PSYCHOMETRIKA—VOL. 49, NO. 4, 501–519
DECEMBER 1984

A RESPONSE MODEL FOR MULTIPLE CHOICE ITEMS

DAVID THISSEN AND LYNNE STEINBERG

UNIVERSITY OF KANSAS

We introduce an extended multivariate logistic response model for multiple choice items; this model includes several earlier proposals as special cases. The discussion includes a theoretical development of the model, a description of the relationship between the model and data, and a marginal maximum likelihood estimation scheme for the item parameters. Comparisons of the performance of different versions of the full model with more constrained forms corresponding to previous proposals are included, using likelihood ratio statistics and empirical data.

Key words: item response theory, multiple choice items, marginal maximum likelihood.

Introduction

In practical ability measurement, binary item response models have been applied routinely to data from multiple-choice tests with four or five alternatives per item. To use the binary models, the data have been dichotomized (correct and incorrect) and the distinct identity of the incorrect alternatives has been lost. This procedure logically follows the tradition of "scoring" a test by considering only the correct answers; however, the binary models are incomplete as theories of the item responses as useful information may be lost in the dichotomization.

It is possible to conceive of a unidimensional latent variable model which completely explains the item response data, and several have been proposed. The model which will be introduced here is an extension of that proposed by Bock (1972) and subsequently extended by Samejima (1979); a different parameterization of the same model has been discussed by Sympson (1983). Bock (1972), Thissen (1976), Sympson (1983), and Levine and Drasgow (1983) 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. In the next section we will introduce the model and describe a workable scheme for maximum likelihood estimation of its parameters. In a subsequent section, we will discuss the gain in information with a new model, as well as some attendant difficulties.

The Model

The model for multiple-choice items to be described makes heavy use of the multivariate logistic transformation suggested as an item response model by Bock (1972), so it is useful to begin with a clear understanding of that model. We do not derive the model as

This research was supported in part by ONR contract N00014-83-C-0283 to the University of Chicago, R, Darrell Bock, Principal Investigator, and AFHRL contract F41689-82-C-00020 to the Educational Testing Service, Howard Wainer, Principal Investigator. This project was facilitated by a sabbatical leave provided David Thissen by the University of Kansas, during which time some of this work was done with the Research Statistics Group of the Educational Testing Service. Conversations with Darrell Bock, Howard Wainer, Paul Holland, Frederic Lord, Donald Rubin, James Ramsay and Malcolm Ree have also been important and useful in shaping the research and its presentation here. While all those mentioned have helped us gain wisdom in these matters, any error that remains is, of course, our own.

Requests for reprints should be sent to David Thissen, Department of Psychology, University of Kansas, Lawrence, KS, 66045, (913)864-4131.

0033-3123/84/1200-6013$00.75/0

a process model, although that might be possible. We develop the model as a flexible system for producing trace lines closely approximating item response data, without extensive semantic interpretation of the parameters.

For m categorical responses, we specify m response functions $z_k = a_k\theta + c_k$, each of which is a linear function of the latent ability variable θ. Usually z_{correct} would have a positive slope and the other z_k's would have negative or less positive slopes. The linear functions describe one of the simplest possible relationships between the response and θ. To make this a model for categorical item responses, the z_k (which lie on the real line) must be mapped into [0, 1]. This is accomplished by the multivariate logistic transformation, so

$$P(x_j = h \mid \theta; \mathbf{a}, \mathbf{c}) = \frac{\exp(z_h)}{\sum_{k=1,\, m_j} \exp(z_k)}. \tag{1}$$

The function (1) is Bock's (1972) model for an item response, $x_j = h$, in which $h = 1, 2, \ldots m_j$ for a multiple choice item j with m_j (classes of) response alternatives. The item parameters are the vectors $\mathbf{a}$ and $\mathbf{c}$, subject to two suitable linear constraints (see below), giving $2(m_j - 1)$ free parameters. The model described by (1) is moderately flexible, but lacks flexibility in certain crucial respects at the extremes. Specifically, one of the response alternatives must have the largest positive value of a_k; the trace line for that alternative is then monotonic increasing. That may be theoretically acceptable, since it is probably the correct response. However, another alternative must have the largest negative value of a_k ($\sum a_k = 0$ is one of the linear constraints), and that alternative's trace line must be monotonic decreasing. The latter aspect of the model is less acceptable: it implies that as ability decreases the probability of selecting one particular incorrect response approaches unity, and all the others go to zero. That is unlikely.

Samejima (1979) proposed a solution to this problem in a conceptual modification of Bock's (1972) model in which she added a completely latent *response category* labelled "zero." Here, we will sometimes refer to this category as "don't know" (DK). Lord (1983) has introduced a similar conceptual entity which he describes by saying the examinee is "(totally) undecided." In Samejima's model, DK is not an *observed* response, but a latent one, multi-logit-linearly dependent on a latent trait (giving two layers latent); the idea was that some proportion $d_h (= 1/m_j$ in Samejima, 1979) "guessed" each of the observable response alternatives and were "mixed in" with the examinees who chose those alternatives intentionally. So the model becomes

$$P(x_j = h \mid \theta; \mathbf{a}, \mathbf{c}, \mathbf{d}) = \frac{\exp(z_h)}{\sum_{k=0,\, m_j} \exp(z_k)} + \frac{d_h \exp(z_0)}{\sum_{k=0,\, m_j} \exp(z_k)},$$

in which the second term adds a proportion d_h of those who "don't know" into each trace line; for future brevity we refer to the model as $P_j(h)$:

$$P_j(h) = \frac{\exp(z_h) + d_h \exp(z_0)}{\sum_{k=0,\, m_j} \exp(z_k)}, \tag{2}$$

in which h takes the values $1, 2, \ldots m_j$. Samejima's (1979) model had two more parameters than Bock's (1972) model: a_0 and c_0. In her presentation, the d_h were fixed and equalled $1/m_j$; this represents the hypothesis that those of sufficiently low ability assign their responses randomly with equal probability to each of the response alternatives. We found that unlikely; later we will show that it is not empirically the case. So we extend the version of the model given in (2) to allow the d_h, $h = 1, 2, \ldots m_j$ to be functions of estimated parameters.

Indeterminacies and Constraints

The model expressed in (2) has a number of indeterminacies and requires the imposition of some constraints to become identifiable. The linear constraints required on **a** and **c**,

$$\sum_{k=0,\, m_j} a_k = \sum_{k=0,\, m_j} c_k = 0$$

are imposed by reparameterization:

$$\mathbf{a} = \mathbf{T}\boldsymbol{\alpha}$$

and

$$\mathbf{c} = \mathbf{T}\boldsymbol{\gamma},$$

where $\mathbf{T}$ is an $(m_j + 1) \times m_j$ transformation matrix, and $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are vectors of free parameters of order m_j; $\mathbf{T}$ is the transpose of that in Bock (1972) throughout. There is a further indeterminacy with respect to the *sign* of $\boldsymbol{\alpha}$; this is an indeterminacy of reflection of the latent variable θ which is logically identical to the rotational indeterminacy of common factor analysis. The reflection indeterminacy is not explicity solved, but rather the iterative solution of the likelihood equations is started with a_h positive for $h =$ the correct response; this has been sufficient to keep the estimated solution "right-side-up" in all applications of the model to date.

