The Nominal Response Model in Computerized Adaptive Testing

R. J. De Ayala
University of Maryland

Although most computerized adaptive tests (CATs) use dichotomous item response theory (IRT) models, research on the use of polytomous IRT models in CAT has shown promising results. This study implemented a CAT based on the nominal response model (NR CAT). Item pool requirements for the NR CAT were examined. The performance of the NR CAT and a CAT based on the three-parameter logistic (3PL) model was compared. For two-, three-, and four-category items, items with maximum information of at least .16 produced reasonably accurate trait estimation for tests with a minimum test length of approximately 15 to 20 items. The NR CAT was able to produce trait estimates comparable to those of the 3PL CAT. Implications of these results are discussed. *Index terms: adaptive testing; computerized adaptive testing; EAP estimation; nominal response model; polytomous models.*

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

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

Dichotomous model-based CATs have not incorporated findings from human cognition studies (e.g., Brown & Burton, 1978; Brown & VanLehn, 1980; Lane, Stone, & Hsu, 1990; Tatsuoka, 1983). For example, Tatsuoka's (1983) analysis of student misconceptions in solving mathematics problems showed that incorrect responses could be of more than just one kind; however, dichotomous scoring uniformly assigned a score of zero to all incorrect responses. Moreover, Nedelsky (1954) demonstrated from a classical test theory perspective, and Levine & Drasgow (1983) from an IRT perspective, that the distribution of incorrect answers over the options of multiple-choice items differed across trait levels. However, an item's incorrect alternatives may augment the estimate of an examinee's trait level (θ) by providing information about the examinee's level of understanding (i.e., provide diagnostic information). Bock (1972) and Thissen (1976) found that for examinees with θ estimates in the lower half of the θ range the nominal response (NR) model provided from one-third to nearly twice the information furnished by a dichotomously scored two-parameter model; there was no difference in information yield between these two models for θ estimates above the

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 4, December 1992, pp. 327-343

0146-6216/92/040327-17$2.10

median θ. In an application to multiple-choice and free-response items, Vale & Weiss (1977) found that the NR model provided more information for middle θ examinees than that shown in the Bock (1972) and Thissen (1976) studies.

In classical test theory, the use of proper scoring techniques to assess partial knowledge yields increases in the reliability of multiple-choice tests (e.g., Coombs, Milholland, & Womer, 1956). Frary (1989), Haladyna & Sympson (1988), and Wang & Stanley (1970) provided reviews of the literature on option scoring strategies. It is obvious that the dichotomization of an examinee's response ignores any partial knowledge that the examinee may have of the correct answer and, as a result, this information cannot be used for θ estimation.

Some research has explored the benefits and operating characteristics of CATs based on polytomous IRT models (e.g., Dodd, Koch, & De Ayala, 1989; Koch & Dodd, 1989; Sympson, 1986). Research on the use of polytomous IRT models in CAT has shown promising results. For instance, Sympson (1986) found that an adaptive test based on a polytomous model (Model 8) could be 15-20% shorter than a paper-and-pencil test without sacrificing test reliability. These studies have shown that (1) using item pools smaller than those used with dichotomous model-based CATs leads to satisfactory estimation; (2) using the standard error of the θ estimate (SEE) for terminating an adaptive test is preferred to using the minimum item information termination criterion; and (3) using a variable stepsize instead of a fixed stepsize tends to minimize nonconvergence of θ estimation. The models used in these studies included Masters' (1982) partial credit model, Andrich's (1978) rating scale model, and Samejima's (1969) graded response model.

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

Purpose

This study concerned the implementation of an NR model-based CAT (NR CAT) and had several objectives:

1. Because the NR model is written in terms of slope and intercept parameters (a form not typically used; cf., Hambleton & Swaminathan, 1985; Lord, 1980; Weiss, 1983), formulas for the location parameters were derived in order to facilitate understanding of the model's formulation. The NR model's relationship with the dichotomous two-parameter logistic (2PL) model is presented. In addition, the effect of varying the location parameters on item information is examined.
2. The quality of the item pool is paramount to CAT performance. Two factors that determine the item pool's quality are the locations of the items and their discrimination indices. Items should be evenly and equally distributed throughout the θ continuum of interest (Patience & Reckase, 1980; Urry, 1977; Weiss, 1982). Because there was no reason to believe that this would not hold for the NR model, this factor was not studied. The minimum item information (i.e., the effect of item discrimination) that would allow reasonably accurate θ estimates by the NR CAT was investigated. This investigation (Study 1) was limited to two-, three-, and four-category cases.
3. The comparative performance of the NR CAT and a CAT based on a dichotomous 3PL model was assessed (Study 2). Furthermore, because of the existence of option informa-

tion, an exploratory simulation was conducted in which items were selected on the basis of option information.

The NR Model

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

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

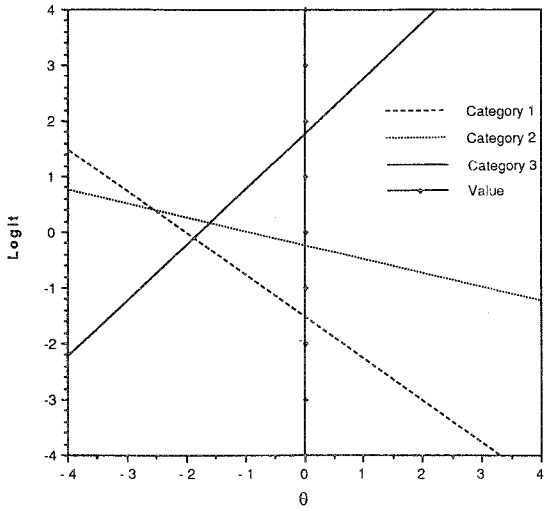
where a_{ij} is the slope parameter,
c_{ij} is the intercept parameter of the nonlinear response function associated with the *j*th category of item *i*, and
m_i is the number of categories of item *i* (i.e., $j = 1, 2, \ldots, m_i$).

