# 3
# A Response Model for Multiple-Choice Items

David Thissen and Lynne Steinberg

## Introduction

In the mid-1960s, Samejima initiated the development of item response models that involve separate response functions for all of the alternatives in the multiple choice and Likert-type formats. Her work in this area began at the Educational Testing Service and continued during a visit to the L.L. Thurstone Psychometric Laboratory at the University of North Carolina. Both Samejima's (1969; this volume) original model for graded item responses and Bock's (1972; this volume) model for nominal responses were originally intended to produce response functions for all of the alternatives of multiple-choice items. For various reasons, neither model has proved entirely satisfactory for that purpose, although both have been applied in other contexts. Using a combination of ideas suggested by Bock (1972) and Samejima (1968, 1979), a multiple-choice model was developed that produces response functions that fit unidimensional multiple-choice tests better (Thissen and Steinberg, 1984); that model is the subject of this chapter.

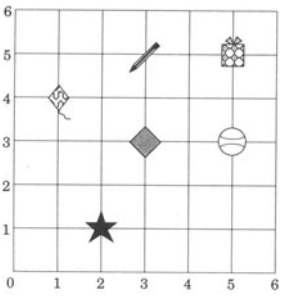
## Presentation of the Model

To give content to an illustration of the multiple-choice model, consider the mathematics item, shown in Fig. 1, that was administered in 1992 to about a thousand third-grade students as part of an item tryout for the North Carolina End of Grade Testing Program. Calibration of this item, embedded in an 80-item tryout form, gives the response functions shown in Fig. 2. As expected, the response function for the keyed alternative $B$, increases monotonically, and has approximately the form of the three-parameter logistic curves commonly used for the right–wrong analysis of multiple choice items. The response functions for alternatives $A$ and $D$, which are fairly obviously incorrect, are monotonically decreasing.

The interesting feature of this analysis involved alternative $C$, which is

W. J. van der Linden et al. (eds.), *Handbook of Modern Item Response Theory*

**The pencil is found at which ordered pair?**




A (3,3)

B (3,5)

C (5,3)

D (5,5)

FIGURE 1. Item 62 from an 80-item tryout form administered to third and fourth grade students as part of the development of the North Carolina End of Grade Testing Program.

a very popular distractor; it was selected by about 27% of the test takers, or about 60% of those who answered incorrectly. Further, the response functions in Figure 2 show that alternative $C$ tended to be selected by examinees of relatively higher proficiency. There is also some interesting anecdotal evidence about this item: When it was used as an illustration of the material to be on the new testing program, a number of classroom teachers complained that the item had two right answers. And when this item was presented informally to some (un-named) psychometricians, they tended to choose alternative $C$. The problem appears to be that the item-writers assumed that the only acceptable meaning of the phrase "ordered pair" involved the Cartesian coordinate system that is included in the third grade mathematics curriculum; that assumption makes $B$ the keyed alternative. However, in other contexts (e.g., using the row-column rule of matrix algebra and table construction, as in spreadsheets), alternative $C$ is equally the "ordered pair." While it is impossible to tell if any of the third-grade test takers followed such sophisticated rules in responding to this item, it would probably be regarded as more fair, and it would probably be more discriminating, if the context asked for, say, "graphical coordinates," instead of an "ordered pair."

The primary purpose of the multiple choice model is facilitation of graphically-based item analysis; curves like those in Fig. 2 may be very useful for item analysis and test construction. Wainer (1989) has described the use of such graphics in a dynamic item analysis system. Thorough analysis

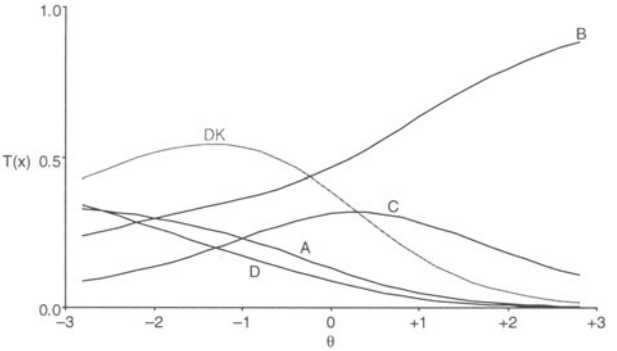



FIGURE 2. Multiple-choice-model response functions for item 62 (shown in Fig. 1), for third grade students.