Since the vector **d** represents a set of m_j proportions, the d_h must be constrained to sum to unity and lie on the interval [0, 1]. The dual constraint is imposed in two parts. In the first, the same multivariate logistic used in the model is also employed to make the d_h proportions from a set of psuedo-parameters d_h^*, which lie on the real line:

$$d_h = \frac{\exp(d_h^*)}{\sum_{k=1,\, m_j} \exp(d_k^*)}. \tag{3}$$

The parameters d_h^* are on the real line, but like **a** and **c**, they must satisfy $\sum d_k^* = 0$. So

$$\mathbf{d}^* = \mathbf{T}_2\,\boldsymbol{\delta} \tag{4}$$

in which $\mathbf{T}_2$ is an $m_j \times (m_j - 1)$ transformation matrix of the same form as $\mathbf{T}$ (above) but of lesser order, and $\boldsymbol{\delta}$ is a vector of length $(m_j - 1)$ of parameters which (through (4) and (3)) give the proportions d_h.

So the set of free parameters for an item consists of $m_j \alpha_k$'s ("a-contrasts"), $m_j \gamma_k$'s ("c-contrasts"), and $(m_j - 1)\delta_k$'s ("d-contrasts"), for a total of $3m_j - 1$ parameters: 11 for a four-alternative multiple-choice item.

The estimation system we use permits any of the elements of $\{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta}\}$ to be fixed at any constant value, or to be constrained to be equal to any other parameter. With this facility, our model (2) includes the previous models hierarchically as subsets. If $\boldsymbol{\delta}$ is fixed at **0**, than all the $d_h = 1/m_j$ and (2) becomes Samejima's (1979) model. If the parameters a_0 and c_0 are fixed at zero, the d_h become irrelevant (and are deleted mechanically) and the model becomes Bock's (1972) model. In our examples, we will compare the goodness of fit of all of these models, as well as versions of (2) that include equality constraints on particular parameters imposed across items.

The Relationship of the Model to Data

Models in item response theory (IRT) are intended to explain observed covariation among test-item responses. Unidimensional "latent trait" theories explain that covariation by appeal to an underlying latent variable (usually denoted θ) on which the probabilities

TABLE 1

Cross Classification of Responses to Two Four-Choice Items.

		Item 2				
		A	B	C*	D	
Item 1	A	20(22)	107(23)	79(22)	44(31)	250
	B*	32(36)	157(34)	143(40)	35(25)	367
	C	7(8)	36(8)	24(7)	19(14)	86
	D	30(33)	161(35)	112(31)	42(30)	345
		89	461	358	140	1048

*=Correct; column percentages are in parentheses.

of the responses are functionally dependent. Each pair of item responses is theorized to have non-zero covariance because the probability of both item responses depends on θ. The models usually hold that θ is the *sole* cause of covariation among the item responses; conditional on θ, the item responses are theorized to be independent ("local independence").

A complete IRT model for the item response data for a multiple-choice test in which n items each have m alternatives would be a model for the covariation in the m^n contingency table containing the counts of respondents giving each response pattern. Such a 4^2 table for two four-alternative multiple-choice items is given in Table 1. Evidence that there is some information in the examinees' incorrect choices is apparent (or at least can be found) in the covariance in the table. (The column percentages for the counts are given in parentheses to aid in interpretation.) Alternative C is correct for item 2. Respondents who select C on item 2 are most likely to choose B on item 1; B is correct, so that is double evidence of their high ability. But examinees who choose A or B on item 2 are more likely to respond correctly to item 1 (36% and 34%) than those who pick D on item 2 (only 25%). A possible explanation is that those who select A or B on item 2 are higher on the ability continuum than are those who pick D. Even with 16 cells, this sort of argument may be lengthy, but the idea is that wrong responses are related to each other; if the responses are locally independent, this implies a relationship to "the trait."

The goodness of fit of a multiple category IRT model to the data in such a table may be evaluated with conventional likelihood ratio χ^2 statistics. For sets of data with more examinees than cells in the table, the likelihood ratio test against a general multinomial alternative may be used. For larger tables (35 four-choice items gives a 4^{35} table), there is no "general multinomial alternative"; but the likelihood ratio test between hierarchically nested models still may be used to evaluate the significance of the additional parameters of the larger model.

Estimation

In general, the data for the estimation of the parameters $\{\alpha, \gamma, \delta\}_j$ for each item j, $j = 1, \ldots, n$ consist of the counts of response patterns $r_{\mathbf{x}}$ (where each x_j may take values in $[1, 2, \ldots m_j]$) in an $m_1 \times m_2 \times \cdots \times m_n$ contingency table like Table 1. The probability of observing a particular response pattern $\mathbf{x}$ when drawing an examinee from a population

in which θ follows a distribution $\phi(\theta)$, assumed $N(0, 1)$ in the examples, is

$$P(\mathbf{x}) = \int_{-\infty}^{\infty} \prod_{j=1,n} P_j(x_j)\phi(\theta)\, d\theta \tag{5}$$

in which $P_j(x_j)$ is from equation (2).

The likelihood for the entire set of observed data is

$$L = C \prod_{\mathbf{x}} P(\mathbf{x})^{r_{\mathbf{x}}} \tag{6}$$

in which C is a normalizing constant not dependent on the parameters, and the product runs over all possible response patterns. In practice, of course, for more than a handful of items the number of possible response patterns is astronomical, and the "count" in each cell is one or zero. In such cases, only those response patterns actually observed need to be considered in the computation of (6), or (7), a function proportional to the loglikelihood:

$$\ell_0 \sim \sum_{\mathbf{x}} r_{\mathbf{x}} \log P(\mathbf{x}). \tag{7}$$

For a very small number of items (3 or 4; the parameters of the full model are not identifiable with only 2 items, as there are fewer cells than parameters), it may be possible to obtain the MLEs by directly maximizing (7) with a Newton-type algorithm. Bock (1972) reported such results for his original model with procedures similar to those used by Bock and Lieberman (1970).

Bock and Aitkin (1981) described an extremely simple and elegant algorithm for maximizing functions like (7) for binary models, and we extend that algorithm to the estimation of multiple category models. The Bock-Aitkin algorithm is a two-step procedure like the "EM-algorithm" (Dempster, Laird & Rubin, 1977), so we describe first the "E-step" and then the "M-step." The procedure is iterative, and EM-pairs are repeated until the process converges.

The E-step

The Bock-Aitkin algorithm is based on a discrete representation of $\phi(\theta)$ and the integrand of $P(\mathbf{x})$, both continuous densities, over Q "quadrature points," θ_q, with $q = 1, 2, \ldots Q$. Such a discrete representation of the continuous densities may be made arbitrarily close to continuous reality by choosing Q large, just as numerical integration may be made arbitrarily accurate by using sufficient quadrature points. However, large values of Q slow the computations; we use $Q = 10$ over the range $\theta = -4.5$ to $\theta = 4.5$ in unit steps.

Under the assumption that the population is composed of individuals who are members of Q discrete "classes" with values $\theta_1, \theta_2, \ldots, \theta_Q$ on the latent variable, "complete data sufficient statistics" for the estimation of the item parameters $\{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta}\}_j$ for item j would consist of a table of counts $\mathbf{R}_j^*$, in which each element r_{jkq}^* is the number of individuals in class θ_q selecting response alternative k on item j. So the E-step of the Bock-Aitkin algorithm consists of computing the expected values of the r_{jkq}^*, conditioned on the data and the current provisional estimates of the item parameters.

