Computerized Adaptive Testing With Polytomous Items

Barbara G. Dodd, University of Texas at Austin

R. J. De Ayala, University of Maryland at College Park

William R. Koch, University of Texas at Austin

Polytomous item response theory models and the research that has been conducted to investigate a variety of possible operational procedures for polytomous model-based computerized adaptive testing (CAT) are reviewed. Studies that compared polytomous CAT systems based on competing item response theory models that are appropriate for the same measurement objective, as well as applications of polytomous CAT in marketing and educational psychology, also are reviewed. Directions for future research using polytomous model-based CAT are suggested. *Index terms: computerized adaptive testing, polytomous item response theory, polytomous scoring.*

Computerized adaptive testing (CAT) is one of the major innovations in measurement applications that has benefitted from developments in item response theory (IRT). The advantages of CAT over traditional paper-and-pencil (P&P) tests have been discussed by many researchers (e.g., Kingsbury & Weiss, 1983; McBride & Martin, 1983) and more recently explicated by Wainer, Dorans, Flaugher, Green, Mislevy, Steinburg, & Thissen (1990). The major benefit of CAT derives from procedures designed to administer items that are matched in difficulty level to the examinee's estimated trait level. CATs result in the administration of considerably fewer items and have equal or greater measurement precision than full-length P&P versions of the same tests (McBride & Martin, 1983; McKinley & Reckase, 1980; Weiss, 1982).

Although CAT using polytomous items is the focus of this paper, most CAT implementations to date have been limited to dichotomous items. CAT versions of many tests have been developed from the procedural guidelines recommended for multiple-choice items that are scored dichotomously (Green, Bock, Humphreys, Linn, & Reckase, 1984; Reckase, 1981; Weiss, 1981, 1983, 1985). For example, the Psychological Corporation has published an adaptive version of the Differential Aptitude Test (Henly, Klebe, McBride, & Cudeck, 1989); the College Board has released the Computerized Placement Tests (College Board, 1993); American College Testing has operational math, reading, and writing adaptive tests in their COMPASS program (American College Testing, 1993); and Educational Testing Service has developed an adaptive version of the Graduate Record Examination (Educational Testing Service, 1993). Licensure boards such as the American Society of Clinical Pathologists (Lunz, Bergstrom, & Wright, 1992), the National Council of State Boards of Nursing (Zara, 1988), and the American Board of Internal Medicine (Reshetar, Norcini, & Shea, 1993) have been researching CAT for certification examinations. Based on the results of such research, several licensure boards have implemented CAT versions of their certification tests. The U.S. Department of Defense also has implemented a CAT version of the Armed Services Vocational Aptitude Battery (Curran & Wise, 1994). In addition, school districts such as the Portland Public School District have developed a variety of CATs to administer to students annually (Kingsbury & Houser, 1993).

To date, all CATs that have been implemented on a wide scale in practical settings have been based on dichotomous IRT models that require each item to be scored either correct or incorrect. As the assessment

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 1, March 1995, pp. 5–22

0146-6216/95/010005-18$2.15

field moves away from techniques that are based solely on dichotomously scored multiple-choice items, the use of IRT models that are designed for item responses that are scored using more than two categories should increase. For example, Likert-type attitude scale items are typically scored using an ordered set of response categories. Also, items in mathematics, physics, and chemistry can be designed for partial-credit scoring in which points are awarded for the completion of steps leading to the correct answer. In addition, essay items are typically scored with integers (ordered categories) to represent the degree of quality of the written response.

Polytomous IRT Models

Several polytomous IRT models have been developed over the last 25 years. In this section, polytomous models that have been used in research with CAT as well as some promising models that have received little or no attention in CAT research are discussed. No attempt is made to provide an exhaustive survey of all known polytomous models: the lack of coverage of a particular model does not constitute a judgment on the usefulness or importance of that model.

A useful way to describe these models is to place them within the taxonomy that was developed by Thissen & Steinberg (1986). This taxonomy consists of a five-category classification scheme for grouping dichotomous and polytomous IRT models. Three of the classification categories identified for polytomous models include: difference models, divide-by-total models, and left-side added divide-by-total models. The third category is based on a nominal class of models with the addition of parameters for a latent response category for modeling examinees who are "totally undecided" (Lord, 1983; cited in Thissen & Steinberg, 1984) as to which item response they should select. The multiple-choice model (Thissen & Steinberg, 1984), Samejima's (1969) multiple-choice model, and Sympson's (1983) Model 6 are included in the left-side added divide-by-total category. Given the fact that none of the models in this classification category has been used for CAT, they are not discussed further.

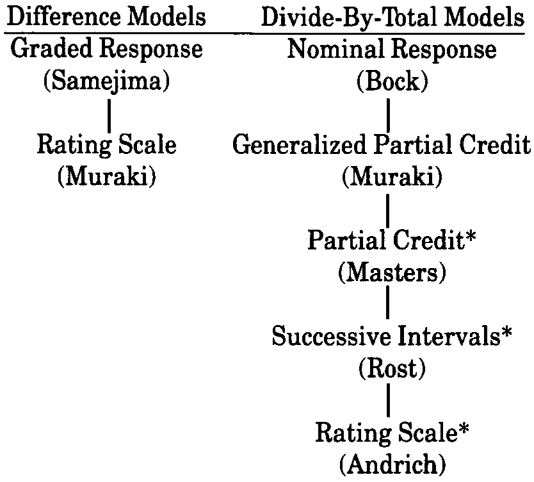
The difference models and divide-by-total models are summarized in Figure 1. The models within each category are arranged so that the most general model is listed at the top of the figure and the most constrained or simplistic models appear at the bottom. The line that connects two models indicates that, by imposing certain constraints on the upper model, the lower model can be obtained. De Ayala (1993) provided a nontechnical introduction to these models. For an in-depth explanation of these models, the reader is referred to the citations that accompany each model.

Difference Models

In difference models, subtraction is used to obtain the probability of a response in a particular category. Samejima's (1969) graded response model (GRM) and Muraki's (1990) rating scale model (MRSM) are two examples of difference models. For both the GRM and MRSM, the probability of responding in a particular category is calculated by subtracting the probability of responding in a given category or higher (conditional on trait level, θ) from the probability of responding in the adjacent or lower category (conditional on θ). The equation describing the probability of responding in each category is referred to as the operating characteristic function (OCF) for the model.