For convenience, the slope and intercept parameters are sometimes represented in vector notation, where $\mathbf{a} = (a_{i1}, a_{i2}, \ldots, a_{im})$ and $\mathbf{c} = (c_{i1}, c_{i2}, \ldots, c_{im})$, respectively.

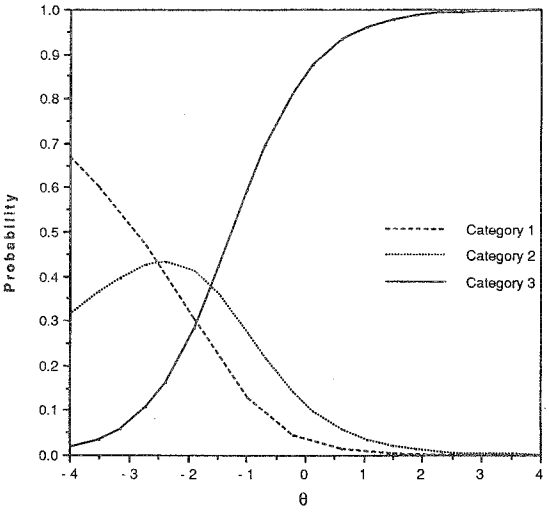
To aid in the interpretation of these parameters, Figure 1a shows a logistic space plot of the (multivariate) logit (i.e., $c_{ij} + a_{ij}\theta$) against θ for a three-category ($m = 3$) item with $\mathbf{a} = (-.75, -.25, 1.0)$ and $\mathbf{c} = (-1.5, -.25, 1.75)$. As the figure shows, c_{ij} is the *y* intercept (i.e., $\theta = 0.0$), and a_{ij} is the slope of the category's response function. The a_{ij}s are analogous to, and have an interpretation similar to, traditional option discrimination indices. That is, a cross-tabulation of θ groups by item alternatives shows that a category with a large a_{ij} reflects a response pattern in which progression from the lower θ groups to the higher θ groups results in a corresponding increase in the number of persons who answered the item in that category; for categories with negative a_{ij}s, this pattern is reversed. Large values of c_{ij} seem to be associated with categories with large frequencies. As the value of c_{ij} becomes smaller, the frequencies for the corresponding categories decrease.

Figure 1
A Three-Category Item, With $\mathbf{a} = (-.75, -.25, 1.0)$ and $\mathbf{c} = (-1.5, -.25, 1.75)$

a. Multivariate Logit Plot



b. Example of Corresponding ORFs



The probability of responding in a particular category as a function of θ is depicted by the category or option response function (ORF). Figure 1b contains the ORFs corresponding to the three-category item presented in Figure 1a. The intersection of the ORFs can be obtained by setting the multivariate logits of adjacent categories

equal to one another and solving for θ. Therefore,

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

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

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

This formulation is analogous to that of the partial credit model in which step difficulties are defined at the intersection of adjacent category response functions.

Bock (1972) compared the NR model with a binary version (i.e., the items consist of correct and incorrect categories). When $m_i = 2$, Equation 1 becomes

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

Given Equation 4, note that the two linear constraints imposed on the item parameters—$\Sigma a = 0$ and $\Sigma c = 0$, to address the indeterminacy of scale—imply that in the two-category case

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

and

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

Therefore, given Equations 5 and 6, for $m_i = 2$

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

By solving Equation 7 for c_2 and substituting the equality into Equation 4,

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

By substituting Equation 5 into Equation 8 and simplifying,

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

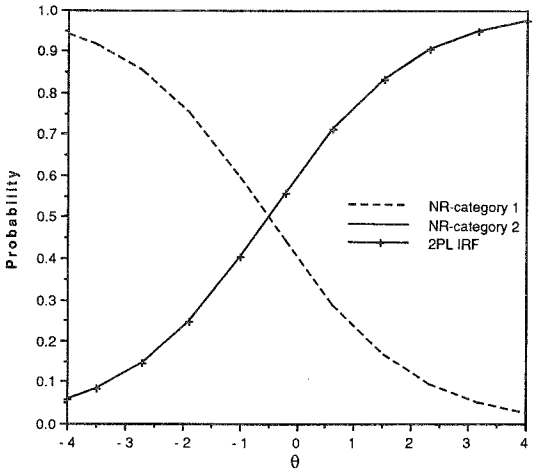
is obtained.

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

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

For $m_i = 2$, the 2PL and NR models are equivalent. For example, Figure 2 shows the NR model's ORFs for an item with $a_2 = .40$, $a_1 = -.40$, $c_2 = .2$, and $c_1 = -.20$ and the item response function (IRF) for the 2PL model with $a = .80$ and $b = -(.4/.8) = -.5$.

