAT conditions and the full scale calibration for each dataset are presented in Table 3. The average θ estimates for each of the six CAT conditions and for the full scale calibration within each dataset were very similar. For the ADCOM and AWS datasets, the mean standard errors were virtually identical for the six adaptive conditions; yet the scale value item selection procedure administered two to three fewer items on the average than the corresponding item information selection method. The information selection procedure gave slightly more items on the average for the artificial data, but resulted in an average standard error that was slightly higher than the corresponding scale value item selection technique. This somewhat surprising result led to an inspection of the standard errors associated with the various θ levels for each CAT condition for the three datasets.

Standard error plots. For each CAT condition and the full scale calibration, the standard errors were plotted against the θ estimates for each of the datasets. It was found that for the scale value selection procedure, regardless of the stepsize used, the standard errors were uniformly quite low, except for the part of the θ scale in which the total information function was low. The plots were also very similar for the three stepsize methods under the item information selection procedure. Because the same result was found across all datasets, only the graph for the artificial dataset is presented. Figure 3 presents the standard error plotted against the θ estimates that were obtained under (1) the variable stepsize for the scale value selection method, (2) the variable stepsize used in conjunction with the item information selection procedure, and (3) the full scale calibration. The standard errors obtained with the CAT procedure using a scale value selection method were slightly higher than those obtained

Table 3
Mean and Standard Deviation of θ, Standard Error, and Number of Items Administered for Three Datasets Under Six Adaptive Conditions and Full-Scale Calibration

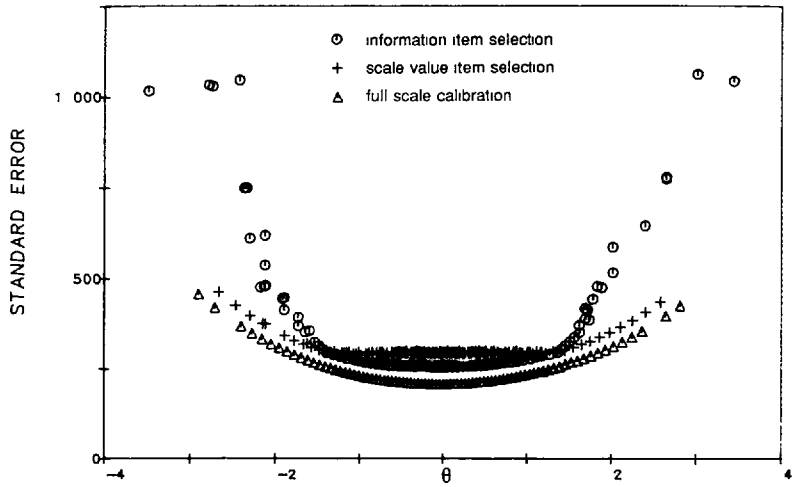
Dataset, Item Selection Method, and Stepsize	θ Estimate		Standard Error		Number of Items	
	Mean	SD	Mean	SD	Mean	SD
ADCOM (N = 484)						
Information						
.4	.01	1.13	.31	.18	18.15	4.34
.7	.03	1.14	.30	.18	18.18	4.28
variable	.02	1.13	.31	.18	18.21	4.29
Scale Value						
.4	−.01	1.04	.30	.03	14.98	2.80
.7	−.03	1.04	.30	.03	14.88	2.81
variable	−.02	1.04	.30	.03	14.89	2.81
Full Scale	.01		.21		39.00	
AWS (N = 520)						
Information						
.4	−.02	.90	.31	.13	17.32	4.07
.7	−.02	.92	.31	.14	17.23	4.28
variable	−.02	.92	.31	.15	17.22	4.30
Scale Value						
.4	−.05	.83	.31	.05	15.30	2.09
.7	−.06	.84	.31	.05	14.38	2.27
variable	−.06	.84	.31	.05	14.37	2.28
Full Scale	−.04		.27		24.00	
Artificial (N = 491)						
Information						
.4	−.03	1.17	.34	.19	17.67	5.05
.7	−.02	1.20	.35	.22	17.32	5.57
variable	−.01	1.19	.34	.19	17.55	5.20
Scale Value						
.4	.00	1.04	.30	.03	16.61	1.96
.7	.01	1.03	.30	.03	16.64	1.96
variable	.00	1.04	.30	.03	16.63	1.97
Full Scale	.00		.23		32.00	

with the full scale calibration. This was a function of the value of .3 used for the standard error stopping rule.

Figure 3 also shows that the standard errors were quite high for both high and low θ estimates obtained for the item information selection procedure. The audit trails of the CAT procedures show that the high standard errors were due to the CAT being terminated after only three to four items had been administered, because no informative items remained in the item pool for these extreme θ estimates. This finding is a possible explanation for the somewhat larger standard deviations of the standard errors and numbers of items administered that were obtained under the information selection procedure than were obtained for the scale value item selection method.