The graded response model. The GRM is appropriate when responses to an item can be classified into more than two ordered categories to represent varying degrees of attainment of a solution to a problem or agreement with an attitude statement. The responses to item i are classified into $(m_i + 1)$ ordered categories so that lower-numbered categories represent less of the trait measured by the item than do higher-numbered categories. The category scores for item i are successive integers, denoted x, where $x = 0, 1, ..., m_i$. Samejima (1989) developed a two-stage process to obtain the probability that a given individual with a certain θ level will receive a given category score. In the first stage, the probability that an individual will

Figure 1
Hierarchy of Polytomous IRT Models
(*Model is a member of the Rasch family of IRT models)



receive a category score of x or higher on item i is expressed by

$$P^*_{ix}(\theta) = \frac{\exp\left[\mathrm{D}a_i(\theta - b_{ix})\right]}{1 + \exp\left[\mathrm{D}a_i(\theta - b_{ix})\right]}, \tag{1}$$

where

D is the scaling constant 1.7,
a_i is the discrimination parameter of item i,
θ is the trait level, and
b_{ix} is the category boundary associated with category x for item i.

For each item, one a_i term and a set of m_i category boundaries are estimated.

The second stage in obtaining the probability of responding in a particular category, $P_{ix}(\theta)$, involves the subtraction of the cumulative probabilities for adjacent categories conditional on θ:

$$P_{ix}(\theta) = P^*_{ix}(\theta) - P^*_{i,x+1}(\theta). \tag{2}$$

In order to use Equation 2 to obtain the probability of responding in either of the two extreme categories, it is necessary to define the probability of responding in the lowest category or higher, $P^*_{i0}(\theta)$, as 1.0 and the probability of responding in category $m_i + 1$ or higher, $P^*_{i,m+1}(\theta)$, as 0. Equation 2 is the OCF for the GRM.

The Muraki rating scale model. Muraki (1990) demonstrated that the MRSM is a restricted case of the GRM for attitude scales. Muraki reparameterized the category boundary parameters (b_{ix}) of the GRM to include a location parameter for the item (b_i) and a set of threshold parameters for the scale (t_x). With the MRSM, the probability of an examinee with a given θ responding in category x or higher on item i is defined as

$$P^*_{ix}(\theta) = \frac{\exp\left[\mathrm{D}a_i(\theta - b_i + t_x)\right]}{1 + \exp\left[\mathrm{D}a_i(\theta - b_i + t_x)\right]}. \tag{3}$$

Therefore, under the MRSM an item with $m_i + 1$ categories is characterized by its location on the scale (b_i), its discrimination power (a_i), and a set of m_i thresholds (t_x) for the entire scale. The restriction that the t_x parameters be constant across items is consistent with the common practice of using a common rating scale for all items. As is the case with the GRM, the probability of responding in category x on item i is obtained by subtracting the adjacent $P_{ix}^*(\theta)$ functions. When an item has only two categories (incorrect and correct), the GRM and MRSM reduce to the two-parameter model (Samejima, 1969). As a result, the GRM may be applied to tests that include both dichotomously and polytomously scored items.

Divide-by-Total Models

Unlike the difference models, the OCF is obtained directly in the divide-by-total models. For these models, the probability of responding in a given category is obtained by dividing the numerator by the sum of all category probability numerators so that the probabilities conditional on θ sum to unity.

The nominal response model. The nominal response model (NRM) developed by Bock (1972) is the most general model in the divide-by-total model classification category, and all the models discussed in this section are restricted cases of the NRM. In contrast to the difference models, the NRM may be applied to items that have alternatives that cannot be ordered to represent varying degrees of the trait measured by the item. The NRM usually is used with multiple-choice items in which it is difficult to order distractors according to their relative degree of correctness or to the relative amount of knowledge required to recognize that each alternative is incorrect. The NRM attempts to increase the precision of the θ estimates for individuals, particularly those with low θ levels, by using information from their incorrect responses.

For the NRM, the probability of an examinee with a given level of θ responding in category x on item i is defined as

$$P_{ix}(\theta)=\frac{\exp[c_{ix}+a_{ix}\theta]}{\sum_{h=1}^{n_i}\exp[c_{ih}+a_{ih}\theta]}, \qquad (4)$$

where

a_{ix} is the slope (discrimination) parameter for category x of item i,

c_{ix} is the intercept parameter of the nonlinear response function associated with category x of item i, and

n_i is the number of categories of item i (i.e., $x = 1, ..., n_i$).

Therefore, in the NRM each category's ability to discriminate among examinees is captured by a_{ix}. c_{ix} reflects the interaction between the difficulty of the category and how well that category discriminates. As a result, the description of an item consists of n_i discrimination and intercept parameters. When an item has only two options (scored as incorrect and correct), the NRM reduces to the two-parameter logistic model (2PLM), where the item difficulty parameter is equal to $-c_{ix}$ divided by a_{ix}.

The partial credit model. Thissen & Steinberg (1986) showed that by imposing the constraint that the NRM slope parameters increase in steps of 1.0, it is possible to apply the NRM to ordered category scores for an item and obtain Master's (1982) partial credit model (PCM). Similar to the GRM, the PCM is appropriate for items that are scored in a graded fashion. The examinees' responses are categorized into $m_i + 1$ scores (i.e., $x = 0, 1, ..., m_i$) to represent varying degrees of the trait measured by item i. Masters proposed the following general expression for the probability that an examinee with a given θ will obtain a category score of x on item i

$$P_{ix}(\theta)=\frac{\exp\left[\sum_{k=0}^{x}(\theta-b_{ik})\right]}{\sum_{h=0}^{m_i}\exp\left[\sum_{k=0}^{h}(\theta-b_{ik})\right]}, \qquad (5)$$

where b_{ik} is the item step difficulty parameter associated with the transition from one category to the next and there are m_i step difficulties for item i. For notational convenience, Masters defined $\Sigma(\theta - b_{ik})$ as being equal to 0.0 when k is 0.

The PCM was conceptualized under the requirement that the steps within an item be completed in order, although the steps need not be ordered in terms of difficulty (e.g., Step 2 can be easier than Step 1). When the steps are not ordered in terms of difficulty, a reversal (Dodd & Koch, 1987) is said to exist. The PCM also assumes that all items are equally effective in discriminating among examinees with varying θ levels. When an item is scored dichotomously, the PCM reduces to the Rasch model and thus can be used with tests composed of both dichotomously and polytomously scored items.