Using provisional estimates of the item parameters for each item (as starting values we use $\boldsymbol{\delta} = \mathbf{0}$, $\boldsymbol{\gamma} = \mathbf{0}$, and $\alpha_k = 1$ for categories k which are incorrect and $\alpha_k = 2$ for category k correct), compute the elements of the $m_j \times Q$ table $\mathbf{R}_j^*$ containing:

$$E(r_{jkq}^* \mid \text{data}; \{\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}\}) = \sum_{(\mathbf{x} \text{ in } \xi)} r_{\mathbf{x}}[P(\mathbf{x}; \theta_q)/\sum_{q} P(\mathbf{x}; \theta_q)] \tag{8}$$

in which ξ is the set of $\mathbf{x}$ in which $x_j = k$ and

$$P(\mathbf{x}; \theta_q) = \prod_{j=1,n} P_j(x_j)\phi(\theta_q). \tag{9}$$

Note that, while (9) and (8) are computed in a (potentially long) loop over the observed response patterns, the values $P_j(k)$ for each item from (2) are required only for a fixed set of Q values of θ. If those values are placed in a table before the E-step is begun, the computations involved in (8) and (9) are limited to table look-up, multiplication, and addition. The E-step yields a set of n $m_j \times Q$ tables of non-integral artificial "counts" which are used as data in the M-step.

The M-step

The M-step consists of maximum likelihood estimation of the parameters $\{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta}\}_j$ for all items $j = 1, 2, \ldots n$, using the tables of expected values $\mathbf{R}_j^*$ as data. It is simply nonlinear regression.

In terms of the "data" in $\mathbf{R}_j^*$, the loglikelihood for item j is

$$\ell_j \sim \sum_k \sum_q r_{jkq}^* \log P_j(k; \theta_q) \tag{10}$$

in which $P_j(k; \theta_q)$ is equation (2) evaluated at θ_q. Standard gradient methods may be used to maximize ℓ_j over the parameter space. The MLEs of the parameters are obtained where

$$\frac{\partial \ell_j}{\partial \boldsymbol{\zeta}} = \mathbf{0}$$

for all parameters in $\boldsymbol{\zeta} = \{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\delta}\}$. Since for any set of parameters constrained by $\mathbf{T}$ such that $\boldsymbol{\kappa} = \mathbf{T}\boldsymbol{\zeta}$,

$$\frac{\partial \ell_j}{\partial \boldsymbol{\zeta}} = \mathbf{T} \frac{\partial \ell_j}{\partial \boldsymbol{\kappa}}, \tag{11}$$

we will state the required components of the gradient for the constrained parameters $\{\mathbf{a}, \mathbf{c}, \mathbf{d}\}$; the gradients used to maximize (10) are then given by (11). Further,

$$\frac{\partial \ell_j}{\partial \kappa} = \sum_k \sum_q r_{jkq}^* \frac{1}{P_j(k; \theta_q)} \frac{\partial P_j(k; \theta_q)}{\partial \kappa}. \tag{12}$$

So what we really need are the derivatives of (2) with respect to $\mathbf{a}$, $\mathbf{c}$, and $\mathbf{d}^*$ (substitute all for κ above). To write these, a useful bit of shorthand is

$$e_k = \exp(a_k \theta + c_k),$$

which is always assumed to be evaluated at the appropriate value of θ; all summations of e_k are over categories $k = 0, 1, \ldots m_j$. The elements of the gradient (P_h is short for $P_j(h; \theta_q)$ hereafter; h' refers to categories other than h) to be substituted in (12), then (11), and

finally zeroed to maximize (10) are:

$$\frac{\partial P_h}{\partial a_0} = \frac{[(\sum e_k)\theta d_h e_0 - (e_h + d_h e_0)\theta e_0]}{(\sum e_k)^2}$$

$$\frac{\partial P_h}{\partial a_h} = \frac{[(\sum e_k)\theta e_h - (e_h + d_h e_0)\theta e_h]}{(\sum e_k)^2}$$

$$\frac{\partial P_h}{\partial a_{h'}} = \frac{-\theta e_{h'}}{\sum e_k}$$

$$\frac{\partial P_h}{\partial c_0} = \frac{[(\sum e_k)d_h e_0 - (e_h + d_h e_0)e_0]}{(\sum e_k)^2}$$

$$\frac{\partial P_h}{\partial c_h} = \frac{[(\sum e_k)e_h - (e_h + d_h e_0)e_h]}{(\sum e_k)^2}$$

$$\frac{\partial P_h}{\partial c_{h'}} = \frac{-e_{h'}}{\sum e_k}$$

$$\frac{\partial P_h}{\partial d_h^*} = \frac{e_1}{\sum e_k} \cdot \frac{\sum d_k^* - d_h^{*2}}{(\sum d_k^*)^2}$$

The gradient does not simplify very much since the model is not part of an exponential family. However, it is possible to locate the maximum of ℓ_j using these derivatives, and use the resulting set of parameters in the next E-step. Our estimation procedure allows equality constraints to be placed across as well as within items; this requires some of the gradients to be accumulated across items, but causes no serious problem beyond book-keeping. We use a conditioned Newton-type algorithm (MINIM, described by Haberman, 1974) to locate the maximum of (10); the conditioning is useful since the matrix of second derivatives may be nearly singular for some items.

Convergence and Local Minima

Using the updated item parameters from the M-step, the sequence (E-step, M-step) is repeated until either (a) the parameters stabilize or (b) a fixed number of cycles is reached (we usually use 15 or 20). With some sets of data, the change between cycles for all parameters becomes small; in other cases, most of the parameters and the loglikelihood remain roughly constant after ten or fifteen cycles, while one or a few parameters change linearly and indefinitely. The changing parameter is frequently a particular α_k (and its associated γ_k), changing as the associated a_k rises toward a very high (possibly infinite) MLE. It might be noted that this situation arises with the 2-PL and 3-PL models as well; there are some configurations of data, analogous to those giving "Heywood cases" in factor analysis, with which the "true" MLE is infinite and the essentially linear iterative system proceeds in that direction indefinitely. The difference in goodness of fit with such a slope high (3. to 5. or so) and much higher is negligible, so there is no loss in simply stopping the estimation procedure arbitrarily. In such a case, the trace line is fairly well-determined, but the numerical value of the associated a_k is not.

A second case in which the parameters fail to clearly converge arises with γ_k's associated with a_k's near zero. The parameter c_k becomes ill-defined when a_k is zero, and the associated γ_k "wanders." Again, there is no loss of fit if the estimation procedure is stopped. The Bock-Aitkin algorithm may be speeded slightly using acceleration as in

Thissen (1982), but we limit the value of the acceleration parameter to -1, which doubles the step-size at each cycle, to avoid oscillation.

The likelihood surface for the model clearly has more than one "local" minimum. The indeterminacy with respect to reflection of θ gives two equal, identical modes. With poorly chosen starting values, the estimation procedure has located other (apparent) stationary points; usually they seem to be located in a peculiar region corresponding to some items using both reflections of θ in the same solution—these may be modes or saddle points. Good starting values and the EM-like nature of the Bock-Aitkin algorithm provides a solution in this case, because EM-algorithms only climb local modes. If the algorithm starts near the desired part of the likelihood surface, it will end there.

The likelihood surface is multi-model and may be ill-behaved, producing parameter estimates that are sometimes undefined or are on the boundary. The entire system begs for full Bayesian treatment, with a prior distribution restricting the parameters to a reasonable part of the space under all circumstances. At this stage however, it is not clear what sort of prior may be appropriate. After some experience with the model is gained beyond that in the examples which follow, we may be in a position to specify reasonable parameters for a prior.

Alternatively, a different model, in the exponential family and with a unimodal likelihood (preferably with similar properties to this one in terms of fitting the data), would also solve all of these problems. Such a model has yet to be proposed.