Intercorrelations of θ estimates. The intercorrelations of the θ estimates obtained under the six CAT conditions and the full scale calibration for each of the datasets were very high, and the scattergrams revealed that the relationships were essentially linear. The correlation between the θ estimates obtained under the scale value item

Figure 3
Standard Errors of the θ Estimates for the Artificial Data Under the Full Scale Calibration and the Two Variable Stepsize Adaptive Testing Conditions



selection procedure and the full scale calibration were generally slightly higher (r = .96 to .98) than the correlations between the θ estimates obtained under the item information selection procedure and the full scale calibration (.94 to .97).

For the artificial data, it was possible to determine the relationship between the known z values used to generate the data and the θ estimates yielded by the various adaptive testing conditions. Again, the correlation coefficients obtained for the scale value item selection procedure (.96) were slightly higher than those obtained for the item information selection technique (.93 to .95).

Discussion

In general, all six rating scale CAT procedures performed well. The correspondence between the θ estimates from the six CAT conditions, the full scale calibration θ estimates, and the known θ values were very high. These results were particularly impressive given the fact that one of the item pools contained only 24 items. Investigations of CAT procedures for other polychotomous IRT models (Dodd et al., 1989; Koch & Dodd, 1989) have also found that CAT procedures perform well with item pools as small as 30 items. The reason the CAT procedures perform successfully with such small item pools is that the polychotomous scoring of each item provides information across almost the entire range of the θ scale. Therefore, few gaps of appropriate items are likely to exist in the item pool.

There were some problems with nonconvergence of the θ estimates for several of the rating scale CAT conditions. Nonconvergent cases were found under almost all CAT conditions that used a fixed stepsize prior to MLE. Using a variable stepsize, however, eliminated this problem altogether.

It was interesting that the two item selection techniques did not consistently select the same item for administration for a given θ estimate because the item information function for the rating scale model peaks near—but not exactly at—the scale value for the item. It should be noted, however, that the difference in the amount of information provided by the selected items for a given θ estimate using the two methods was extremely small.

The standard error plots and CAT summaries revealed that the use of the maximum information item selection procedure in conjunction with

the minimum information stopping rule produced some undesirable results. High standard errors were obtained for extreme θ estimates because the CAT was terminated after only three to four items had been administered. There were no items left in the item pool that could meet the minimum information of .44 to be administered. Although it would have been a simple matter to lower the information cutoff value, it would have resulted in a maximum test length of 20 items for almost all individuals in the middle of the θ range. Alternatively, a standard error stopping rule could have been used in conjunction with the information item selection procedure to alleviate the problem.

The scale value item selection method used in conjunction with the standard error stopping rule administered two to three fewer items on the average than the information item selection method used in conjunction with the minimum information stopping rule. The scale value selection technique also outperformed the information item selection method in terms of the frequencies of nonconvergent cases, and it produced slightly higher correlations of the CAT θ estimates with the full scale θ estimates and the known θ values. In addition, the calculations involved in the scale value item selection procedure were less complex and faster than the calculation of maximum item information.

The findings of the present research suggest several guidelines for CAT using the rating scale model. First, the CAT procedures should perform well with item pools as small as 25 to 30 items. Second, the use of a variable stepsize prior to MLE should produce fewer cases of nonconvergence than the use of a fixed stepsize. Third, the scale value item selection technique used in conjunction with a standard error stopping rule should perform better than the information item selection procedure used in conjunction with an information stopping rule.

The rating scale CAT system outlined above could prove very useful to researchers interested in the measurement of people's attitudes. The adaptive testing procedure would enable the researcher to assess an individual's attitude very accurately with relatively few items. Koch, Dodd, and Fitzpatrick (1990) have implemented an adaptive version of an attitude scale using the procedures suggested in the present research. In general, the respondents preferred the computerized adaptive assessment to the conventional pencil-and-paper version of the attitude scale.

References

Andrich, D. (1978a). Application of a psychometric 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.

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 IL, U.S.A.

Dodd, B. G., Koch, W. R., & DeAyala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13,* 129–143.

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. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education, 2,* 335–357.

Koch, W. R., Dodd, B. G., & Fitzpatrick, S. J. (1990). Computerized adaptive measurement of attitudes. *Measurement and Evaluation in Counseling and Development, 23,* 20–30.

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

Patience, W. M., & Reckase, M. D. (1979). *Operational characteristics of a one-parameter tailored testing procedure.* (Research Report 79-2). Columbia: University of Missouri, Educational Psychology Department, Tailored Testing Research Laboratory.

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

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

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

Spence, J. T., Helmreich, R., & Stapp, J. (1973). A short version of the Attitude Toward Women Scale (AWS). *Bulletin of the Psychonomic Society, 2,* 219–220.

Valentine, R. J. (1978). *Audit of administrator communication.* Columbia MO: Jerry W. Valentine.

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

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

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

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago IL: MESA Press.

Author's Address

Send requests for reprints or further information to Barbara G. Dodd, Measurement and Evaluation Center, University of Texas at Austin, Box 7246, Austin TX 78713, U.S.A.