The Andrich rating scale model. When the PCM is applied to a set of Likert-type items that share a fixed set of rating points, the PCM may be simplified to obtain Andrich's (1978a) rating scale model (ARSM). Masters & Wright (1984) showed how each item step difficulty parameter from the PCM could be decomposed into two components

$$b_{ik} = b_i + t_k \,, \tag{6}$$

where b_i is the location (i.e., scale value) of item i, and t_k is the threshold parameter for the kth category over the entire set of items. By substituting Equation 6 into Equation 5, simplifying, and letting

$$K_x = -\sum_{k=1}^{x} t_k \tag{7}$$

and $K_0 = 0.0$, the ARSM is derived from the PCM. Andrich (1978a, 1978b) defined the probability that a person with a given θ level will respond in category x to item i as

$$P_{ix}(\theta) = \frac{\exp\left[K_x + x(\theta - b_i)\right]}{\sum\limits_{h=0}^{m_i} \exp\left[K_h + h(\theta - b_i)\right]} \,, \tag{8}$$

where K_x is the negative sum of the thresholds passed.

Similar to the MRSM, the t_ks are estimated for the entire item set, whereas the item scale values (b_i) are estimated individually for each item (Andrich, 1978a). As is the case with the PCM and unlike the MRSM, the ARSM assumes that items are equally effective at discriminating among examinees.

The successive intervals model. Rost (1988) developed the successive intervals model (SIM), which is another polytomous Rasch model that is appropriate for attitude measurement. The probability that a person with a given θ level will respond in a particular category for an item may be expressed as

$$P_{ix}(\theta) = \frac{\exp\left\{K_x + x\theta - \left[xb_i + x(m - x)d_i\right]\right\}}{\sum\limits_{h=0}^{m_i} \exp\left\{K_h + h\theta - \left[hb_i + h(m_i - h)d_i\right]\right\}} \,, \tag{9}$$

where

b_i is the scale value (location parameter) for item i,

d_i is the dispersion parameter for item i, which reflects the degree to which the threshold distances for the item deviate from the threshold parameters for the entire scale, and

K_x is the negative sum of the threshold parameters associated with Categories 1 to x.

For notational convenience, t_0 is defined as being equal to 0.0 so that Equation 9 also can be used to obtain the probability of responding in category 0.

As is the case for the ARSM, the SIM is a special case of the PCM and estimates a scale value or item location parameter for each item, as well as a single set of response threshold values for the entire set of items. Unlike the ARSM, however, the SIM contains a second item parameter, d_i, for each item that reflects the degree of difference between the threshold distances for the item and the threshold distances for the entire scale. By imposing the restriction that all d_i values equal 0.0, the SIM simplifies to the ARSM.

The generalized partial credit model. Muraki (1992) extended the PCM by removing the assumption that all items discriminate equally well. The generalized PCM (GPCM) was developed in a fashion parallel to the development of the PCM by substituting the two-parameter model for the Rasch model. The GPCM is expressed as

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{k=0}^{x} a_i(\theta - b_{ik})\right]}{\sum_{h=0}^{m_i} \exp\left[\sum_{k=0}^{h} a_i(\theta - b_{ik})\right]}, \tag{10}$$

where

$P_{ix}(\theta)$ is the probability of an examinee with a given θ responding in category x of item i with $m_i + 1$ categories,

a_i is the item discrimination, and

b_{ik} is the step difficulty parameter associated with category k ($k = 1, ..., m_i$).

Muraki defined $\Sigma(\theta - b_{ik})$ as being equal to 0.0 when k is 0. Similar to the PCM, the b_{ik} terms are not necessarily ordered and, therefore, reversals may occur. When a_i equals 1.0, the GPCM simplifies to the PCM. By further assuming that b_{ik} can be split into its component parts–the item's location (b_i) and the threshold parameters for the entire scale (t_k)–the GPCM becomes the ARSM. Muraki (1992) also demonstrated that the GPCM is a special case of the NRM for ordered response categories.

Information

In contrast to dichotomous models in which the concept of information is defined at the item level, the information function for polytomous models may be estimated for each response category as well as for the item. Samejima (1969) proposed information functions for polytomous items that are applicable to all of the models discussed here. Although other formulas for information that are computationally simpler have been derived for specific models, Samejima's formulation is presented because of its generality.

Samejima (1969) defined the category information function $[I_{ix}(\theta)]$ for item i as

$$I_{ix}(\theta) = \frac{[P'_{ix}(\theta)]^2}{[P_{ix}(\theta)]^2} - \frac{P''_{ix}(\theta)}{P_{ix}(\theta)}, \tag{11}$$

where $P_{ix}(\theta)$ is the probability of obtaining a category score of x for a fixed θ, and $P'_{ix}(\theta)$ and $P''_{ix}(\theta)$ are the first and second derivatives of $P_{ix}(\theta)$, respectively. Samejima (1969) defined the item information $[I_i(\theta)]$ for a polytomous item as

$$I_i(\theta) = \sum_{x=0}^{m_i} I_{ix}(\theta) P_{ix}(\theta). \tag{12}$$

Substituting the equality from Equation 11 into Equation 12 and simplifying yields

$$I_i(\theta) = \sum_{x=0}^{m_i} \frac{[P'_{ix}(\theta)]^2}{P_{ix}(\theta)} - \sum_{x=0}^{m_i} P''_{ix}(\theta). \tag{13}$$

Samejima (1969) demonstrated that the second term in Equation 13 equals 0.0 and thus can be deleted from the equation for the item information function. The test or scale information function is simply the sum of the item information functions. It should be noted that polytomous scoring of an item provides more information than dichotomous scoring of the item (Samejima, 1969); this is true of all the polytomous models discussed here.

Model Selection

The selection of a particular polytomous model involves a number of factors: the type of data, model-data fit, philosophical considerations, model assumptions, and parsimony. If the data consist of items that have unordered alternatives, then the NRM is appropriate. When responses to an item are classified into more than two categories that can be ordered to represent varying degrees of the trait measured by the item, then either the GRM, GPCM, or the PCM could be used. If the ordered data are ratings, then more constrained versions of these models, such as the MRSM, the SIM, or the ARSM, would be appropriate.