of the behavior of the alternatives of a multiple-choice item is beyond the convenient reach of the traditional test theory because the relation between proficiency and the probability of most incorrect alternatives is not monotonic. The traditional theory rests on correlations, which summarize only monotonic relationships well. Item response theory need not be concerned with monotonicity, and furthermore, IRT facilitates graphical display of the concepts of interest. Wainer (1989) argues that graphical description of the data for each of the item's alternatives is superior to the numerical summaries of traditional item analysis. The multiple-choice model uses a parametric function as the mechanism for producing graphics such as that shown in Fig. 2; see Ramsay (1991; this volume; Abrahamowicz and Ramsay, 1992) for alternative, less parametric approaches.

Response functions for each category of the item for the multiple choice model are

$$T(u = h) = \frac{\exp[a_h\theta + c_h] + d_h \exp[a_0\theta + c_0]}{\sum_{k=0}^{m} \exp[a_k\theta + c_k]} . \tag{1}$$

The latent variable (*ability* in this volume; also often called *proficiency*) is denoted by $\theta$, and the response functions describe the probability of a response in category $h$ as a function of $\theta$. The observed data are in categories $h = 1, 2, \ldots, m$, where $m$ is the number of multiple-choice alternatives.

Elsewhere in the literature, the response functions are often referred to as *trace lines*; Lazarsfeld (1950, pp. 363ff) introduced that term to describe a curve representing the probability that a respondent selects a particular alternative as a response to an interview question or test item, as a function of the underlying variable that is being measured. That is the reason for the use of the notation $T$ to denote the response functions (trace lines), both as a mnemonic and to emphasize the fact that, while the response function has probabilistic interpretations, it does not describe an observable relative-frequency probability. $T(u = h)$ is a hypothetical type of probability: It can be described as the proportion of a (hypothetical) group of test takers with $\theta = \theta^*$ who select alternative $h$, or the proportion of (hypothetical) items

just like this one (in some sense) for which a particular test taker would select alternative $h$. It is used to describe these hypothetical probabilities, and the notation $P$ is reserved for later use to describe the model for observable, relative-frequency proportions.

There is no *a priori* ordering of the response categories. The response function model is identical to that for Bock's (1972; this volume) model for nominal responses, with the addition of the second term in the numerator in Eq. (1). Thus, the underlying stochastic model for the response process is identical to that described by Bock (1972; this volume), with the addition of responses in a latent category, referred to as *category 0* or *don't know* (DK). The idea is that the examinees who do not know the correct answer to a multiple-choice item comprise a latent class, and membership in that latent class has a response function that is:

Examinees in the latent DK class give an observed response, by selecting one of the multiple-choice alternatives; the model in Eq. (1) describes this process by adding

$$T(u=0) = \frac{\exp[a_0\theta + c_0]}{\sum_{k=0}^{m} \exp[a_k\theta + c_k]} \tag{2}$$

some proportion ($d_h$) of the DK responses to the response functions for each of the observed categories. The DK curve is shown in Fig. 2.

The parameters denoted by $a$ reflect the order, as well as discrimination, for the categories (see Thissen et al., 1989). In the nominal model, the category with the largest algebraic value of $a$ has a monotonically increasing response function; generally, that is the keyed alternative. After the addition of the DK curve in Eq. (1), the response function for the keyed response may be nonmonotonic, although it usually approaches 1.0 as $\theta$ becomes large. The category with the lowest algebraic value of $a$ has a monotonically decreasing response function; generally, that is DK. (The reader will note that the DK curve in Fig. 2 is not monotonically decreasing; that aspect of the model for this item will be considered below.) Response alternatives associated with intermediate values of $a$ have nonmonotonic response functions. The parameters denoted by $c$ reflect the relative frequency of the selection of each alternative; for alternatives with similar values of $a$, those with larger values of $c$ are selected more frequently than those with lower values of $c$.