Figure 2
NR Model ORFs ($a_2 = .40$, $a_1 = -.40$, $c_2 = .2$, and $c_1 = -.20$) and the 2PL IRF ($a = .80$ and $b = -.5$)



Information

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

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

and item information is

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

For a given item i,

$$\mathbf{W} = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_m \\ -p_2p_1 & p_2(1-p_2) & \cdots & -p_2p_m \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ -p_mp_1 & -p_mp_2 & \cdots & p_m(1-p_m) \end{bmatrix} . \quad (13)$$

For the $m_i = 2$ case, the location of maximum item information (I_{max}) is

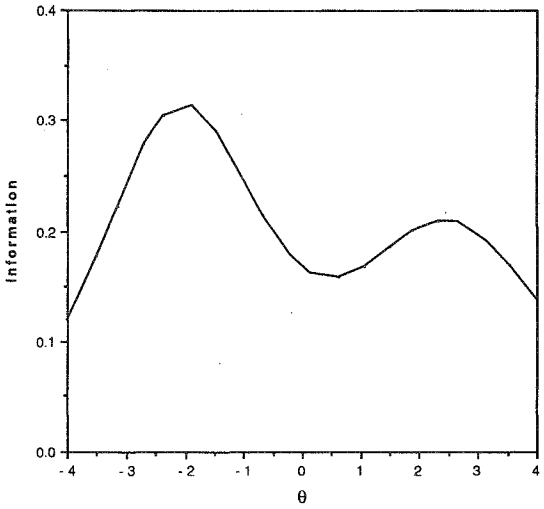
$$\theta_{max} = \frac{c_1 - c_2}{a_2 - a_1} \quad (14)$$

with $I_{max} = .25(a_2 - a_1)^2$. Due to the number of unknowns, a formula for the location of I_{max} cannot be determined for $m_i > 2$. For $m_i = 2$ and a given **a**, changing the values of **c** forces the location of I_{max} to shift along the θ continuum, but the maximum amount of information remains constant.

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

For $m_i = 4$ and a given **a**, if the *b*s are in ascending order, then the item information function becomes more platykurtic as the difference between *b*s becomes less extreme. This pattern holds if the last two *b*s are reversed. When the *b*s are in descending order, then relative to the item information function when the *b*s are in ascending order, the function becomes more leptokurtic as the difference between *b*s becomes less extreme. This is also true if the first two *b*s are transposed. For the other two possible *b* patterns, the information function becomes comparatively more leptokurtic as the distance among the *b*s decreases. It is possible to obtain bimodal item information functions. For instance, Figure 3 contains the information function for an item with **a** = (1, .1, -.1, -1) and **c** = (.1, 2.4, -2.6, .1).

Figure 3
Bimodal Information Function for an Item With **a** = (1, .1, -.1, -1) and **c** = (.1, 2.4, -2.6, .1)



As Equation 12 implies, item information is a function of the magnitude of the elements of **a**, and the order of the elements of **a** [i.e., for a given **c**, **a** = (-.25, 1.0, -.75), **a** = (-.25, -.75, 1.0), and **a** = (-.75, -.25, 1.0)] will produce three different I_{max}s at three different θ_{max}s. For a given **a**, the signs of the elements are irrelevant as long as $\Sigma\mathbf{a} = 0$ (and $\Sigma\mathbf{c} = 0$). For instance, given two items that have the same **c** [e.g., **c** = (.25, -.15, -.1)] but that differ only in the sign of the elements, such as **a** = (.4, .25, -.65) and **a** = (-.4, -.25, .65), the items will have the same $I_{max} = .245$ but at different θ_{max}s; specifically, $\theta_{max} = .84$ for **a** = (-.4, -.25, .65), and $\theta_{max} = -.84$ for **a** = (.4, .25, -.65). This is also true in the four-category case. Given the same **c**, two items that differ only in the sign of the elements of **a** (and satisfy $\Sigma\mathbf{a} = 0$), such as **a** = (.55, .4, -.35, -.6) and **a** = (-.55, -.4, .35, .6), will yield $I_{max} = .26$ at $\theta_{max} = .059$ and $\theta_{max} = -.059$, respectively.

Method

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

Programs. A program was written for per-

forming adaptive testing with the NR model. The program used expected a posteriori (EAP) estimation (Bock & Mislevy, 1982) of θ. Item selection was based on information. The adaptive testing simulation was terminated when a maximum of 30 items was administered. θ estimates at test lengths of 10, 15, 20, 25, and 30 items were recorded. The initial θ estimate for an examinee was the population's mean, and a uniform prior with 10 quadrature points was used. An additional program for generating data according to the NR model was written and is discussed below.

Data. A series of item pools was created. The item pools differed on two factors: I_{max} and the number of item alternatives (m = 2, 3, or 4 options). The item pool contained 90 items (cf., Dodd, Koch, & De Ayala, 1989; Koch & Dodd, 1989).

Although Urry's (1977) guidelines for the discrimination parameter were stated in terms of the magnitude of a, the importance of an item's a value is its effect on I_{max}. When the number of categories is three or more, different combinations of **a** and **c** can produce the same I_{max} value; therefore, establishing guidelines in terms of the magnitudes of the elements of these vectors was not pursued. Rather, specified values for I_{max} were set a priori, and the **a** vector to obtain a specific I_{max} was determined. The I_{max} values studied were .25, .16, .09, and .04.

When m_i = 2, the **a** vectors may be specified a priori. For I_{max} values of .25, .16, .09, and .04, the corresponding **a** vectors were (.50, -.50), (.40, -.40), (.30, -.30), and (.20, -.20), respectively. (For the 2PL model, these **a** vectors are equivalent to a values of 1.0, .8, .6, and .4, respectively.) Because Urry (1977) recommended the use of items with $a \geq .80$ in CAT, for m_i = 2 the use of **a** = (.40, -.40) was expected to be equivalent to the use of a = .80 with a 2PL model-based CAT. For each I_{max} level of the m_i = 3 and m_i = 4 conditions, the **a** vectors for the items were selected using a trial-and-error procedure to approximate the relevant I_{max} value.

