Modeling Incorrect Responses to Multiple-Choice Items With Multilinear Formula Score Theory

Fritz Drasgow, Michael V. Levine, Bruce Williams, Mary E. McLaughlin, and Gregory L. Candell
University of Illinois

Multilinear formula score theory (Levine, 1984, 1985, 1989a, 1989b) provides powerful methods for addressing important psychological measurement problems. In this paper, a brief review of multilinear formula scoring (MFS) is given, with specific emphasis on estimating option characteristic curves (OCCS). MFS was used to estimate OCCS for the Arithmetic Reasoning subtest of the Armed Services Vocational Aptitude Battery. A close match was obtained between empirical proportions of option selection for examinees in 25 ability intervals and the modeled probabilities of option selection. In a second analysis, accurately estimated OCCS were obtained for simulated data. To evaluate the utility of modeling incorrect responses to the Arithmetic Reasoning test, the amounts of statistical information about ability were computed for dichotomous and polychotomous scorings of the items. Consistent with earlier studies, moderate gains in information were obtained for low to slightly above average abilities. *Index terms: item response theory, marginal maximum likelihood estimation, maximum likelihood estimation, multilinear formula scoring, option characteristic curves, polychotomous measurement, test information function.*

Multilinear formula score theory or multilinear formula scoring (MFS; Levine, 1984, 1989a, 1989b) is a nonparametric item response theory for which consistent and asymptotically efficient estimators of ability densities have been derived and programmed. This article describes and illustrates the use of MFS for nonparametric estimation of option characteristic curves (OCCS). It will be shown that the nonparametric methods based on MFS can discover or verify complex relations between ability and the inclination to select an option.

MFS provides a powerful new approach to substantive questions of long standing. These questions include determining the shapes of ability distributions and the magnitudes of differences between ability distributions of various groups, discovering the shapes of item characteristic curves (ICCS) for unidimensional and multidimensional tests, identifying biased and other faulty items, and assessing the extent to which two tests measure the same ability or trait.

MFS, the nonparametric theories of Samejima (1984), and the models developed by Sympson (1983, 1986; Sympson & Haladyna, 1988) provide alternatives to the parametric models devised by Bock (1972), Samejima (1979), Thissen and Steinberg (1984), and Masters (1982), and to the general class of parametric

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 3, September 1989, pp. 285–299
© Copyright 1989 Applied Psychological Measurement Inc.
0146-6216/89/030285-15$2.00

models included in Thissen and Steinberg's family of "**T**-matrix" models. Powerful estimation methods have been developed and programmed for the parametric models (e.g., MULTILOG; Thissen, 1987). This paper describes and illustrates the application of an estimation algorithm for nonparametric OCCs.

When the assumptions of one of the parametric models are correct (including, of course, the parametric form of the OCCs), estimates obtained from the MFS analysis described in this paper are obviously inferior to estimates of parameters of the correct parametric model. However, in applied measurement it is rare that a researcher *knows* that a specific parametric model is correct. Furthermore, standard methods for evaluating the fit of a model are not without problems of their own (e.g., the likelihood ratio test comparing a specific model to a general multinomial alternative is not feasible even for tests of modest length).

Nonparametric models, and MFS in particular, offer a way for researchers to *discover* the shapes of OCCs. This seems especially important for polychotomous measurement models (i.e., models that differentiate among incorrect response options) because a very large variety of shapes for OCCs are plausible. Moreover, it is anticipated that MFS estimates of OCCs will be proven to be consistent, that is, the estimates will be shown to converge with probability 1 to the true curves as sample size increases. (The methods used to study MFS density estimation are currently being modified in an attempt to devise and prove theorems about consistency for ICC and OCC estimates.)

Polychotomous models can be used to address a variety of problems in applied measurement. For example, such models can be used to obtain high rates of detection of inappropriate response patterns (Drasgow, Levine, & McLaughlin, 1987; Drasgow, Levine, McLaughlin, & Earles, 1987). Polychotomous models can also be used to study item bias, evaluate translations of tests and scales into different languages, and provide specific feedback to item writers about which distractors were effective and which were ineffective.