From a statistical perspective, a likelihood ratio for fit to the data using a general model (e.g., the GPCM) could be obtained; then, by imposing constraints on the general model, the PCM and its respective likelihood ratio could be determined. The difference between the likelihood ratios could be tested for significance, and the simplest model that does not differ significantly in fit compared to the more complicated model would be used. In a similar fashion, the SIM could be compared with the ARSM. This approach, or some similar fit analysis method, can be extended to include a cross-validation sample (e.g., see Drasgow, Levine, Tsien, Williams, & Mead, 1995).

An alternative approach is to use the ideal observer index (IOI) (Levine, Drasgow, Williams, McCusker, & Thomasson, 1992). The IOI allows statistical models to be compared on the basis of response vector probabilities. Using the IOI, competing models can be compared in terms of correct classification rates. In contrast to the likelihood ratio approach described above, the IOI allows comparison of nonhierarchical models.

Both the likelihood ratio approach and the IOI approach ignore philosophical differences among families of models, such as the Rasch family (PCM, ARSM, SIM); differences concerning the interpretation of model-data fit (e.g., whether the model should be selected to fit the data or the data should be selected to fit the model); the fact that some models may not be identified; and particular model characteristics. For example, the PCM assumes that items have equal discriminations; however, some individuals may prefer to use models that allow items to vary in terms of discrimination, such as the GPCM or GRM.

Research on CAT With Polytomous Items

Basic research has investigated a variety of possible operational procedures for CAT based on one of the difference models (GRM) and four of the divide-by-total models (NRM, PCM, ARSM, and SIM). The general procedural guidelines that have emerged from these studies are discussed in terms of the major components of a CAT system. Studies comparing the performance of CAT systems based on competing IRT models that are appropriate for the same measurement objective and several live-testing applications of polytomous CAT procedures are discussed after the basic research on operational procedures. Table 1 presents a brief description of the polytomous CAT research studies.

Operational Procedures Research

There are four major components of an adaptive test: (1) the item bank, (2) the item selection procedure, (3) the trait estimation procedure, and (4) the stopping rule (Kingsbury & Zara, 1989, 1991; Reckase, 1989; Wainer et al., 1990; Weiss, 1982). These four components are common to all CATs regardless of the IRT model used.

Item bank. The size of the item bank and the characteristics of the items included in the bank can im-

Table 1
Summary of Polytomous CAT Research

Model(s) and Reference	Item Banks	Type of Data	Variables	Findings
Partial Credit				
Koch & Dodd (1985)	39 items 24 items 50 items	real Likert-type attitude real Likert-type attitude simulated attitude	information functions; stopping rules: 10- or 20-item fixed length vs. variable length	item bank information functions were moderately peaked; 20-item fixed length tests had the lowest standard errors for θ estimates
Partial Credit				
Koch & Dodd (1989)	30, 60, and 120 items with both peaked and flat item information functions	simulated achievement test items scored 0 to 3	peaked vs. flat item information functions; item bank size; fixed vs. variable stepsize for θ estimation	peaked information functions yielded CAT θ estimates with lower standard errors; fewer items were administered with the 30-item banks; variable stepsize performed better than fixed stepsize
Partial Credit				
Dodd et al. (1993)	30 and 60 items with peaked, bimodal, positively and negatively skewed item bank information functions	simulated achievement test items scored 0 to 3	item bank size; distribution of item difficulty (information); stopping rules	30-item banks were sufficient for peaked information banks but were problematic for skewed information banks; minimum prespecified standard error stopping rule was recommended
Graded Response				
Dodd et al. (1989)	30 and 60 items	simulated achievement test items scored 0 to 4	item bank sizes; fixed vs. variable stepsizes for θ estimation; stopping rules	30-item banks were adequate; variable stepsize for preliminary θ estimates performed best; minimum standard error stopping rule performed best
Graded Response				
Singh et al. (1990)	12 items	real Likert-type marketing survey	CAT efficiency; live field tests of CAT feasibility; branching among scales	25% to 50% reduction in test length for CAT compared to conventional paper-and-pencil marketing surveys
Singh (1993)	12, 18, and 24 items	real Likert-type attitude		
Nominal Response vs. Three-Parameter Logistic				
De Ayala (1989)	multiple choice 50 items	mathematics achievement test items; real response	model comparison; stopping rules; relative efficiency	67% reduction in test length; nominal model was more informative than three-parameter logistic for low θ examinees; both models performed equally well for CAT
Nominal Response vs. Three-Parameter Logistic				
De Ayala (1992)	90 items and 150 items	simulated multiple-choice items	RMSE and bias of θ estimates; number of item response categories (from 2 to 4); item selection methods	nominal and three-parameter CATs did not differ in RMSE or bias; nominal CATs administered fewer items than three-parameter CATs

continued on next page

Table 1, continued
Summary of Polytomous CAT Research

Model(s) and Reference	Item Banks	Type of Data	Variables	Findings
Andrich's Rating Scale				
Dodd (1990)	39 items 24 items 32 items	real Likert-type attitude real Likert-type attitude simulated attitude	item selection based on scale value vs. item information function; fixed vs. variable stepsizes for θ estimation	item banks of 25 to 30 items are adequate; variable stepsize method is preferable; recommended use of scale value to select items and use of minimum standard error stopping rule
Andrich's Rating Scale				
Koch et al. (1990)	40 items	field test of real Likert-type attitude items	CAT efficiency; students' attitudes toward CAT	item bank information was peaked; 60% reduction in test length; students' attitudes were very favorable toward CAT
Andrich's Rating Scale				
Dodd & De Ayala (1994)	39 items 24 items 32 items	real Likert-type attitude real Likert-type attitude simulated attitude	item selection using scale values vs. item information functions	both methods performed equally well for CATs
Successive Intervals				
Koch & Dodd (in press)	61 items and 30 items with random, large, or small item dispersion parameters; 39 items	simulated Likert-type attitude items real Likert-type attitude	size of item banks; types of item dispersion parameters; item selection using scale values vs. item information	the 61-item bank produced CATs that were 3 items shorter than the 30-item banks; the types of dispersion parameters did not affect the CAT results; scale value item selection was recommended
Graded Response vs. Andrich's Rating Scale				
Dodd et al. (1988)	40 items 25 items 30 items	real Likert-type attitude real Likert-type attitude simulated attitude	model comparison; CAT θ estimates vs. full-scale θ estimates	rating scale CATs outperformed graded response CATs based on accuracy of CAT θ estimates; graded response CATs administered fewer items than rating scale CATs; skewed item pool information function adversely affected graded response CATs
Graded Response vs. Partial Credit				
De Ayala et al. (1992)	150 items; 3 different graded response item banks; 5 different partial credit item banks	simulated achievement test items scored 0 to 4	robustness of graded response and partial credit CATs to the inclusion of misfitting items into the CAT item pool; model comparison	even with 45% of the items misfitting, the partial credit CATs produced accurate θ estimates; graded response model fit 97% of the items; graded response CATs outperformed the partial credit CATs