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

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

1,300 examinee θs were generated to be evenly distributed between -3.0 and 3.0 using a .5 logit interval between successive θ levels (i.e., for 100 examinees θ = -3.0, for 100 examinees θ = -2.5, etc.). These true θs plus the 90 item parameters for each condition were used to generate polytomous response strings with a random error component for each simulated examinee (i.e., 12 response datasets were created). Generation of an examinee's polytomous response string was accomplished by calculating the probability of responding to each alternative of an item according to the NR model. Based on the probability for each alternative, cumulative probabilities were obtained for each alternative. A random error component was incorporated into each response by selecting a random number from a uniform distribution [0,1] and comparing it to the cumulative probabilities. The ordinal position of the first cumulative probability that was greater than the random number was taken as the examinee's response to the item.

Analysis. Study 1 was designed to determine the minimum I_{max} value that would result in a

significant improvement in the estimation of θ. The accuracy of θ estimation was assessed by root mean square error (RMSE) and bias. RMSE and bias were calculated as

$$\text{RMSE}(\theta) = \left[\frac{\sum(\hat{\theta}_k - \theta)^2}{n_f}\right]^{1/2} \quad (15)$$

and

$$\text{bias}(\theta) = \frac{\sum(\hat{\theta}_k - \theta)}{n_f} \quad , \quad (16)$$

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

The analysis of the two-, three-, and four-category cases was treated separately. Therefore, a one-group repeated measures design with two dependent variables—RMSE and bias—was used. $I_{\max}$ was the between persons factor, and test length was the within persons factor. The test length factor was included because the accuracy of θ estimation is influenced by both the adaptive test length and the information content of the items administered. Because the Bonferroni method was used to control for familywise Type I error, α was set at .008 (i.e., .05/6). Post hoc analysis was performed with the Scheffé test using a critical F of 13.26, based on 3 and 48 degrees of freedom (Hays, 1988).

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

Programs. The NR CAT program used in Study 1 also was used in Study 2. The program was designed to select items on the basis of either item or option information. An additional CAT program based on the 3PL model (3PL CAT) was written. The 3PL CAT program estimated θ by EAP and selected items on the basis of information. The adaptive testing simulation was terminated when either of two criteria was met—a maximum of 30 items was administered or a predetermined SEE was obtained (SEE termination criteria of .20, .25, .30 were used). The initial θ estimate for an examinee was the mean of the population. Both CATs used a 10-point uniform prior distribution for θ.

A data generation program based on a linear factor analytic model (Wherry, Naylor, Wherry, & Fallis, 1965) was written and is discussed below. The linear factor analytic approach for generating the data was used to minimize any bias in favor of either the 3PL or NR model; this procedure has been used previously (De Ayala, Dodd, & Koch, 1992; Dodd, 1984; Koch, 1981; Reckase, 1979). MULTILOG (Thissen, 1988) was used to obtain item parameter estimates for the NR and 3PL models using default program parameters.

Data. 1,300 examinee θs were generated to be evenly distributed between −3.0 and 3.0 using a .5 logit interval between successive θ levels. The examinees' responses to 150 four-alternative items were generated according to the linear factor analytic model:

$$z_{ki} = a_i\theta_k + (1 - h_i^2 z_{e_{ki}})^{1/2} \quad , \quad (17)$$

where θ_k is examinee ks latent θ level,
a_i is the factor loading of item i,
h_i^2 is the communality of item i, and
$z_{e_{ki}}$ is a random number generated from a N(0,1) distribution and is the error component of examinee k and item i.

All factor loadings were uniformly high and ranged from .62 to .84. z_{ki} was compared to prespecified category boundaries to determine the category response for examinee k to item i.

MULTILOG was used to obtain item parameter estimates for both the NR and 3PL models. Based on the results of Study 1, item pools for the NR and the 3PL CATs were constructed by identifying items with values of $I_{max} \geq .16$ and with $\theta_{\max}$ values distributed evenly throughout the −2.0 to 2.0 θ range. These items were replicated to produce item pools of 152 items.

Analysis. Study 2 was designed to determine whether there were any psychometric advantages to using the polytomous NR model as opposed to the dichotomous 3PL model. The quality of the θ estimation provided by the two CATs was

analyzed by calculating RMSE and bias. Moreover, the number of items administered (NIA) in obtaining $\hat{\theta}$ also was used for comparing the two types of CATs. A one-group repeated measures design with three dependent variables—RMSE, bias, and NIA—was used. The between persons factor was the type of CAT used (NR, 3PL), and the repeated measures or within persons factor was the SEE termination criterion used (.20, .25, .30). The Bonferroni method was used to control for familywise Type I error, and α was set at .0056. Post hoc analysis was performed with the Scheffé test using a critical F of 10.223 based on 1 and 16 degrees of freedom (Hays, 1988).

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

Results

Study 1

Table 1 contains descriptive statistics on the NR adaptive tests. As expected, there was a direct relationship between the fidelity coefficient, $r_{\hat{\theta}\theta}$, and $I_{\max}$, and between $r_{\hat{\theta}\theta}$ and test length. For $I_{\max} = .25$, there was a slight increase in $r_{\hat{\theta}\theta}$ as the number of categories increased for a given test length (10, 15, 20, or 25 items)—this increase in $r_{\hat{\theta}\theta}$ tended to diminish with increasing test length. For instance, for a test length of 10 items, $r_{\hat{\theta}\theta}$ increased from .935 ($m = 2$) to .942 ($m = 4$).

The repeated measures analyses are presented in Tables 2 through 5. For the two-category condition, the average RMSE improved significantly as test length and $I_{\max}$ increased. Post hoc analysis of the $I_{\max}$ factor (Table 4) showed that for the two-category case there was a significant reduction in RMSE as $I_{\max}$ increased from .04 to .09 to .16 across test lengths. Increasing the item information from .16 to .25 did not produce a significant improvement in θ estimation as assessed by RMSE. For the 10-item test, there was a significant improvement in the accuracy of estimation from $I_{\max} = .16$ to .25 (Table 4). That is, for the shorter test length (10 items) more informative items were needed than at longer test lengths.