The parameters $a_h$ and $c_h$ are not identified with respect to location, so these parameters are estimated subject to some linear constraint, most often that $\sum a_h = \sum c_h = 0$; the constraint is most easily imposed using reparametrization (Bock, 1972, 1975, pp. 239–249):

$$\mathbf{a}' = \alpha'\mathbf{T}_\mathbf{a},$$

where $\alpha$ contains the unconstrained parameters that are actually estimated. The $c_h$s are estimated using a similar reparametrization:

$$\mathbf{c}' = \gamma'\mathbf{T}_\mathbf{c}.$$

In some contexts, it may be desirable to use other types of contrasts as the rows of the matrix $\mathbf{T}$ [see Thissen (1991)]. The parameters represented by $d_h$ are proportions, representing the proportion of those who don't know who respond in each category on a multiple-choice item. Therefore, the constraint that $\sum d_h = 1$ is required. This constraint may be enforced by estimating $d_h$ such that

$$d_h = \frac{\exp[d_h^*]}{\sum \exp[d_h^*]}$$

and contrasts among the $d_h^*$ are the parameters estimated. Then reparametrization gives

$$\mathbf{d}^{*\prime} = \delta' \mathbf{T}_{\mathbf{d}}.$$

The elements of the vector $\delta$ are the parameters estimated.

The use of the parametric model brings with it all of the benefits now associated with IRT: parametric (very smooth) curves for item analysis, information functions for use in test assembly, straightforward computation of scaled scores (but see cautions, below), and likelihood-based procedures for linking scales for different groups, and for testing hypotheses, e.g., about differential item functioning [DIF; see Thissen et al. (1993)]. In addition, the multiple choice model has been used as the generating model in simulation studies of the performance of other IRT systems (Abrahamowicz and Ramsay, 1992; Wainer and Thissen, 1987). However, the use of such a complex parametric model also brings the burden of parameter estimation.

## Parameter Estimation

While there are no doubt many conceivable approaches to parameter estimation for the multiple-choice model and closely related IRT models, the only approaches that have been described in the literature are maximum likelihood estimation (that gives the MLEs) and the closely related *maximum a posteriori* (Bayesian) approach, for estimates referred to as MAPs.

Using maximum likelihood estimation, the item parameter estimates maximize the likelihood

$$L = \prod_{\{u\}} P(u)^{r_u}, \tag{3}$$

in which the product is taken over the set $\{u\}$ of all response patterns, $r_{\mathbf{u}}$ is the frequency of response pattern $\mathbf{u}$, and $P(\mathbf{u})$ represents the probability of that response pattern in the data,

$$P(\mathbf{u}) = \int \prod_{i=1}^{n} T_i(u_i \mid \theta)\phi(\theta)d\theta. \tag{4}$$

The population distribution for $\theta$, $\phi(\theta)$, is usually assumed to be $N(0,1)$; given the complexity of the model, it is not clear to what degree aspects

of the population distribution might be jointly estimable with the item parameters.

In general, there are no closed-form solutions for the parameter estimates, and so some iterative solution is required. In the context of the normal-ogive and two-parameter logistic models, Bock and Lieberman (1970) described an algorithm for direct maximization of Eq. (3); however, that approach was limited to a small number of items, with many fewer parameters per item than the multiple-choice model.