pact the properties of an adaptive test. With dichotomously scored items, it has been recommended that an item bank consist of at least 100 items for use with the three-parameter logistic model (3PLM; Urry, 1977). Considerably larger item banks (e.g., 500 to 1,000 items) may be desirable if content balancing and fixed length CATs are used for high-stakes testing, such as certification examinations, in which security of the item bank is a concern.

One of the problems that frequently occurs with small item banks is the nonconvergence of θ estimates to a finite value when maximum likelihood estimation is used. With nonconvergence, even if finite estimates can be obtained, they are typically accompanied by very large standard errors. This is typically due to insufficient item information to estimate the given θ at certain points within the item bank. Therefore, approximately uniform (rectangular) distributions of item difficulty values for the items in the bank have been recommended for CAT systems based on dichotomously scored items (Reckase, 1981; Urry, 1977). When the item difficulty values are uniformly distributed, the item bank information function will be relatively constant across the range of θ with a moderate peak at $\theta = 0.0$ (Dodd, Koch, & De Ayala, 1989).

In contrast to the recommendations for dichotomously scored items, research on polytomous CAT based on various models has shown that substantially smaller item banks may be used successfully. This research has found that item banks with 30 items may be sufficient for accurate θ estimation, with few nonconvergence problems, for the GRM (Dodd et al., 1989), PCM (Dodd, Koch, & De Ayala, 1993; Koch & Dodd, 1989), SIM (Koch & Dodd, in press), and ARSM (Dodd, 1987, 1990; Dodd & De Ayala, 1994). Several of the studies in the context of Likert-type attitude measurement have found that item banks with as few as 24 items have worked very well for the PCM (Koch & Dodd, 1985) and the ARSM (Dodd, 1990; Dodd & De Ayala).

However, these findings do not imply that any item bank composed of 30 or more items will be sufficient for CAT based on polytomous IRT models. The characteristics of the individual items that comprise the item bank have an impact on the success of any CAT system. Dodd et al. (1993) found that an item bank of 30 items worked well for a CAT based on the PCM, if the item bank information function was moderately peaked at a point close to $\theta = 0.0$ or if the total information function was bimodal. Skewed item bank information functions with predominantly easy or difficult items, however, proved problematic for item banks of only 30 items. In addition, pragmatic issues concerning content validity, item exposure, and test security for high stakes testing may require considerably larger item banks.

The finding that relatively small item bank size works well for polytomous CAT is due to the fact that the information provided by a polytomous item is considerably more than that provided by a dichotomously scored item. Not only is the modal level of information higher, but the information is typically distributed across a wider range of the trait being measured. In essence, each pair of adjacent categories in the polytomous item serves as a single dichotomous item and thus the set contributes more to the total item bank information function than the typical dichotomously scored item (Dodd, 1987; Dodd & De Ayala, 1994; Dodd & Koch, 1994; Koch, 1983).

A major limitation of much of the polytomous CAT research to date is that the item banks have been simulated rather than real. The advantage of simulated item banks is that the known parameters can be manipulated systematically to investigate basic variables of interest, but much more research is needed with field tests of real items and real examinees.

Item selection procedure. The goal of item selection in CAT is to administer the next unused item remaining in the item bank that provides the most information at the examinee's current θ estimate. To achieve this, most CAT systems use item information functions as the basis for item selection. For the polytomous models that have been studied for CAT—the GRM (Dodd et al., 1989), the NRM (De Ayala, 1989, 1992), and the PCM (Dodd et al., 1993; Koch & Dodd, 1985, 1989)—CATs have performed very well using item information to select the next item for administration. De Ayala (1992) found that using category information, rather than item information, as the item selection procedure resulted in one less item, on average, being administered for NRM CAT simulations. To date, no other studies have investigated the use of category information for item selection in CAT. Other than information (item or category), no other item selection procedure has been investigated for these models because they contain no location (scale value) parameter.

An alternative item selection procedure has been studied for the ARSM (Dodd & De Ayala, 1994) and SIM (Koch & Dodd, in press). Because both of these models contain a scale value item parameter for each item that represents the location of the item along the θ continuum, the two studies compared the method of selecting the item with the closest scale value to the current θ estimate with the maximum item information selection procedure. Although the item information function for the ARSM (Dodd & De Ayala) and the SIM (Dodd & Koch, 1994) does not peak precisely at the scale value of the item, it was found that selecting items for administration based on the closest scale value did not substantially diminish the performance of the CATs relative to the maximum item information selection procedure. Dodd & De Ayala and Koch & Dodd (in press) recommended the scale value item selection procedure for these models over the maximum item information selection procedure because it is much simpler and requires much less computation time.

Item selection procedures for polytomous CAT have not been studied under conditions in which it is necessary to ensure content balancing of the items presented during the CAT. Selecting items based strictly on item information or item scale values will frequently impact content validity negatively. More studies are needed that investigate item selection strategies in realistic contexts.

Trait estimation procedure. For CATs based on dichotomous IRT models, θ levels can be estimated with either a maximum likelihood method or with one of several Bayesian methods. To date, maximum likelihood has been the only trait estimation method used in CATs based on the PCM (Dodd et al., 1993; Koch & Dodd, 1985, 1989), the ARSM (Dodd, 1987, 1990; Dodd & De Ayala, 1994), the SIM (Koch & Dodd, in press), and the GRM (Dodd et al., 1989).

With maximum likelihood estimation, no maximum likelihood estimate is possible after the administration of the first item if the examinee responds in either the lowest or highest category. However, a maximum likelihood estimate can be calculated after only one item if the examinee responds in any category other than the two extreme categories. Because such an estimate will be very unstable and will have a high standard error associated with it, all of the research to date on polytomous CAT has used a systematic procedure to estimate a preliminary θ level based on either a fixed or variable stepsize until the examinee receives item scores in two different categories, as an alternative to maximum likelihood estimation. With a fixed stepsize, the new θ estimate is increased or decreased by a prespecified amount (e.g., .4 or .7) depending on whether the response to the previously administered item was in the upper or lower half of the response scale.