For all $I_{\max}$ values, there was a significant improvement in the accuracy of estimation as test length increased from 10 to 15 to 20 items (see Table 5). At higher item information levels (e.g., .16 and .25), increasing test length from 20 to 25 items or from 25 to 30 items did not yield a significant reduction in RMSE. For $I_{\max} = .09$, estimation accuracy was significantly improved by increasing the test length from 20 to 25 items, but not from 25 to 30 items. These results suggest that using items with $I_{\max} \geq .16$ (i.e., $a \geq .80$) provides reasonable θ estimation for tests of 20 (possibly 15) or more items. Shorter tests require more informative items than longer tests.

Results of the ANOVAs (Table 2) show that test length and $I_{\max}$ did not have a significant effect on bias. This is, in part, a function of the way bias was calculated and the potential for cancellation of negative bias by positive bias. Figure 4 contains RMSE and bias plots for selected NR CATs—these plots were typical of all the NR CAT plots.

For the three-category condition and test lengths of 20 or more items, the results were similar to the two-category condition (Tables 2–5). There was a significant reduction in RMSE as $I_{\max}$ increased from .04 to .09 to .16, but not from .16 to .25. However, for the 10- and 15-item test lengths, the results were the reverse of those of the two-category condition.

In general, the results for the four-category condition paralleled those of the two- and three-category conditions. There was a significant reduction in RMSE as $I_{\max}$ increased from .04 to .09 to .16 to .25 for tests of 20 or fewer items. There was no significant reduction in RMSE as $I_{\max}$ increased from .16 to .25 for tests of 25 or 30 items.

Study 2

Table 6 contains descriptive statistics for the

Table 1
Mean $\hat{\theta}$, Standard Deviation of $\hat{\theta}$ (SD), and Pearson Product-Moment Correlations Between θ and $\hat{\theta}$ (r) for NR Adaptive Tests with Mean $\theta = 0.0$ and $s_\theta = 1.872$ for $m = 2$, 3, and 4, and I_{max} From .25 to .04 at Test Lengths of 10 to 30 Items

		Test Length				
I_{max}	Statistic	10	15	20	25	30
$m = 2$						
.25	Mean	.021	.010	.002	−.002	−.003
	SD	1.936	1.921	1.906	1.898	1.900
	r	.935	.956	.967	.973	.977
.16	Mean	.027	.007	.001	−.009	−.009
	SD	1.949	1.927	1.923	1.918	1.914
	r	.910	.938	.954	.962	.968
.09	Mean	.052	.006	−.003	−.140	−.003
	SD	1.952	1.948	1.948	1.950	1.937
	r	.863	.905	.926	.939	.949
.04	Mean	.068	.061	.020	.014	.009
	SD	1.875	1.908	1.932	1.945	1.956
	r	.759	.818	.855	.880	.900
$m = 3$						
.25	Mean	−.003	.003	0.000	.004	.001
	SD	1.951	1.936	1.929	1.924	3.670
	r	.936	.958	.968	.974	.978
.16	Mean	−.003	−.014	−.004	.010	.009
	SD	1.959	1.951	1.952	1.942	1.938
	r	.918	.939	.956	.964	.971
.09	Mean	−.004	−.006	−.009	−.008	0.000
	SD	1.965	1.963	1.958	1.951	1.950
	r	.863	.903	.929	.941	.950
.04	Mean	−.020	0.000	.009	.015	.003
	SD	1.881	1.922	1.939	1.950	1.954
	r	.763	.831	.868	.890	.907
$m = 4$						
.25	Mean	−.007	−.008	−.013	−.014	−.016
	SD	1.969	1.951	1.941	1.943	1.934
	r	.942	.960	.969	.974	.977
.16	Mean	−.034	−.025	−.028	−.031	−.035
	SD	1.979	1.974	1.973	1.960	1.961
	r	.912	.939	.951	.959	.964
.09	Mean	−.015	−.007	−.006	−.016	−.017
	SD	1.978	1.979	1.975	1.986	1.985
	r	.855	.902	.925	.938	.945
.04	Mean	−.034	−.008	−.001	−.002	−.009
	SD	1.902	1.941	1.963	1.976	1.982
	r	.752	.816	.847	.876	.892

NR and 3PL adaptive tests. These results for the NR and 3PL CATs are comparable, but a meaningful difference did occur in $r_{\hat{\theta}\theta}$ at a termination SEE of .30. However, the NR CAT tended to administer adaptive tests which, on average, were shorter than those administered by the 3PL CAT.

Table 7 contains results for the ANOVAs. The NR CAT and 3PL CAT did not differ significantly with respect to RMSE and bias. The NR CAT

Table 2
Results of the Repeated Measures ANOVAs for RMSE and Bias for the Two-, Three-, and Four-Category NR CAT Conditions