The direct approach to maximization of Eq. (3) has never been attempted for the multiple-choice model. The reason is that the number of parameters involved is very large: For five-alternative multiple-choice items, each $T_i(u_i \mid \theta)$ in Eq. (4) is a function of 14 parameters (five $\alpha$s, five $\gamma$s, and four $\delta$s); thus, for a 50-item test, Eq. (3), taken as a function of the item parameters (as it is for likelihood estimation), is a function of 700 parameters. In principle, it might be possible to maximize Eq. (3) directly as a function of those 700 parameters—but no commonly-used approach to likelihood maximization appears promising for this purpose. Derivative-free methods, maximizing Eq. (3) by brute force, would almost certainly flounder in a 700-dimensional parameter space. The Newton–Raphson approach used by Bock and Lieberman (1970) for the simpler IRT models with few items would require the computation of the first and second partial derivatives of the log of Eq. (3) with respect to each of the 700 free parameters—where the item parameters are within the product in Eq. (3) of the integrals of the products in Eq. (4); within that, the parameters are within the response function model [Eq. (1)], as well as the linear transformation used to ensure identifiability. Given the dimensionality of the parameter space, a direct approach to maximization of the likelihood for the multiple-choice model appears unlikely to be manageable. The likelihood in Eq. (3) [using the cell probabilities in Eq. (4)] is stated here only because it is the likelihood that is (indirectly) maximized in practical applications.

Bock and Aitkin (1981) described an indirect approach to the maximization of such IRT likelihoods using a variation of the EM algorithm, and that approach can be generalized to the multiple-choice model (Thissen and Steinberg, 1984). This algorithm is implemented in the computer program Multilog (Thissen, 1991). In broad outline, the iterative procedure involves a repeated sequence of so-called E-steps (for *expectation*) followed by M-steps (for *maximization*) until convergence. The steps are:

> **E-Step:** For each item, using the current estimates of the item parameters and some prespecified list of values of $\theta$, $\theta_q$, compute the expected frequencies of responses for each alternative as
>
> $$r^*_{hq} = \sum_{\{u_i=h\}} \prod_{i=1}^{n} T_i(u_i \mid \theta^*_q)\phi(\theta^*_q),$$
>
> where the sum is taken over all examinees with $u_i = h$.

**M-Step:** Using some suitable general function maximization technique, for each item maximize

$$\ell^* = \sum_h \sum_q r^*_{hq} \log T_i(u_i \mid \theta^*_q) \tag{5}$$

as a function of $\alpha$, $\gamma$, and $\delta$.

The E-step is completely defined above. The M-step defines a set of loglikelihoods that must be maximized; if no constraints are imposed on the model that involve the parameters of more than one item, there is one loglikelihood per item, involving only the parameters for that item. To reconsider the case of 50 multiple-choice items with five alternatives each, the M-Step involves 50 14-parameter estimation problems, maximizing the loglikelihood in Eq. (5). This can be done using derivative-free methods, or using a Newton–Raphson approach with the derivatives of Eq. (5) given by Thissen and Steinberg (1984, pp. 506–507). If constraints are imposed on the model that involve the parameters of two or more items, then the M-step must be done for the entire set of items over which the constraints have been imposed.

There is no proof of the uniqueness of the parameter estimates for the multiple-choice model (or, to our knowledge, of the parameters of any IRT model outside the so-called Rasch family). This is probably because it is easy to show that the parameter estimates are *not* unique. The obvious demonstration that the parameter estimates that maximize the likelihood are not unique arises from the observation that the goodness of fit remains the same if the signs of all of the slope parameters ($a$s) are reversed; this is equivalent to reversing the direction of the latent variable ($\theta$). This lack of uniqueness is common to all IRT models that are parameterized with slopes, and it is the same as the indeterminacy of reflection in the single common factor model. While this obvious indeterminacy precludes the development of any formal proof of the uniqueness of the parameter estimates, it does not appear to affect the usefulness of the model in applied contexts, where the orientation of $\theta$ is usually determined by the selection of starting values for the iterative estimation procedure.

Similarly, there is no general proof of the existence of the maximum likelihood estimates of the parameters, nor is one possible, because there are obvious patterns of data for which finite estimates of the parameters (taken as a set) do not exist. For example, if $x$ independent, then it follows that the MLEs of all of the $a$s are zero, and the $c$s for the observed categories parametrize the marginal totals; there is no information left in the data for estimation of the latent DK class of the $d$s. As is generally the case with complex latent-variable models, the behavior of the parameters is highly data-dependent.