With a variable stepsize, the new θ estimate is set halfway between the current θ estimate and one of the two most extreme item parameter estimates in the item bank. Whether the highest or lowest item parameter is used depends on the individual's response to the previously administered item. If the individual responded in the upper half of the response scale, then the highest item parameter is used. If the response is in the lower half of the response scale, then the lowest item parameter is used. The particular item parameter estimate in the variable stepsize procedure depends on the particular IRT model that is being used in the CAT. Although the extreme step values are used for the PCM, the extreme category boundaries are used for the GRM.

Both fixed and variable stepsize methods have been investigated to measure their impact on the operational characteristics of CATs using the PCM (Koch & Dodd, 1989), the ARSM (Dodd, 1990), and the GRM (Dodd et al., 1989). In those studies, the use of the variable stepsize outperformed the fixed stepsize procedure (i.e., there were fewer cases of nonconvergence of the θ estimate with the variable stepsize procedure). CAT systems based on the ARSM and the SIM use the extreme scale value parameters to update the current θ estimate prior to maximum likelihood estimation.

Only two studies have used Bayesian methods to estimate θ during a CAT based on a polytomous model. De Ayala (1992) used expected a posteriori (EAP) estimation (Bock & Mislevy, 1982) in a CAT based on the NRM. Chen, Hou, Fitzpatrick, & Dodd (1995) compared EAP and maximum likelihood estimation procedures in CAT based on the ARSM. One advantage of EAP over maximum likelihood is that estimates still can be obtained for individuals who respond in either the lowest or highest category score for every item. Another advantage of EAP estimation is that the mean squared error associated with the estimator across the population of θ levels is smaller than that associated with maximum likelihood estimates (Bock & Mislevy).

Stopping rule. In addition to specifying some minimum/maximum static stopping rule (e.g., fixed test length), two different dynamic stopping rules have been studied in research on polytomous CATs. The minimum information stopping rule terminates the CAT when no remaining item in the bank has a prespecified minimum level of item information given the examinee's current θ estimate. The second stopping rule that has been used terminates the CAT when the standard error associated with the current θ estimate falls below a prespecified level. In these studies, if the specified stopping rule was not met after a given number of items had been administered (usually 20), the CAT was terminated. Comparisons of these two stopping rules in CATs based on the GRM (Dodd et al., 1989), the PCM (Dodd et al., 1993), the ARSM (Dodd, 1990), and the NRM (De Ayala, 1989) revealed that using the standard error stopping rule was superior to the minimum item information rule in terms of the mean number of items administered, frequencies of nonconvergence of θ estimates, and correlations of CAT θ estimates with full-scale calibration θ estimates and known θ levels.

Two studies used a static stopping rule that terminated the CAT when a prespecified number of items had been administered. De Ayala (1992) employed a fixed test length of 30 items to investigate a variety of operational characteristics of CAT based on the NRM. Koch & Dodd (1985) also used fixed CAT lengths in their initial investigation of the operational characteristics of CAT based on the PCM. In general, dynamic stopping rules result in more efficient use of the item bank in terms of item exposure and development cost than fixed-length stopping rules (Kingsbury & Houser, 1993).

Comparison Studies

To date, only four studies have compared the performance of CATs based on competing polytomous IRT models. De Ayala (1989, 1992) compared CATs based on the NRM and the 3PLM in the context of achievement testing. Maximum likelihood estimation was used for θ estimation in the 1989 study and Bayesian estimation was used in the 1992 study. Although both studies revealed that the two models performed equally well, considerably fewer items were administered by the NRM CAT than the 3PLM CAT. This is because the NRM provides more information than the 3PLM for low θ level examinees.

De Ayala, Dodd, & Koch (1992) compared the PCM and GRM. The purpose of their study was to determine the impact on a CAT of including misfitting items. The results showed that, although the GRM had substantially better fit to more items, the CAT based on the PCM produced θ estimates that were as accurate as those produced by the GRM even though 45% of the items in the PCM item bank had poor fit to the model.

Dodd, Koch, & De Ayala (1988) compared the ARSM and the GRM in terms of CAT attitude measurement for both real and simulated datasets. The ARSM CAT outperformed the GRM CAT in terms of the accuracy of the CAT θ estimates relative to the full scale (all items) calibration θ estimates and known θs used to generate the data. The GRM CAT did not perform as well as the ARSM CAT when the item bank information function was skewed. The authors suggested that the ARSM CAT be used rather than the GRM CAT for attitude measurement because (1) the ARSM CAT procedures showed less degradation of performance when the item bank information function was skewed, (2) there are fewer item parameters to estimate with the ARSM, and

(3) it is often reasonable to assume approximately equal discriminations for all items in a conventionally constructed Likert-type scale.

The De Ayala et al. (1992) and Dodd et al. (1988) comparison studies illustrated the need to research the properties of the item bank thoroughly before implementing a real CAT system, because the characteristics of the item bank can have a profound effect on the performance of CAT. Researchers also should investigate the appropriateness of competing IRT models prior to implementing any CAT system. If several competing models perform equally well, the researcher should consider parsimony when selecting an IRT model. More research also is needed to compare the entire array of polytomous models that are appropriate for the same measurement problem for CAT applications in a wide variety of measurement situations.

Applications

Five CAT applications were found in the marketing literature. Kamakura & Balasubramanian (1989) demonstrated the potential usefulness of CAT procedures in marketing research by using real data collected on a personality measure to simulate a CAT based on Birnbaum's (1968) 2PLM. In a second study, Balasubramanian & Kamakura (1989) used the same CAT procedures as in their other study to implement two live adaptive marketing surveys. Both studies demonstrated that CAT based on the 2PLM could save considerable time and money for those individuals working in marketing research and survey fields. The only drawback to these two studies was that polytomous responses to each item had to be dichotomized in order to use the 2PLM. Polytomous IRT models could have used all the information in the response scale to each item and thus most likely would result in even more efficient measurement.

Another application in the marketing research area was a real-data simulation study of CAT based on the GRM by Singh, Howell, & Rhoads (1990). Although they used a Likert-type scale of consumer discontent, their study was limited by their extremely small item bank of only 12 items.