Source	SS	*df*	MS	*F*	*p*
RMSE, Two Categories					
Between Persons					
I_{max}	10.591	3	3.530	167.500*	0.000
Persons Within Groups	1.012	48	.021		
Within Persons					
Test Length	4.237	4	1.059	553.451*	0.000
I_{max} × Test Length	.121	12	.010	5.288	0.000
Test Length × Persons Within Groups	.367	192	.002		
Bias, Two Categories					
Between Persons					
I_{max}	.041	3	.014	.110	.954
Persons Within Groups	5.964	48	.124		
Within Persons					
Test Length	.074	4	.018	2.237	.067
I_{max} × Test Length	.017	12	.001	.176	.999
Test Length × Persons Within Groups	1.580	192	.008		
RMSE, Three Categories					
Between Persons					
I_{max}	9.492	3	3.164	135.396*	0.000
Persons Within Groups	1.122	48	.023		
Within Persons					
Test Length	4.326	4	1.081	495.580*	0.000
I_{max} × Test Length	.168	12	.014	6.407	0.000
Test Length × Persons Within Groups	.419	192	.002		
Bias, Three Categories					
Between Persons					
I_{max}	.002	3	.001	.007	.999
Persons Within Groups	5.114	48	.107		
Within Persons					
Test Length	.005	4	.0013	.157	.960
I_{max} × Test Length	.010	12	.0008	.091	1.000
Test Length × Persons Within Groups	1.677	192	.0087		
RMSE, Four Categories					
Between Persons					
I_{max}	11.713	3	3.904	135.736*	0.000
Persons Within Groups	1.381	48	.029		
Within Persons					
Test Length	3.731	4	.933	556.861*	0.000
I_{max} × Test Length	.177	12	.015	8.826	0.000
Test Length × Persons Within Groups	.322	192	.002		
Bias, Four Categories					
Between Persons					
I_{max}	.018	3	.006	.059	.981
Persons Within Groups	4.888	48	.102		
Within Persons					
Test Length	.004	4	.0010	.134	.970
I_{max} × Test Length	.008	12	.0007	.078	1.000
Test Length × Persons Within Groups	1.599	192	.0083		

*Significant at overall $\alpha = .05$.

Table 3
Mean RMSE for Levels of I_{max} for Two-, Three-, and Four-Category Conditions at Five Test Lengths

	Test Length				
I_{max}	10	15	20	25	30
RMSE, Two Categories					
.04	1.298	1.136	1.018	.931	.854
.09	.999	.833	.733	.667	.612
.16	.809	.668	.575	.521	.478
.25	.684	.561	.486	.441	.402
RMSE, Three Categories					
.04	1.286	1.095	.973	.890	.824
.09	1.001	.843	.724	.659	.609
.16	.773	.672	.570	.513	.464
.25	.687	.556	.482	.438	.402
RMSE, Four Categories					
.04	1.321	1.151	1.058	.957	.893
.09	1.033	.856	.750	.686	.650
.16	.810	.678	.609	.555	.520
.25	.661	.546	.481	.441	.413

administered, on average, fewer items than the 3PL CAT to achieve the same accuracy in estimation, but this difference was not significant using the Bonferroni criterion. That is, the θ estimation of the NR CAT was comparable to that of the 3PL CAT.

With a polytomous model, item information is the sum of the information functions for individual responses (i.e., category or option information function); therefore, an exploratory study that selected items on the basis of category information was conducted, in which items were selected based on which item provided the maximum information for the particular alternative selected by the examinee. This was based on the assumption that selecting items on the basis of category information would be more consistent with the concept of polytomous scoring of examinee responses than selecting items on the basis of item information, which ignores the particular

Table 4
Post Hoc Comparison *F*s for NR CAT I_{max} at Five Test Lengths for Two-, Three-, and Four-Category Conditions

	Test Length				
Comparison	10	15	20	25	30
Two Categories					
$\mu_{.25}$ vs. $\mu_{.16}$	17.749*	13.005	8.998	7.270	6.561
$\mu_{.16}$ vs. $\mu_{.09}$	41.007*	30.926*	28.357*	24.213*	20.397*
$\mu_{.09}$ vs. $\mu_{.04}$	101.553*	104.288*	92.265*	79.169*	66.524*
Three Categories					
$\mu_{.25}$ vs. $\mu_{.16}$	7.487	13.622*	7.400	5.694	3.891
$\mu_{.16}$ vs. $\mu_{.09}$	52.625*	29.602*	24.008*	21.579*	21.284*
$\mu_{.09}$ vs. $\mu_{.04}$	82.226*	64.287*	62.766*	54.019*	46.795*
Four Categories					
$\mu_{.25}$ vs. $\mu_{.16}$	20.337*	15.961*	15.008*	11.905	10.488
$\mu_{.16}$ vs. $\mu_{.09}$	45.553*	29.024*	18.212*	15.720*	15.481*
$\mu_{.09}$ vs. $\mu_{.04}$	75.979*	79.718*	86.898*	67.274*	54.091*

*Significant at overall $\alpha = .05$

Table 5
Post Hoc Comparison *F*s for NR CAT Test Length at Levels of I_{max} for Two-, Three-, and Four-Category Conditions

	I_{max}			
Comparison	.04	.09	.16	.25
Two Categories				
μ_{30} vs. μ_{25}	19.269*	9.831	6.009	4.943
μ_{25} vs. μ_{20}	24.599*	14.157*	9.477	6.581
μ_{20} vs. μ_{15}	45.253*	32.500*	28.109*	18.281*
μ_{15} vs. μ_{10}	85.293*	89.557*	64.613*	49.169*
Three Categories				
μ_{30} vs. μ_{25}	14.224*	8.163	7.840	4.232
μ_{25} vs. μ_{20}	22.495*	13.796*	10.609	6.322
μ_{20} vs. μ_{15}	48.601*	46.240*	33.972*	17.881*
μ_{15} vs. μ_{10}	119.122*	81.515*	33.309*	56.036*
Four Categories				
μ_{30} vs. μ_{25}	13.375*	4.232	4.000	2.560
μ_{25} vs. μ_{20}	33.309*	13.375*	9.522	5.224
μ_{20} vs. μ_{15}	28.242*	36.689*	15.546*	13.796*
μ_{15} vs. μ_{10}	94.357*	102.299*	56.895*	43.184*

*Significant at overall $\alpha = .05$

response an examinee provided (although the likelihood function is a function of an examinee's particular responses). This exploratory study used the same simulated data and programs as Study 2, except that items were selected on the basis of category information rather than on the basis of item information.