Fitting the multiple choice model, using procedures described by Thissen and Steinberg (1984) and Thissen et al. (1989), is both computationally in-

tensive and fairly demanding of the data analyst. The focus of the problem is the large number of parameters involved: There are 11 free parameters per item for the four-alternative multiple-choice format, and 14 parameters per item for five-choice items. For the 80-item test that included the item shown in Fig. 1, fitting the model involved optimization of the likelihood in an 880-dimensional (parameter) space! The dimensions of the problem alone account for a good deal of the computer time required; using Multilog (Thissen, 1991), the 80-item calibrations for the examples used here, with about 1000 examinees, each required 1–2 hours on a 33MHz 80486-based MS-DOS computer.

The computational problem is exacerbated by the fact that the model is somewhat overparameterized for most test items and their associated data. Many multiple choice items, especially in the unselected sets usually used for item tryout, include distractors that are chosen by very few examinees. It is not unusual, even with a calibration sample size of 1000, for an alternative to be selected by fewer than 100 test takers. Nevertheless, the multiple choice model involves fitting a nonlinear function giving the expected proportion choosing that alternative; that would be unstable with a sample size of 100, even if the independent variable ($\theta$) was fixed and known, but in this case it is a latent variable.

Further, for a four-alternative item, the model devotes three free parameters to distinguishing among the guessing levels for the four choices, and two more parameters to distinguish the DK curve from the rest of the response functions. However, if the item is relatively easy, most test takers know the answer and respond correctly; only a subset of the sample don't know, and guess. It is not possible to say precisely how large the guessing subset is; but it is not simply the proportion of the sample who respond incorrectly, because some of those select the correct response. Nevertheless, the effective sample size for the guessing parameters (the $d_h$s) and the DK parameters (the contrasts that yield $a_0$ and $c_0$) can be very small.

The result is that there may be very little information in the data about some of the parameters for many of the items, and the speed of the convergence of the EM algorithm is proportional to the information about the parameters (Dempster et al., 1977). The solution to the problem is obviously to reduce the number of parameters; but it is not clear, *a priori*, which parameters are needed to fit which items. A very difficult four-alternative item may well have sufficient information for all 11 parameters, while many of the easier items in a test may require different subsets of the parameters to be fitted properly. A partial solution to the problem is the use of subjective prior densities to constrain the estimates, computing MAPs in place of MLEs. In the item calibrations for the examples in this chapter, Gaussian prior densities were used for all three classes of parameters: the $\alpha$s and $\gamma$s were assumed $N(0,3)$, and the $\delta$s were assumed $N(0,1)$. These densities multiply the likelihood in Eq. (3), and their logs are added to the loglikelihood in Eq. (5); otherwise, the estimation algorithm is unchanged.

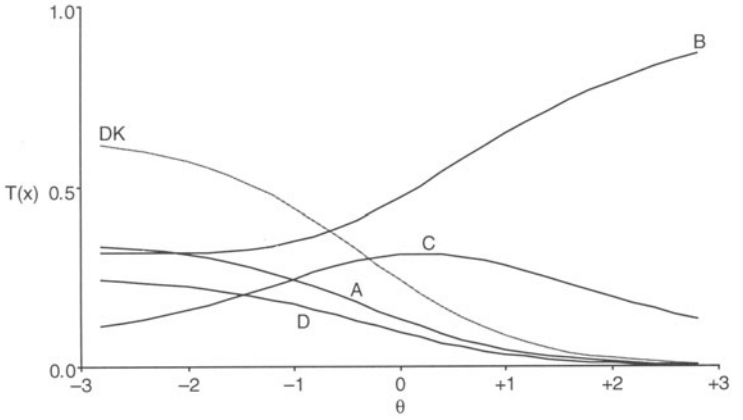



FIGURE 3. Multiple choice model response functions for item 62 (shown in Fig. 1), for third grade students, with the constraints on the parameters described in the text.