Characterizing θ

The parameter estimation described above usually has the goal of "calibrating" a set of test items, after which the parameters are to be taken as known and used to characterize θ for examinees who produce a particular response pattern $\mathbf{x}$. Given a set of item parameters, the posterior density for θ is

$$P(\mathbf{x};\,\theta) = \prod_{j=1,\,n} P_j(x_j)\phi(\theta), \tag{13}$$

in which $P_j(x_j)$ is from equation (2). If the model is correct, (13) describes the distribution of examinees who respond with pattern $\mathbf{x}$. It can be characterized graphically, and two examples will be presented below in Figure 5.

For more than a few items, (13) is roughly Gaussian in shape, and so it may also be described by estimates of its location and spread. An extension of the procedure commonly used in binary IRT is to use the mode as an estimate of the location of (13), where

$$\frac{\partial \log P(\mathbf{x};\,\theta)}{\partial\theta} = \sum_{j=1,\,n} \frac{\partial \log P(x_j;\,\theta)}{\partial\theta} + \frac{\partial \log \phi(\theta)}{\partial\theta}$$

is zero, with the variance approximated by the negative inverse of

$$E\,\frac{\partial^2 \log P(\mathbf{x};\,\theta)}{\partial\theta^2}$$

The modal estimate is practical to compute as long as $\phi(\theta)$ is a reasonable function, and easy if $\phi(\theta)$ is normal. There is no guarantee that (13) is unimodal; for small numbers of items it is likely to have more than one mode. Multimodality presents potential problems for mechanical use of modal estimates.

However, it is also fairly straightforward to numerically integrate (13) to obtain its mean and variance. The mean has been called the "EAP" (Expected A Posteriori) estimate of θ by Bock and Mislevy (1982). An advantage of the EAP procedure over modal esti-

mation is that the derivative of $\phi(\theta)$ is not required; therefore $\phi(\theta)$ may take any form describable as a histogram with finite variance.

Examples

To illustrate application of the model to item response data, we have analyzed several subsets of items from an operational military accession test, total $N = 1178$. All items are four-alternative multiple choice. For small illustrations of the performance of variously constrained forms of the model, including the Bock (1972) and Samejima (1979) proposals, we use two small tests consisting of four relatively difficult items from a verbal test (called items, *R*, *S*, *T*, and *U*) and four items from a test of technical knowledge (called items *W*, *X*, *Y*, and *Z*). In the sequel, we refer to the four-item tests as *RSTU* and *WXYZ*. Four items were used because, with roughly a thousand examinees, that is the largest number for which the number of cells in the m^n table is less than N ($4^4 = 256$), so significance tests of the overall goodness of fit of the model against a general multinomial alternative are feasible.

For the four-item examples, only examinees with complete data (no non-response) for all four items were used. For items *WXYZ*, $N = 1048$ and 214 of the 256 cells of the table were filled (one of the 4×4 marginal tables of this table makes up Table 1); for items *RSTU*, $N = 976$ and only 156 response patterns were observed.

We also fitted several forms of the model to the entire set of 35 verbal items, and a subset of 12 of the technical knowledge items called "physical science" hereafter. In these analyses, all examinees were used ($N = 1178$) and non-response data to individual items were ignored (not placed in any category).

Results

Table 2 gives summary goodness of fit results for the four-item examples. Bock's (1972) model and Samejima's (1979) modification are both rejected at the $p < 0.05$ level for both sets of data. The model called "*ABCD*" is a form of (2) in which the vector δ is estimated, but restricted to be equal across items: this represents the hypothesis that a (constant) proportion of those who "don't know" select alternative *A*, a different proportion *B*, and so on; but the distribution over *A*, *B*, *C*, and *D* (regardless of the "correct"

TABLE 2

G^2 Values for Selected Models for the Four-Item Examples.

		Items RSTU		Items WXYZ	
Model	d.f.	G^2	p	G^2	p
Bock (1972)	231	307	<.01	277	<.02
Samejima (1979)	223	271	<.02	264	<.05
"ABCD"	220	258	<.04	262	<.02
"ABCD(C),ABCD(D)"	217	248	>.05	---	--
"11 per item"	211	245	<.05	---	--
"Correct vs. incorrect"	222	---	--	253	=.05
"Correct vs., each item"	219	---	--	253	=.05

TABLE 3
Estimated Parameters for Items R, S, T, and U.
Estimates for Four Items Above, Whole Test Below.
Parameters for Correct Response Underscored.

		Response DK	A	B	C	D
Item						
R	a_k	-1.1	-2.9	1.5	0.2	2.2
		-1.7	-1.0	1.1	0.3	1.9
	c_k	0.1	-3.8	2.9	-1.9	2.7
		0.3	-2.3	2.4	-2.5	2.1
	d_k		0.1	0.2	0.3	0.4
			0.25	0.25	0.25	0.25
S	a_k	-3.0	-0.3	1.2	2.8	-0.7
		-2.1	-0.6	1.2	2.3	-0.8
	c_k	0.1	0.6	-1.6	0.1	0.8
		2.1	0.5	-3.0	-0.6	1.0
	d_k		0.2	0.2	0.5	0.1
			0.2	0.2	0.4	0.2
T	a_k	-0.8	-1.8	-0.1	1.9	0.7
		-1.3	-0.9	-0.2	1.9	0.5
	c_k	-1.6	-3.2	0.2	2.6	2.0
		-0.9	-2.5	-0.1	1.8	1.6
	d_k		0.2	0.2	0.5	0.1
			0.2	0.2	0.4	0.2
U	a_k	-1.6	-0.4	-0.2	-0.8	3.0
		-1.9	0.5	0.0	-0.6	1.9
	c_k	0.0	-1.3	0.4	0.4	0.5
		-0.1	-2.0	0.5	0.8	0.8
	d_k		0.1	0.2	0.3	0.4
			0.25	0.25	0.25	0.25

response alternative) is the same for all items. For the *RSTU* data the model is a great improvement over Samejima's $\delta = 0$ model ($G^2 = 12.6$ on 3 d.f.), and for *WXYZ* it is not. The estimated vector **d** for *RSTU* for [*A*, *B*, *C*, *D*] is [.1, .2, .4, .3].

In *RSTU*, *D* is correct for the first and the last and *C* is correct for the others; so to test the hypothesis that (somehow) the correct alternative "attracts" guessing we tested "*ABCD*(*C*), *ABCD*(*D*)" in which the two pairs of items had different vectors δ. This model fits the *RSTU* data, and the improvement over *ABCD* is significant ($G^2 = 10.6$, d.f. = 3, $p < 0.02$); the parameter estimates are given in Table 3. The "saturated" version of (2), with 11 free parameters per item, does not fit the *RSTU* data significantly better.

Items *W*, *X*, *Y*, and *Z* do not have such a convenient structure of correct alternatives; but for these items "correctness" seemed more important than alternative position, so in the model "correct vs. incorrect" d_h is constrained to be one value (the same for all items) for the correct response and another for all the incorrect responses. This model (barely) fits the *WXYZ* data. The parameter estimates are given in Table 4. Forty percent of the DK guess the correct alternative; there is no evidence that this varies across items (The last line of Table 2 gives the G^2 for "correct vs. incorrect each item").

TABLE 4
Estimated Parameters for Items W, X, Y, and Z.
Estimates for Four Items Above, Physical Science Subset Below.
Parameters for Correct Response Underscored.