In a live-testing study of a 14-item locus of control instrument, Singh (1993) found a 25% test length reduction and a 25% reduction in administration time for a CAT based on the GRM relative to a P&P administration. In a second live-testing study, using three Likert-type marketing scales that contained 24, 18, and 12 items, respectively, Singh (1993) explored the use of bivariate and multiple regression to branch between the three correlated scales, based on a method originally proposed by Brown and Weiss (1977) for use with dichotomous items. He found that using θ estimates from the scales administered first substantially reduced the overall length of the CAT based on the GRM. He observed a 50% reduction in administration time and test length when compared to computer administration of all items from the three scales. Singh concluded from his field trials that adaptive surveys are not only practical but facilitate the quantity-quality tradeoffs that occur in marketing research.

To date, only one real-data polytomous CAT application has been conducted outside of the marketing field. Koch, Dodd, & Fitzpatrick (1990) measured 111 students' attitudes toward alcohol with a CAT using the ARSM and a P&P version of a 40-item attitude scale. Students took the P&P version first, followed by the CAT two weeks later. There were no cases of nonconvergence of θ estimation in the CATs. A correlation of .89 was found between the scores obtained from the CAT and P&P versions. On average, there was a 64% reduction in test length under the CAT condition when compared to the P&P version. A survey of the students' reactions to the CAT revealed that the majority of the students found it more interesting to take the CAT version of the attitude scale than the P&P version, and they preferred the CAT version over the P&P version. The students also thought that CAT procedures would result in more honest answers than either a personal interview or P&P procedure. This suggests that attitude surveys that deal with sensitive issues might benefit from using CAT procedures.

The major limitation of the applications of polytomous CATs is that there have been so few studies using real items with real examinees. Furthermore, the few studies conducted have been restricted to attitude measurement. Substantial real-data research is needed to assess the degree to which the promising results from simulations transfer to live testing.

Directions for Future Research

The research on polytomous model-based CAT is in its early stages, just as dichotomous model-based CAT was in the late 1970s and early 1980s. Much of the CAT research that still needs to be conducted with polytomous models mirrors the research that has been conducted on dichotomous model-based CAT. Although the research that has been conducted suggests promise for this type of CAT, much basic research still needs to be performed before CATs based on polytomous IRT models are implemented on a wide scale.

Operational procedures for CAT based on polytomous models such as the GPCM and MRSM need to be studied. Bayesian methods of θ estimation need to be explored for various polytomous models and compared to the maximum likelihood estimation methods that are currently being used. CATs based on competing IRT models that are appropriate for the same measurement objective need to be compared. Researchers should evaluate the performance of the CATs in terms of precision of measurement, number of items administered, bias of the θ estimate, and parsimony. The use of multiple scoring schemes within a CAT system warrants investigation. For example, some items may be scored in a graded fashion whereas others may be nominally or dichotomously scored. It should be straightforward to develop CATs based on several hierarchical polytomous models that allow items within a test to be scored differently. Although research that integrates multiple nonhierarchical polytomous models in CAT will be considerably more difficult to conduct, it could potentially produce CATs that would be better suited for certain types of tests than hierarchical polytomous models.

The impact of mode of presentation and content balancing needs to be assessed. Mode of presentation could have an important effect on polytomous CAT if the items require multiple screens to present the stimuli. In that case, using item responses from P&P versions of the test or instrument to calibrate the item bank may be inappropriate. In addition, the need for content balancing might increase the item bank size requirements considerably above the minimal level of 30 items used in simulations. For example, the table of specifications for a mathematics CAT could require content balancing that could drastically increase the item bank size requirements in order to have adequate content representation in the CAT and yet not have problems with over-exposure of the item bank for test security reasons.

If mode of presentation effects are found for a given measurement instrument, the validity evidence for the P&P version may not be generalizable to the CAT version of the instrument. Even if mode effects are not present, it cannot be assumed that the validity evidence for the P&P version of the instrument will extend to the CAT version of the same instrument. Validity studies will, therefore, be necessary if the CAT version will replace a P&P version of existing tests or instruments.

Equating procedures and item banking techniques should be explored in greater detail. To date, equating and item banking procedures have been investigated only for the GRM (Baker, 1992), the PCM (Masters & Evans, 1986), and the NRM (Baker, 1993). Much more research is needed in this area if adequate item banks are to be available for CATs based on other polytomous IRT models. The issue of equating is particularly crucial if performance assessment is to be implemented using CAT systems based on polytomous IRT models. In general, a few tasks or items are included in a given performance assessment because it usually takes the examinee considerable time to complete each task. As a consequence, an item bank of 30 performance tasks cannot be constructed without item banking and equating procedures. When considering the possibility of CATs based on a polytomous model for performance tasks, the test developer must weigh the increased costs associated with the development of an adequate item bank against the advantages of CAT. It is quite possible that for certain item types, the cost of item development might prohibit the use of CAT.

Before polytomous CATs are implemented on a wide scale in practical settings, considerably more live-testing studies are necessary. More information is necessary about examinee reactions to CATs and practical administration concerns, such as testing time, so that multiple administrations can be scheduled to effectively use the computers. In addition, live testing is necessary to determine the comparability of scores from P&P and CAT versions of a measurement instrument.

The polytomous models discussed in this paper all require that the test or instrument measure a unidimensional trait. Measurement instruments, however, often measure more than one dimension. For example, word problems in areas such as mathematics have been found to measure not only mathematical skills but also verbal comprehension (Reckase, 1985). Unidimensional polytomous models may have particular difficulties in multidimensional situations. For instance, in a multiple-choice item, different abilities may be used by an examinee in deciding between different response options, so that the option parameters are not on the same scale. Research is necessary on ways to handle these types of measurement problems through the use of multidimensional polytomous IRT models.

The possibility of integrating polytomous CATs with computer-assisted instruction programs needs to be studied. Merging the two computer-based components could result not only in adaptive assessment, but also adaptive instruction. Each student could potentially benefit from such a system because he/she would be receiving material that is appropriate for his/her trait level. Bright students would not be bored by too low a level of instruction and testing, whereas low ability students would not be frustrated by receiving too high a level of material. The system could be used to bring students to a specified level of proficiency and to facilitate diagnostic testing of examinee errors.