Often, upon inspection, the number of parameters for an item can be reduced without reducing the goodness of fit of the model. For instance, for the item in Fig. 1 with the curves in Fig. 2, after we observed that the response functions for alternatives $A$ and $D$ are very similar and roughly proportional to the DK curve, we re-fitted the model with the constraints $a_0 = a_1 = a_4$ and $d_1 = d_4$ (by fixing the values of $\alpha_1$, $\alpha_4$, and $\delta_3$ equal to zero). These constraints reduced the number of free parameters from eleven to eight, and did not change the overall goodness of fit of the model for the 80-item test as measured by $-2$loglikelihood. The curves for the constrained estimation are shown in Fig. 3. There is little change in the forms of the curves between those shown in Figs. 2 and 3, except that the DK curve is monotonically decreasing in Fig. 3, as it should be (showing that the nonmonotonicity of the DK curve observed in Fig. 2 is not a reliable feature of the data).

At the present time, imposing such constraints to reduce the number of parameters estimated can only be done by the data analyst, on an item-by-item basis. It should be possible, in principle, to create an automated system that examines the item response data and imposes appropriate constraints on the multiple-choice model parameters before likelihood-based estimation is attempted. Because such an automated system requires response functions, it might well be based on Ramsay's (Abrahamowicz and Ramsay, 1992; Ramsay, 1991, 1992, this volume) much faster nonparametric system to provide those curves. With appropriate constraints, removing parameters for which there is little information, computational time for estimation may be reduced by factors of between two and ten.

## *Estimating θ*

The response functions of the multiple-choice model may be used to compute scaled scores, in the same way that the response functions for any IRT model are used for that purpose. Given the response functions, which is to say, given the item parameters, and the assumption of local independence, the posterior density over $\theta$ for persons with response pattern $\mathbf{u}$ is

$$f(\mathbf{u} \mid \theta) = \left[\prod_{i=1}^{n} T_i(u_i \mid \theta)\right] \phi(\theta).$$

The mode or the mean of $f(\mathbf{u} \mid \theta)$ may be computed as the *maximum a posteriori* (MAP) or *expected a posteriori* (EAP) estimate of $\theta$, respectively. The MAP is the most likely value of $\theta$ for persons with response pattern $\mathbf{u}$, and the EAP is the average value of $\theta$ for persons with response pattern $\mathbf{u}$; thus, either is a reasonable summary description, or scaled score. The mode of $f(\mathbf{u} \mid \theta)$ may be computed using any of a number of approaches to function maximization, maximizing $f(\mathbf{u} \mid \theta)$ as a function of $\theta$. The mean of the posterior represented by $f(\mathbf{u} \mid \theta)$ requires numerical integration, as described by Bock and Mislevy (1982).

Bock (1972), Thissen (1976), Sympson (1983), Levine and Drasgow (1983), Thissen and Steinberg (1984), and Abrahamowicz and Ramsay (1992) have shown that some increased precision of measurement may be obtained when information in the incorrect alternatives on multiple-choice tests is included in the item response model. For the most part, the additional information obtained with models that take incorrect responses into account is limited to relatively low-ability examinees; with an appropriately-targeted test, most of the higher-ability examinees choose few incorrect alternatives, so there is little information available from that source for them.

There are, however, some potential difficulties with the use of the multiple choice model, or any model that incorporates a model for guessing, to compute scaled scores. As we noted at the time the model was originally presented (Steinberg and Thissen, 1984), the multiple-choice model may produce nonmonotonic response functions for correct responses. The use of nonmonotonic response functions in constructing the posterior density gives rise to potential problems: If the correct response function for an item is nonmonotonic on the right (e.g., it turns down), then examinees with some response patterns on the other items will be penalized by responding correctly to the item; they would have been assigned a higher scaled score if they had selected particular incorrect alternatives. Correct response functions that are nonmonotonic on the left similarly penalize examinees of low ability who respond correctly; conditional on the other item responses, a correct response to an item may not imply higher ability; rather, it may be more likely to be guessing or cheating. As measurement, this is all entirely satisfactory; but it may be a problem if the scoring system must be

explained to the examinees, and the examinees are not adequately trained in psychometrics.