The results, provided in Tables 8 and 9, parallel those presented in Table 7. Specifically, the NR CAT that selected items on the basis of category information provided θ estimation that, in terms of RMSE and bias, was comparable to that of the 3PL CAT. However, unlike the NR CAT results presented previously, selecting items on the basis

Figure 4
RMSE and Bias for NR CAT ($m_i = 3$, NIA = 20)

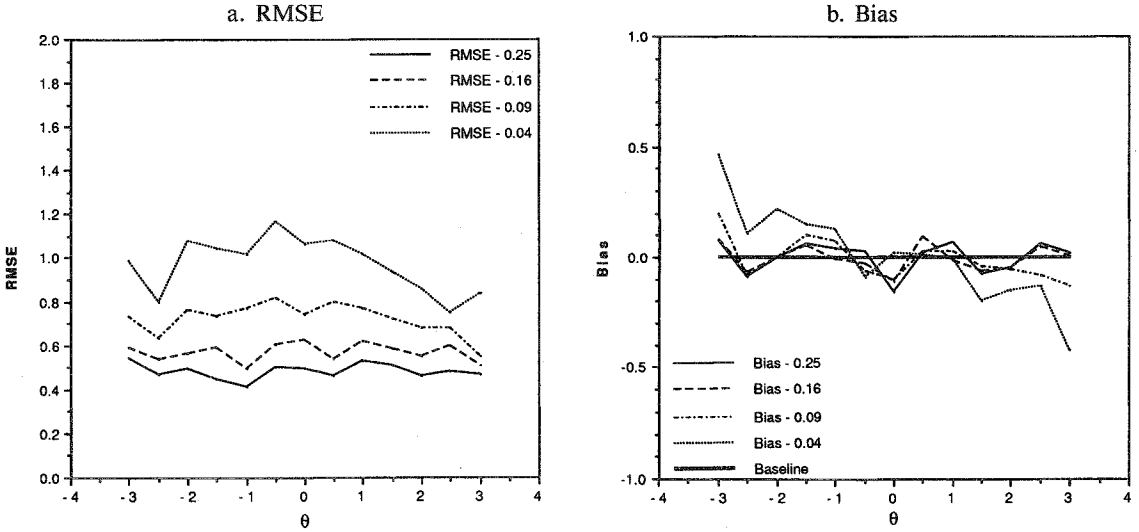


Table 6
Mean and Standard Deviation (SD) of $\hat{\theta}$, Mean, Median, and SD of NIA, and Spearman Rank-Order Correlations Between $\hat{\theta}$ and θ (r) with $\bar{\theta} = 0.0$, $s_\theta = 1.292$ When Items Were Selected by Item Information for NR and 3PL CATs

	$\hat{\theta}$		NIA			
SEE	Mean	SD	Mean	Median	SD	r
3PL CAT						
.30	.168	1.193	12.759	10.000	5.927	.902
.25	.152	1.165	15.073	13.000	6.335	.925
.20	.171	1.164	16.191	13.000	6.879	.928
NR CAT						
.30	.275	1.200	9.682	8.000	5.871	.926
.25	.267	1.190	10.763	9.000	6.472	.926
.20	.269	1.186	12.393	10.000	6.532	.929

of category information did result in the NR CAT administering significantly shorter tests, on average, than did the 3PL CAT for all SEE termination conditions. The post hoc comparison Fs for NIA were all significant at an overall $\alpha = .05$ and were 12.074, 16.225, and 11.357 for the SEE termination criteria of .20, .25, and .30, respectively. Despite this reduction in test length, the NR CAT yielded fidelity coefficients comparable to those of the 3PL CAT (see Table 9).

Table 7
Results of the Repeated Measures ANOVAs for RMSE, Bias, and NIA for NR and 3PL CATs When Items Were Selected by Item Information

Source	SS	df	MS	F	p
RMSE					
Between Persons					
CAT Type	.054	1	.054	.681	.421
Persons Within Groups	1.267	16	.079		
Within Persons					
SEE Term	.022	2	.011	12.584*	0.000
CAT Type × SEE Term	.004	2	.002	2.527	.096
SEE Term × Persons Within Groups	.028	32	.001		
Bias					
Between Persons					
CAT Type	.154	1	.154	.661	.428
Persons Within Groups	3.736	16	.234		
Within Persons					
SEE Term	.001	2	.0005	1.492	.240
CAT Type × SEE Term	.001	2	.0005	.763	.475
SEE Term × Persons Within Groups	.014	32	.0004		
NIA					
Between Persons					
CAT Type	187.638	1	187.638	8.068	.012
Persons Within Groups	372.095	16	23.256		
Within Persons					
SEE Term	85.231	2	42.615	76.371*	0.000
CAT Type × SEE Term	3.455	2	1.728	3.096	.059
SEE Term × Persons Within Groups	17.856	32	.558		

*Significant at overall $\alpha = .05$.