		Response				
		DK	A	B	C	D
Item						
W	a_k	-2.3	-0.2	2.0	0.9	-0.3
	c_k	0.5	0.7	-0.5	-1.9	1.1
	d_k		0.2	0.4	0.2	0.2
X	a_k	-0.8	0.6	-0.5	1.1	-0.4
		-4.0	1.2	1.1	2.5	-0.9
	c_k	1.6	-2.8	1.5	0.0	-0.3
		-0.8	-0.7	1.4	0.2	-0.1
	d_k		0.2	0.2	0.4	0.2
			0.2	0.3	0.3	0.2
Y	a_k	-0.5	-0.2	2.0	-1.2	0.0
		-3.3	0.2	2.6	-0.1	0.5
	c_k	-0.3	0.7	-1.0	0.7	0.0
		-1.8	1.1	-1.0	1.4	0.2
	d_k		0.2	0.4	0.2	0.2
			0.2	0.4	0.2	0.2
Z	a_k	-1.5	-0.7	-0.2	0.1	2.3
	c_k	0.4	0.4	-0.5	0.5	-0.8
	d_k		0.2	0.2	0.2	0.4

When the complete verbal data are fitted, the likelihood ratio test against a "general multinomial alternative" for the 1178 observations in a 4^{35} table is, of course, meaningless. Nevertheless, we report likelihood ratio tests between variously constrained versions of the model. The Samejima (1979) model is the most constrained form that we use for comparison. A model of the "*ABCD*" form of (2), in which $\boldsymbol{\delta}$ is estimated, equal across items, gives a reduction $G^2 = 41$ on 3 d.f., $p < 0.001$; the estimate of $\mathbf{d}$ was [.2, .3, .3, .2] for [*A*, *B*, *C*, *D*]. As with *RSTU*, a model in which $\boldsymbol{\delta}$ is constrained to be equal only among items with the same (letter) correct response fits significantly better: $G^2 = 72$ on 9 d.f., $p < 0.01$. No more complex models were considered, as estimation of 11 parameters per item for the very easy items (to which the majority of examinees responded correctly) would be estimating many parameters with little data. The estimated parameters for *RSTU* obtained in the analysis of the entire verbal set are given directly below their values for the four-item estimation in Table 3.

Analysis of the physical science test gave similar results: (2) in its "*ABCD*" form fitted significantly better than Samejima's model, $G^2 = 29$ on 3 d.f., $p < 0.01$. The estimate of $\mathbf{d}$ was again [.2, .3, .3, .2]. And the model in which $\mathbf{d}$ varied according to the letter-value of

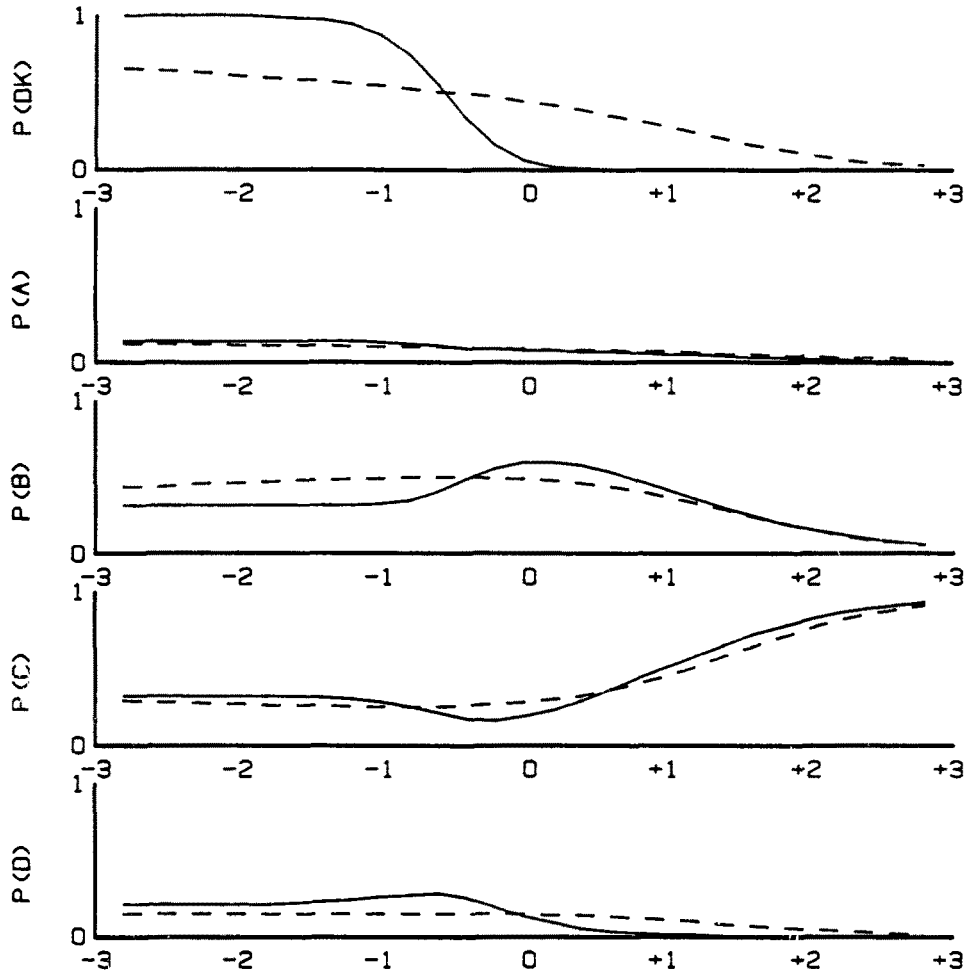

FIGURE 1
Five panels showing $P(k)$, $k = 0, 1, \ldots, m_j$ as a function of θ; the solid lines correspond to the probabilities for DK, and responses A, B, C, and D for item X, estimated as part of "physical science." The dashed curves use the estimates obtained with four items.

the correct alternative was better: $G^2 = 26$ on 9 d.f., $p < 0.01$. Items X and Y are in the physical science set, and their parameter estimates are given in Table 4.

Discussion

In cases in which the fit can be tested—with a thousand examinees and four-alternative items, four items—some forms of the model given by (2) and (5) fit item response data satisfactorily. This represents a major step forward in item analysis, because it is no longer necessary to look at deviations from the fit to examine items; we may examine the fit itself.

The solid curves in Figure 1 illustrate the trace lines for the "physical science" fit to item X, which is item 2 in Table 1, discussed above. From Table 1 we inferred that those who chose A or B might be more able than those who chose D. Indeed, the trace line for D restricts that response to those with values of $\theta < 0$ for all practical purposes, while B-responses come mostly from those with $\theta > 0$, and A is spread all over. The DK trace line indicates that most of those below $\theta = -.5$ have no idea, but more of them guess correctly than any other way.

The dashed curves in Figure 1 illustrate the trace lines for the "four-item" fit to item *X*. Note in Table 4 that the parameter estimates differ a great deal, but the solid and dashed trace lines in Figure 1 differ much less (except for DK) *in the middle.* DK is a completely inferred latent response and is not really very well-defined by four items. The other curves are more similar in the middle, where the data are; 95% of the population distribution lies between θs of -2 and $+2$. The different parameter estimates seem to affect the curves primarily at the extremes.