The software Multilog (Thissen, 1991) provides MML estimates of the item parameters for the multiple-choice model (as well as for those of several other logistic IRT models), or MAP estimates using Gaussian prior distributions. Multilog also computes MAP scaled scores for individual response patterns, or EAP scaled scores associated with response patterns in tabulated frequency data. The currently available version of Multilog operates under the MS-DOS operating system, and is limited to 50 items (with conventional memory), or 200 items (with extended memory), with up to nine response alternatives per item.

## Goodness of Fit

Little formal work has been done on goodness-of-fit testing for any multiple-category models outside of the so-called Rasch family; so far, the problem appears fairly intractable. For fewer than three items, the multiple-choice model is simply not identified, so goodness of fit is an irrelevant concept. For a sufficiently large examinee sample and three, four, or five items, the $m^n$ cross-classification of the responses produces a contingency table that may have expected values sufficiently large that the conventional Pearson's or likelihood ratio test statistics may be used to test the goodness of fit against the general multinomial alternative; that procedure was illustrated previously (Thissen and Steinberg, 1984; Thissen et al., 1989).

For six or more items, with any conceivable sample size, the $m^n$ table becomes too sparse for the conventional test statistics to be expected to follow their conventionally expected $\chi^2$ distributions, so that this approach does not generalize to long tests. The problem is that no specification of the distribution of responses in such a sparse table under some reasonably general alternative hypothesis has been done in any context. This leaves us with no general goodness-of-fit test for the model.

It is possible to use likelihood ratios to test the significance of the differences between hierarchically constructed parametrizations; indeed, that was done earlier in this chapter, when the equality constraints were imposed on the parametrization for the item shown in Fig. 1, to obtain the constrainted fit shown in Fig. 3. The conventional likelihood ratio test for the difference between the unconstrained model shown in Fig. 2 and the constrained model shown in Fig. 3 was 0.0 (computed to one decimal place) on three degrees of freedom (because the values of three contrast-coefficients, $\alpha_1$, $\alpha_4$, and $\delta_3$, were constrained to be zero in the restricted model). Such likelihood ratio tests can be used generally to compare the fit of more versus less constrained models. Their use for examining differential item functioning (DIF) using the multiple choice model was illustrated in a paper by Thissen et al. (1993).

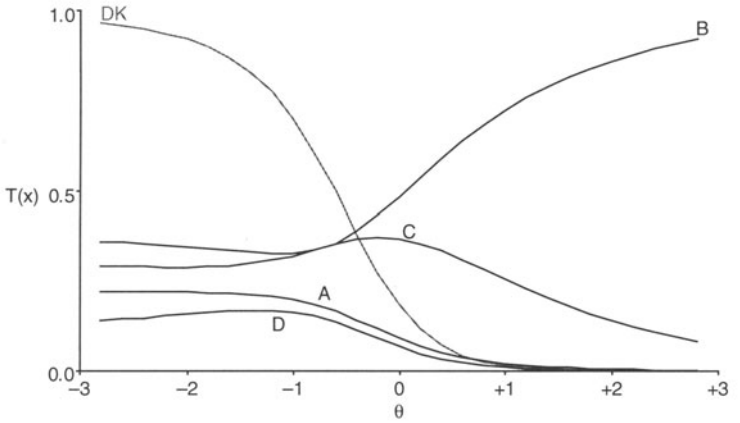



FIGURE 4. Multiple choice model response functions for item 62 (shown in Fig. 1), for fourth grade students.

Thus, as of this writing, the status of goodness-of-fit testing for the multiple-choice model is identical to that for the general linear model: There has been developed no overall goodness of fit against any general alternative; however, assuming that the fully parametrized model fits adequately, there are likelihood ratio tests of the significance of the null hypothesis that any of the component parameters in the model equal zero (or any other appropriate null value). Glas (1988) has developed some refined diagnostic statistics for Rasch family models; but no generalization of those statistics to models with differential slope and/or guessing parameters has yet appeared.