Presently, many CATs simply administer P&P items on a computer. As such, CAT is not fully exploiting the capabilities of the computer. New item types should be explored that will take advantage of the computer's resources and capabilities. For example, in a spatial ability test, the computer could allow the examinee to rotate the items graphically, which cannot be done on P&P instruments, and the degree of accuracy of the solution for each item could be used to score the item polytomously. Allowing the examinee to interact with the computer on problem-solving tasks also could enrich the polytomous scoring of items. Careful construction of items could prove useful for diagnostic testing. Frederiksen, Mislevy, & Bejar (1993) and Bennett & Ward (1993) provide insights into ways to reconceptualize test items, which should facilitate the development of creative item types that could be used in polytomous model-based CAT. The development of innovative items might also make viable the implementation of CAT-based performance assessment.

Although the research on polytomous CAT that has been conducted provides a good foundation, considerably more work is needed in order to develop CAT systems based on polytomous models. As computer technology continues to advance, the ways in which examinees interact with the computer will evolve. Capitalizing on these developments should facilitate the development of new item types and creative scoring algorithms based on polytomous models, and thus broaden the areas of measurement that can take advantage of the benefits of CAT.

References

American College Testing. (1993). *COMPASS user's guide.* Iowa City IA: Author.

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

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

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87–96.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17,* 239–251.

Balasubramanian, S. K., & Kamakura, W. A. (1989). Measuring consumer attitudes toward the marketplace with tailored interviews. *Journal of Marketing Research, 26,* 311–326.

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale NJ: Erlbaum.

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* (pp. 397–479). Reading MA: Addison-Wesley.

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

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

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Rep. No. 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Chen, S., Hou, L., Fitzpatrick, S. J., & Dodd, B. G. (1995, April). *The effect of population distribution and methods of theta estimation on CAT using the rating scale model.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

College Board. (1993). *Accuplacer user's notebook.* New York: Author.

Curran, L. T., & Wise, L. L. (1994, August). *Evaluation and implementation of CAT-ASVAB.* Paper presented at the annual meeting of the American Psychological Association, Los Angeles.

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

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16,* 327–343.

De Ayala, R. J. (1993). An introduction to polytomous item response theory models. *Measurement and Evaluation in Counseling and Development, 25,* 172–189.

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

Dodd, B. G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14,* 355–366.

Dodd, B. G., & De Ayala, R. J. (1994). Item information as a function of threshold values in the rating scale model. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2; pp. 301–317). Norwood NJ: Ablex.

Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement, 11,* 371–384.

Dodd, B. G., & Koch, W. R. (1994). Item and scale information functions for the successive intervals Rasch model. *Educational and Psychological Measurement, 54,* 873–885.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1988, April). *Computerized adaptive attitude measurement: A comparison of the graded response and rating scale models.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

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,* 129–143.

Dodd, B. G., Koch, W. R., & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement, 53,* 61–77.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). *Fitting polychotomous item response theory models to multiple-choice tests.* Unpublished manuscript.

Educational Testing Service. (1993). *GRE 1993-94 guide to the use of the Graduate Record Examinations Program.* Princeton NJ: Author.

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (Eds.). (1993). *Test theory for a new generation of tests.* Hillsdale NJ: Erlbaum.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347–360.

Henly, S. J., Klebe, K. J., McBride, J. R., & Cudeck, R. (1989). Adaptive and conventional versions of the DAT: The first complete test battery comparison. *Applied Psychological Measurement, 13,* 363–371.

Kamakura, W. A., & Balasubramanian, S. K. (1989). Tailored interviewing: An application of item response theory for personality measurement. *Journal of Personality Assessment, 53,* 502–519.

Kingsbury, G. G., & Houser, R. L. (1993). Assessing the utility of item response models: Computerized adaptive testing. *Educational Measurement: Issues and Practice, 12,* 21–27, 39.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York: Academic Press.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359–375.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4,* 241–261.

Koch, W. R. (1983). Likert scaling using the graded response 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. (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. (in press). An investigation of procedures for computerized adaptive testing using the successive intervals Rasch model. *Educational and Psychological Measurement.*

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.

Levine, M. V., Drasgow, F., Williams, B., McCusker, C., & Thomasson, G. L. (1992). Measuring the difference between two models. *Applied Psychological Measurement, 16,* 261–278.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement, 16,* 33–40.

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

Masters, G. N., & Evans, J. (1986). Banking non-dichotomously scored items. *Applied Psychological Measurement, 10,* 355–367.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529–544.

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: Latent trait test theory and computerized adaptive testing* (pp. 224–236). New York: Academic Press.

McKinley, R. L., & Reckase, M. D. (1980). Computer applications to ability testing. *Association for Educational Data Systems Journal, 13,* 193–203.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59–71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

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

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8,* 11–15.

Reshetar, R. A., Norcini, J. J., & Shea, J. A. (1993, April). *A simulated comparison of two content balancing and maximum information item selection procedures for an adaptive certification examination.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional scaling concept. *Applied Psychological Measurement, 12,* 397–409.

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

Singh, J. (1993, September). *Some initial experiments with adaptive survey designs for structured questionnaires.* Paper presented at the New Methods and Applications in Consumer Research Conference, Cambridge MA.

Singh, J., Howell, R. D., & Rhoads, G. K. (1990). Adaptive designs for Likert-type data: An approach for implementing marketing research. *Journal of Marketing Research, 27,* 304–321.

Sympson, J. B. (1983, June). *A new IRT model for calibrating multiple choice items.* Paper presented at the annual meeting of the Psychometric Society, Los Angeles CA.

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

Thissen, D., & Steinberg, L. (1986). Taxonomy of item response models. *Psychometrika, 51,* 567–577.

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

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale NJ: Erlbaum.

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

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 mea-*

surement of intellectual capabilities. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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

Zara, A. R. (1988). Introduction to item response theory and computerized adaptive testing as applied in licensure and certification testing. *National Clearinghouse of Examination Information Newsletter, 6,* 11–17.

Acknowledgments

The authors thank Steven J. Fitzpatrick for his constructive comments on earlier versions of this manuscript.

Author's Address

Send requests for reprints or further information to Barbara G. Dodd, University of Texas at Austin, Measurement and Evaluation Center, P.O. Box 7246, Austin TX 78713-7246, U.S.A. Internet: bgdodd@mail.utexas.edu.