Unlike algorithms which make use of point-estimates of θ in the estimation of item parameters, and require both large numbers of examinees and large numbers of items, the MML estimation procedure used here should be consistent considering the number of examinees alone if the model is correct. That is, as the number of examinees becomes large the estimates for four items or 12 or 35 should all converge to the true values. So the dashed and solid lines in Figure 1 should be the same. They are very similar, and there are three possible reasons for the small differences observed between them. First, one thousand examinees may easily not be "asymptotic" for a table with 256 cells. Second, one or the other solution may not be completely converged; EM-algorithms can be slow near the

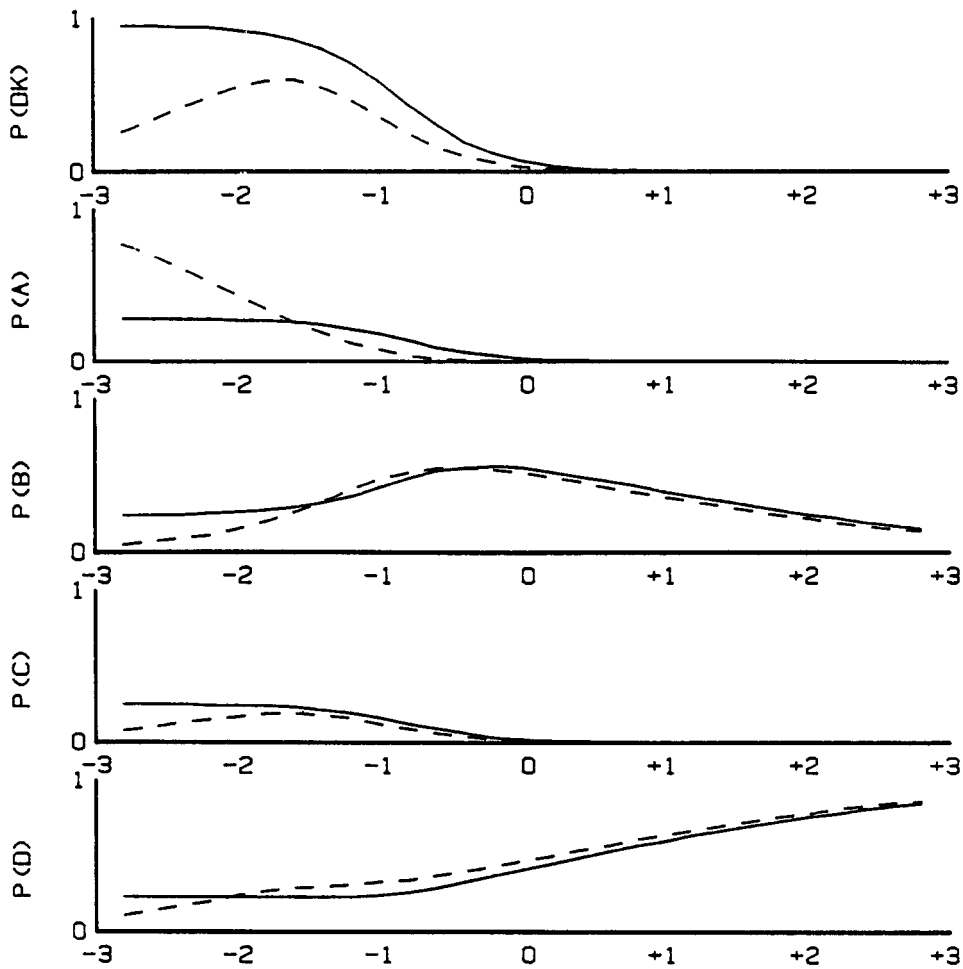

FIGURE 2
Five panels showing $P(k)$, $k = 0, 1, \ldots, m_j$ as a function of θ; the solid lines correspond to the probabilities for DK, and responses *A*, *B*, *C*, and *D* for item *R* estimated in the entire test. The dashed curves use the estimates obtained with four items.

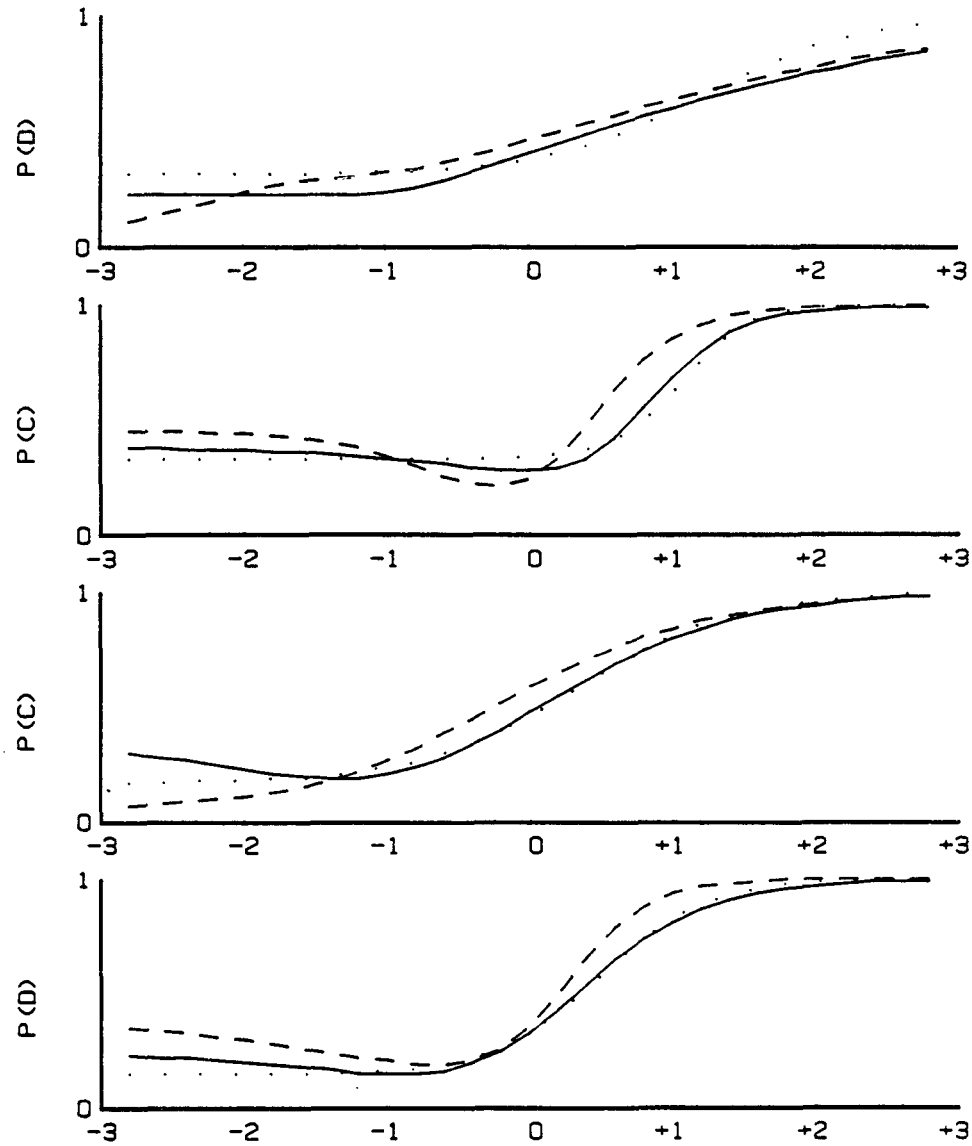

FIGURE 3
Four panels show three alternative fits of the trace line for the correct responses for (top to bottom) items R, S, T, and U: dots are the standard 3-PL curve, the solid line is the model of the present paper fitted with the entire test, and the dashed line is the present model fitted with only those items.

maximum. Third, and most likely, the model may be incorrect in that the latent dimension θ may be defined slightly differently in this set of four difficult items than in the entire test.