## Example

As part of the scale-linking design used in the development of the North Carolina End of Grade Tests, the item tryout form including Item 62 (shown in Fig. 1) was also administered to a sample of about 1000 fourth grade students. The response function obtained in the calibration of that form with the fourth grade sample are shown in Fig. 4. Comparing Fig. 4 with Figs. 2 and 3, it is clear that the main features of the response functions are essentially the same, indicating that the results obtained are not specific to either sample. As expected, the response functions in Fig. 4 are shifted about 0.5 standard unit to the right, relative to Figs. 2 and 3; this is because when they are placed on a common scale, average mathematics ability in the fourth grade is about 0.5 standard unit higher than in the third grade.

## Discussion

The principal use of the multiple-choice model is item analysis: Curves such as those shown in Figs. 2–4 provide more information to item analysts than do numerical summaries of item performance. In addition, the model and its associated graphics may provide information that may help item writers modify items that do not perform well, by indicating which of the distractors might be revised. The model also invites the more detailed analysis of DIF, by permitting the analyst to examine differences between alternative selection for different groups of examinees; see Green et al. (1989) for further motivation of this idea.

Given currently available software, the multiple-choice model is fairly difficult to apply routinely in large testing programs. The task of eliminating redundancies in the parametrization is a taxing one; before the model is likely to be widely applied, further research would be useful, to develop a system to reduce the parametrization for each item to a smaller set that is well identified. Nevertheless, interpretable results may be obtained without exhaustive model fitting and constraints. The multiple-choice model provides a complete IRT item analysis for multiple choice tests, with response functions for all of the response alternatives. Graphical presentation of the response functions provides information for thorough item analysis, which has much to recommend it. The item analysis is sufficiently flexible that even poorly designed items can be fitted and specific flaws can be exposed. These can then be observed and such items modified or eliminated in the course of test development.

## References

Abrahamowicz, M. and Ramsay, J.O. (1992). Multicategorical spline model for item response theory. *Psychometrika* **57**, 5–27.

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

Bock, R.D. (1975). *Multivariate Statistical Methods in Behavioral Research.* New York: McGraw-Hill.

Bock, R.D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika* **46**, 443–449.

Bock, R.D. and Lieberman, M. (1970). Fitting a response model for $n$ dichotomously scored items. *Psychometrika* **35**, 179–197.

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

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, **39**, 1–38.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika* **53**, 525–546.

Green, B.F., Crone, C.R., and Folk, V.G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement* **26**, 147–160.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, and J.A. Clausen, *Measurement and Prediction* (pp. 362–412). New York: Wiley.

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

Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika* **56**, 611–630.

Ramsay, J.O. (1992). *TESTGRAF: A Program for the Graphical Analysis of Multiple Choice Test and Questionnaire Data* (Technical Report). Montreal, Quebec: McGill University.

Samejima, F. (1968). *Application of the Graded Response Model to the Nominal Response and Multiple Choice Situations* (Research Report #63). Chapel Hill, N.C.: University of North Carolina, L.L. Thurstone Psychometric Laboratory.

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

Samejima, F. (1979). *A New Family of Models for the Multiple Choice Item* (Research Report #79-4). Knoxville, TN: University of Tennessee, Department of Psychology.

Steinberg, L. and Thissen, D. (1984, June). *Some Consequences of Non-Monotonic Trace Lines in Item Response Theory.* Paper presented at the meeting of the Psychometric Society, Santa Barbara, CA.

Sympson, J.B. (1983, June). *A New IRT Model for Calibrating Multiple Choice Items.* Paper presented at the meeting of the Psychometric Society, Los Angeles, CA.

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

Thissen, D. (1991). *MULTILOG User's Guide–Version 6.* Chicago, IL: Scientific Software.

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

Thissen, D., Steinberg, L., and Fitzpatrick, A.R. (1989). Multiple choice models: The distractors are also part of the item. *Journal of Educational Measurement* **26**, 161–176.

Thissen, D., Steinberg, L., and Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland and H. Wainer (Eds.), *Differential Item Functioning* (pp. 67–113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement* **26**, 191–208.

Wainer, H. and Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics* **12**, 339–368.