Table 8
Results of the Repeated Measures ANOVA for RMSE, Bias, and NIA for NR and 3PL CATs When Items Were Selected by Category Information for NR CAT and Item Information for 3PL CAT

Source	SS	*df*	MS	*F*	*p*
RMSE					
Between Persons					
CAT Type	.018	1	.018	.203	.658
Persons Within Groups	1.450	16	.091		
Within Persons					
SEE Term	.035	2	.017	9.023	.001
CAT Type × SEE Term	.008	2	.004	2.026	.148
SEE Term × Persons Within Groups	.062	32	.002		
Bias					
Between Persons					
CAT Type	.196	1	.196	.767	.394
Persons Within Groups	4.085	16	.255		
Within Persons					
SEE Term	.004	2	.002	1.206	.313
CAT Type × SEE Term	.007	2	.004	2.416	
SEE Term × Persons Within Groups	.048	32	.001		
NIA					
Between Persons					
CAT Type	335.653	1	335.653	13.883*	.002
Persons Within Groups	386.833	16	24.177		
Within Persons					
SEE Term	102.072	2	51.036	74.531*	0.000
CAT Type × SEE Term	2.123	2	1.062	1.550	.228
SEE Term × Persons Within Groups	21.912	32	.685		

*Significant at overall $\alpha = .05$.

Discussion

In general, the distribution of information was affected by the distance between the item's *b*s, by the order of the *b*s, and by the number of item alternatives. Study 1 showed that for two-, three-, and four-category items, items with an I_{max} value of at least .16 produced reasonably accurate θ

Table 9
Mean and Standard Deviation (SD) of $\hat{\theta}$, Mean, Median, and SD of NIA, and Spearman Rank-Order Correlations Between $\hat{\theta}$ and θ (r) with $\bar{\theta} = 0.000$, $s_\theta = 1.292$ When Items Were Selected by Category Information for the NR CAT and Item Information for the 3PL CAT

	$\hat{\theta}$		NIA			
SEE	Mean	SD	Mean	Median	SD	*r*
3PL CAT						
.30	.168	1.193	12.759	10.000	5.927	.902
.25	.152	1.165	15.073	13.000	6.335	.925
.20	.171	1.164	16.191	13.000	6.879	.928
NR CAT						
.30	.302	1.157	8.121	6.000	4.956	.916
.25	.292	1.170	9.532	8.000	5.116	.918
.20	.259	1.180	11.411	10.000	6.195	.924

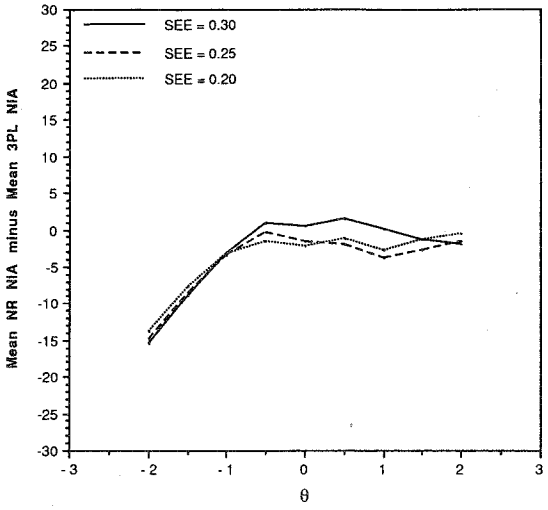
estimation for test lengths of 15 or more items. Shorter length tests required more informative items to maintain reasonable θ estimation.

Results from Study 2 indicated that the NR CAT was able to produce θ estimates comparable to those of the 3PL CAT. To achieve the same level of accuracy (e.g., SEE = .20), the NR CAT administered fewer items, on average, than did the 3PL CAT (e.g., 12.393 versus 16.191, respectively). A plot of the difference in average NIA between the NR and 3PL CATs versus θ (Figure 5) showed that the NR CAT administered substantially fewer average items primarily for examinees with $\theta \leq -1.0$. A relative efficiency comparison of the information content of the item pools of the NR and 3PL CATs showed that although the NR model provided slightly more information than did the 3PL model throughout the θ range, the NR model began to provide substantially more information than the 3PL model below $\theta = -1.0$. However, 3PL item pools can be constructed that are more informative for $\theta < -1.0$ than the item pool used here. The significant NIA results when category information was used for selecting items for the NR CAT might also have resulted from these characteristics of the item pool used.

The present results indicate that a NR model-based CAT can provide θ estimation comparable to a dichotomous model-based CAT. The NR CAT did not provide more accurate $\hat{\theta}$s for examinees with $\theta < 0.0$, relative to the 3PL CAT, because a variable test length was used. That is, the additional information provided by the NR model over a dichotomous model for the lower half of the θ distribution resulted in the adaptive test terminating sooner than it would with the dichotomous model. For a given (reasonable) fixed length test, the NR CAT would be expected to provide more accurate $\hat{\theta}$s for examinees with $\theta < 0.0$ than would a dichotomous model.

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

Figure 5
Mean NIA for NR CAT Minus
Mean NIA for 3PL CAT



References

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

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

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

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

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

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

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

De Ayala, R. J., Dodd, B. G., & Koch, W.R. (1992). A comparison of the partial credit and graded response models in computerized adaptive testing. *Applied Measurement in Education, 5,* 17–34.

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

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

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

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

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

Hays, W. L. (1988). *Statistics.* New York: Holt, Rinehart, & Winston.

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

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

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

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

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

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

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

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 223–237). New York: Academic Press.

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

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.

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

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

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

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

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

Thissen, D. J. (1988). *MULTILOG User's Guide (Version 5.1)* [Computer program and manual]. Mooresville IN: Scientific Software, Inc.

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

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

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

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

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

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

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

Weiss, D. J. (1983). *New horizons in testing: Latent trait*

test theory and computerized adaptive testing. New York: Academic Press.

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.

Acknowledgments

The author thanks William D. Schafer and the editor for constructive and helpful comments.

Author's Address

Send requests for reprints or further information to Ralph De Ayala, University of Maryland, Benjamin Building Room 1230, College Park MD 20742-1115, U.S.A.