Figure 2 similarly illustrates the four-item and complete test estimates of the trace lines for item R, which demonstrates this effect more graphically. With estimates from the entire test, DK goes to unity for low θ. But the four-item (dashed) curve for DK rises for moderately low θ and then goes back down, and A goes to unity for low θ. That can't be right; but it occurs in the region below $\theta = -2$, where there are essentially no data and the model is extrapolating. Extrapolation is bad: the model is too flexible at the extremes. Flexibility was one of the goals of the model, but it seems we may have overdone it.

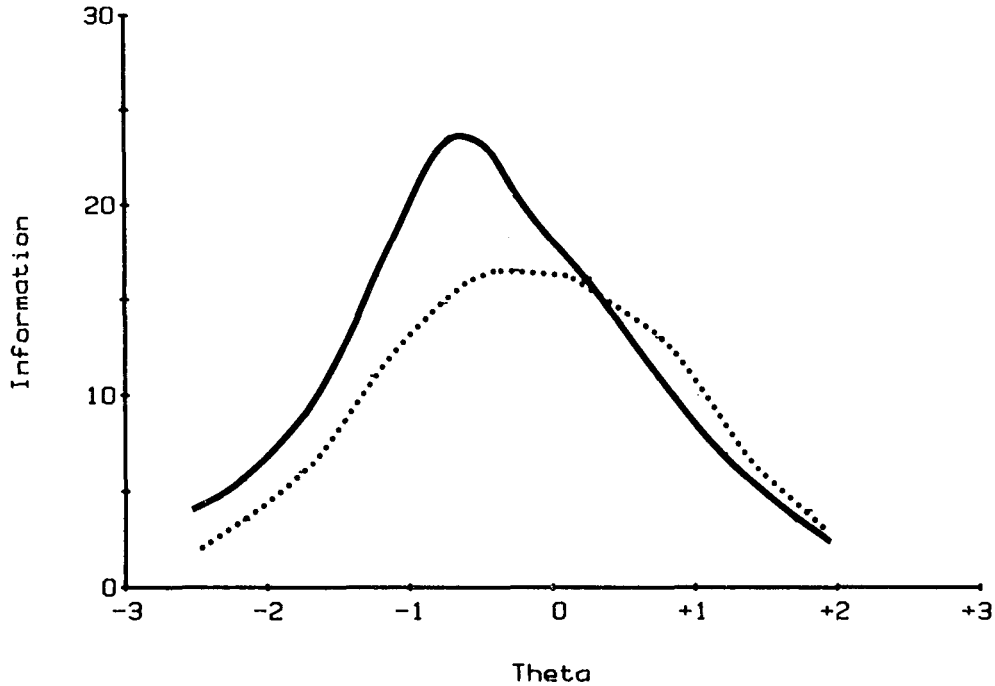

FIGURE 4
Test information curves for binary (dotted line) and multiple category (solid line) scoring for the verbal test, as functions of θ.

This is in marked contrast to traditional IRT models which go to unity, zero, or some asymptote at the extremes, and never misbehave there. Note that even with all of this strangeness, the other trace lines in Figure 2 are essentially identical between θs of -2 and $+2$.

Figure 3 shows the trace lines for the correct responses only for items *R*, *S*, *T*, and *U* from the four item set and the whole test, as well as traditional 3-PL curves estimated for the entire test (dotted lines) for comparison. If the "tail-wagging" to the left of $\theta = -2$ is ignored, the pairs of curves for the whole test are nearly identical and the four-item curves are somewhat deviant. That may be capitalization on chance in four items—or it may be that θ differs mildly in those items from its definition in the entire test. None of the curves are extremely different.

For the verbal test as a whole, the usual result (of Bock, 1972; Thissen, 1976) is obtained with respect to test information: Figure 4 shows the information curves for the multiple category scoring and the binary scoring (3-PL). For $\theta < 0$, information from incorrect responses increases total information by about 50%, equivalent to extending the test (for half the examinees) from 35 to 50 items. As in the earlier studies, the binary model appears to provide slightly more information for $\theta > 0$. This may be illusory; for difficult items with complex trace lines for the correct response, the 3-PL trace line may rise quite sharply from its asymptote to unity, and thus seem to provide more information in this region than the better-fitting trace line from the multiple category model. Over-estimation of the slope in the 3-PL model increases estimated information, but the information is over-estimated as well.

Figure 5 illustrates the process by which the multiple choice model provides increased information. The solid curves show the trace lines associated with a particular set

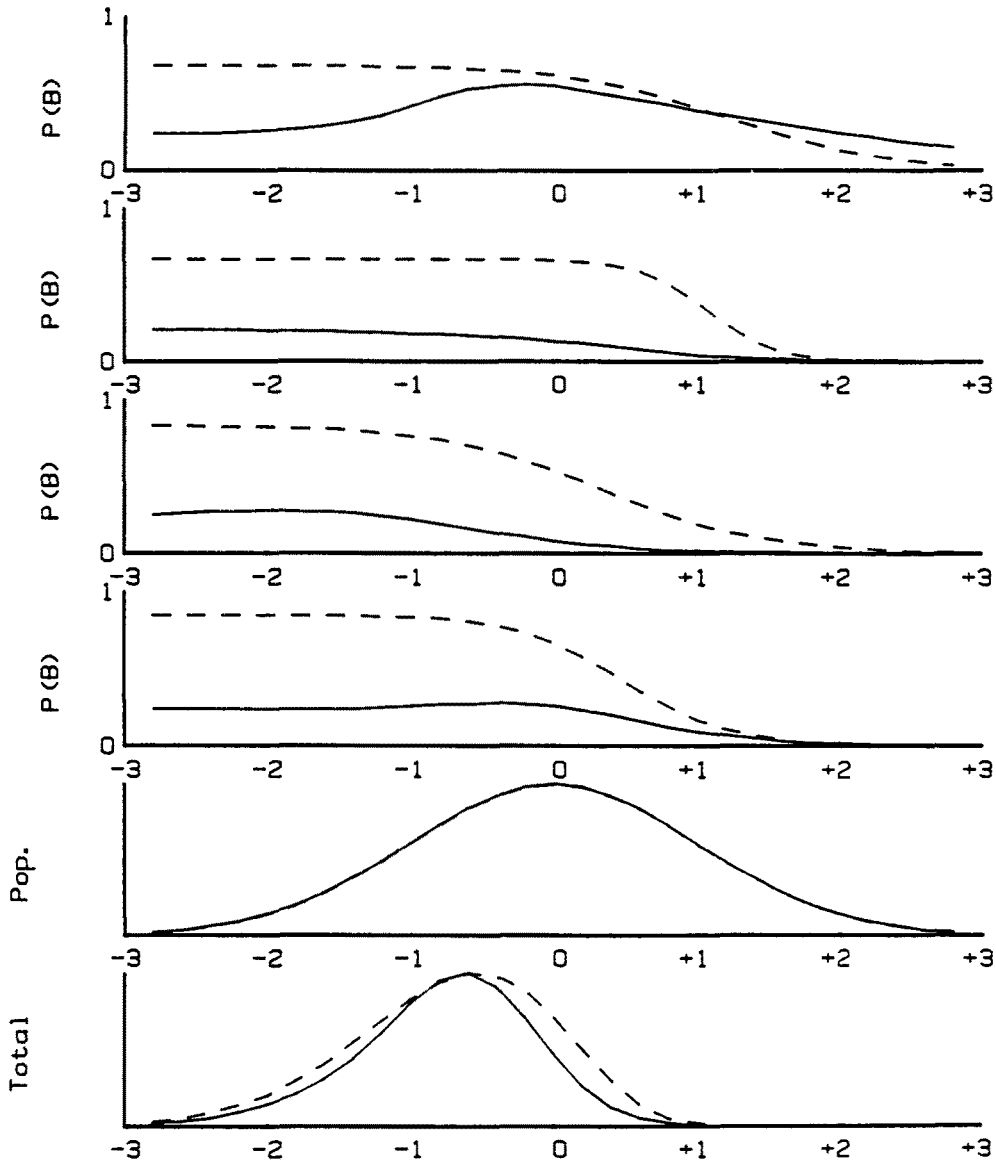

FIGURE 5
The solid lines are the trace lines for response pattern [*BBBB*] (all incorrect) for items *R*, *S*, *T*, and *U* from the present model, the population density, and the posterior density (labelled "Total"). Dashed lines give the corresponding curves for the same response pattern using 3-PL estimates.

of responses [*B*, *B*, *B*, *B*], all incorrect, for items *R*, *S*, *T*, and *U*. Then the $N(0, 1)$ population distribution is plotted with the product of all five curves, or posterior density, labelled "Total." The mode of that density is traditionally $\hat{\theta}$; it is about -0.7 (s.e. $= 0.6$) in this case. The dashed curves show the 3-PL trace lines, all incorrect, so they are all monotonic decreasing. That is not very informative; the (dashed) posterior is broader (s.e. $= 0.7$) and closer to zero ($\hat{\theta} = -0.5$) because there is less information so the estimator is "shrunk" more toward the mean of the population distribution. If the model approximates the world reasonably well, the multiple category scoring gives more information about people responding [*BBBB*] than can 3-PL.

General Discussion

The model proposed here provides the first practical, complete IRT item analysis for multiple choice tests, describing the performance of all of the response alternatives as functions of the trait being measured. The model and its fitting procedure are complex; its use is for "serious testing," not classroom exams. We have used sample sizes between one and two thousand examinees to estimate the parameters of the model for tests ranging in length from four to 35 items. Computer bills to calibrate such tests range through the hundreds of dollars on several systems we have used. Such expenditures are only justified for rigorous item analysis.

The parameters of the model are not readily interpretable for the most part. Semantic interpretation of the parameters is difficult, at best, given the form of the model. Further, the model has in common with the 3-PL extreme non-orthogonality in its parameterization: many apparently different sets of parameters give very similar trace lines. We consider the procedure to be one that estimates trace lines; the parameters are simply a means to that end. We base our interpretation on graphical output. Graphical presentation of the trace lines and/or their products with the population distribution gives thorough item analysis which has much to recommend it.

This sort of item analysis is sufficiently flexible that "bad" trace lines are fitted. These can then be observed and such items modified or eliminated in the course of empirical test development. That is not possible, in general, with simpler models which require examination of the residuals to find bad items. For item analysis, the model of (2) and (5) is clearly an excellent choice; for scoring tests (aka "estimating $\hat{\theta}$"), matters are more complicated.

Non-monotonic trace lines for the correct response have become a popular feature of recent IRT developments: Lord (1983) and Choppin (1983) have proposed item response models which permit "low-ability" non-monotonicity, in work on binary models independent from Samejima's (1979) multiple category proposal. There are a number of possible explanations for non-monotonicity on the left, as for the correct response in Figure 1. The only requirement for candidacy as an explanation for the effect is that it must account for "getting the right answer for the wrong reason." Bock (personal communication, May 9, 1983) suggests a name for the phenomenon: "positive misinformation."

Two sources of positive misinformation come to mind. The first is that the correct response for a particular item differs from the distractors on dimensions other than that which is intended and observable, given sufficient ability. Examinees of medium and high ability perceive the features and attempt the processing intended by the item-writer, while examinees of low ability see other features of the alternatives which cause them to select the correct one. These "other features" may be effectively "invisible" to individuals of higher ability, and therefore to the item-writers as well.

A second possible cause of non-monotonicity in the correct trace line is cheating. If low-ability individuals cheat (e.g., by copying a neighbor's answer, which is more likely to be correct than not for most items), then the resulting correct alternative trace lines will rise for those of very low ability.

It may also be possible that non-monotonic trace lines are estimation artifacts in small sets of items (like four). It may be that, with few monotonic correct-alternative trace lines to "orient" θ, the estimation procedure could "wrap around" and (effectively) place some of the high ability individuals on the left end of the θ-continuum. But this explanation can be discounted with longer tests, as the non-monotonicity there comes from correct response to an item being more likely from examinees who get most of the other items wrong than from those who get fewer of the other items wrong.

Distinguishing among these possibilities would require experimentation with the items and the testing situation. The model does, however, offer an item analysis which permits such phenomena to exhibit themselves in the fit, when they are present.

So the use of non-monotonic curves in item analysis may be required to fit the observed data. However, the use of non-monotonic trace lines in constructing the posterior density, a measure of the central tendency of which is to be called $\hat{\theta}$ and used to "score the test," gives rise to many potential problems. If the correct response trace line for item j is non-monotonic 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 item j; they would have been assigned a higher $\hat{\theta}$ if they had selected certain of the incorrect alternatives. Correct response trace lines which are non-monotonic on the left similarly "penalize" examinees of low ability who respond correctly. As measurement, this is all probably satisfactory: *conditional on the other item responses*, a correct response to item j may *not* imply higher ability; it may be more likely to be guessing or cheating. But this may be a problem for the test "as contest." Further, it might be difficult to defend such a method of test scoring against the onslaught of members of the bar before a jury. A case can, therefore, be made for either (a) rejecting the model or (b) rejecting the use of items with non-monotonic trace lines for the correct response.

Multiple category scoring of this nonlinear sort has certain bizarre properties even if non-monotonic trace lines are eliminated. For certain regions on the θ-continuum, selecting a particular incorrect response will increase $\hat{\theta}$ more than would selecting the actual correct response. For instance, in item R in Figure 2, for θ just below -1, selecting response B will increase $\hat{\theta}$ more than would selecting the correct response. But for θ around 1, selecting B will "penalize" the respondent! The contingencies are sufficiently complex that it is unlikely that the examinees could find a strategy to take advantage of the system. But the possibilities for legal difficulties are considerable under circumstances in which the test and scoring system must be disclosed. On the other hand, the quality of measurement for research purposes and in non-disclosed tests could be improved by these "bizarre" features; that is, after all, where the multiple category model obtains its additional information.

Conclusion

Item response models for multiple choice items have come of age: they fit the data. Questions remain about uses for these models. They produce excellent item analysis, but it is complex and best-represented graphically—that breaks with the tradition of item analysis with a few numerical summaries. Multiple category scoring clearly increases the information obtained in scoring a test—but again at the cost of complexity, and, potentially, controversy. The validity of test scored with such methods has not been examined here, nor can it be with data internal to the test in any event. Consideration of all of these matters is deferred to future work.

References

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

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

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

Bock, R. D., & Mislevy, R. G. (1982, August) Applications of EAP estimation in computerized adaptive testing. Paper presented at the annual meeting of the Psychometric Society, Montreal, Canada.

Choppin, B. (1983). *A two-parameter latent trait model.* (CSE Report No. 197). Los Angeles: University of California, Center for the Study of Evaluation, Graduate School of Education.

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

Haberman, S. (1974). *Subroutine MINIM* [Computer program]. In R. D. Bock & Bruno Repp (Eds.) *MATCAL: Double precision matrix operations subroutines for the IBM System 360/370 computers.* Chicago: National Educational Resources.

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

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika, 48,* 477–482.

Samejima, F. (1979). *A new family of models for the multiple choice item.* (Research Report No. 79-4) Knoxville: University of Tennessee, Department of Psychology.

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

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47,* 175–186.

Manuscript received 8/26/83
Final version received 8/13/84