For testing practitioners to benefit from polychotomous measurement, they must be able to apply a polychotomous model and check its fit. To this end, MFS theory and its application to polychotomous measurement are described. Results obtained by applying MFS algorithms to estimate OCCs are then described. The estimation results were verified by checking them against empirical proportions of examinees selecting each option. A simulation study in which the OCCs were known was also used to verify the estimation method. Finally, increases in measurement precision due to polychotomous measurement were quantitatively evaluated by computing asymptotic standard errors of ability estimates at various ability levels; Bock (1972), Thissen (1976), Thissen and Steinberg (1984), and Sympson (1983, 1986; Sympson & Haladyna, 1988) have also demonstrated increased measurement accuracy with polychotomous measurement.

Review of Multilinear Formula Score Theory

This section reviews MFS as it is used in this paper. The theory is more general than outlined here (see Levine, 1984, 1985, for more details), but for the sake of clarity only the material required for the present application is described here.

Formula Scores

Let u_i denote the response to the ith item of an n-item test scored $u_i = 1$ if correct and $u_i = 0$ if incorrect. The u_is generate the *elementary formula scores*, which can be enumerated as

$$\begin{array}{l} 1 \\ u_1,\ u_2,\ \ldots,\ u_n \\ u_1u_2,\ u_1u_3,\ \ldots,\ u_{n-1}u_n \\ \cdot \\ \cdot \\ \cdot \\ u_1u_2\ \ldots\ u_n \quad . \end{array} \tag{1}$$

Traditional formula scoring (Lord & Novick, 1968, especially chap. 14) generally uses only linear scores. When there is neither omitting nor polychotomous scoring, *linear formula scores* are formulas with a constant term plus a linear combination of the binary item scores u_1, u_2, …, u_n. (When there is omitting and polychotomous scoring, a linear score is a constant plus a linear combination of binary variables indicating omitting and option choice.)

Levine's MFS theory generalizes traditional formula score theory by using quadratic scores (linear scores added to linear combinations of u_1u_2, u_1u_3, …, $u_{n-1}u_n$), cubic scores (quadratic scores plus linear combinations of products of item scores for three different items), and higher order scores. Most of the results in this paper were obtained with fifth-order scores. The new theory is called "multilinear" because frequent use is made of the fact that when all the scores but one are held constant, a "linear" score is obtained.

Regression Functions and the Canonical Space

In this paper it is assumed that the regression of u_i on the latent trait (ability) θ is a three-parameter logistic ogive:

$$\mathrm{E}(u_i|\theta = t) = c_i + \frac{1-c_i}{1+\exp[-Da_i(t-b_i)]} = P_i(t) \quad , \tag{2}$$

where D is a scaling constant set equal to 1.702, a_i is the discrimination parameter, b_i is the difficulty parameter, and c_i is the lower asymptote of the ICC. By local independence, the regressions of the elementary formula scores on the latent trait can then be written

$$\begin{array}{l} 1 \\ P_1(t),\ P_2(t),\ \ldots,\ P_n(t) \\ P_1(t)P_2(t),\ P_1(t)P_3(t),\ \ldots,\ P_{n-1}(t)P_n(t) \\ \cdot \\ \cdot \\ \cdot \\ P_1(t)P_2(t)\ \ldots\ P_n(t) \quad , \end{array} \tag{3}$$

where each $P_i(t)$ is a three-parameter logistic ICC.

There are 2^n regression functions listed above. More can be generated by taking linear combinations of the elementary formula scores and then computing their regressions on the latent trait. For example, the number-correct score

$$X = u_1 + u_2 + \ldots + u_n \tag{4}$$

has the regression

$$\mathrm{E}(X|t) = \sum_{i=1}^{n} P_i(t) \quad . \tag{5}$$

It can be shown (Levine, 1985) that the regression function of any statistic computed from an examinee's item responses is a linear combination of the 2^n regression functions listed above. The collection of regression functions of *all* linear combinations of elementary formula scores is called the *canonical space* (CS) of a test.

OCC Estimation

Identifiability considerations (Levine, 1984, 1985) can be used to show that density estimation can be viewed as selecting a function from the CS. This result greatly simplifies the problem of density estimation. For the present research, Levine's algorithms for density estimation were modified for OCC estimation, and again only functions in the CS are considered. In both density estimation and OCC estimation, a relatively small number of functions are selected from the CS, and the function the researcher wishes to estimate is then represented as a linear combination of the selected functions with unknown coefficients. The coefficients of the linear combination are estimated by marginal maximum likelihood. Clearly, if the estimated OCCs are to be close to the true OCCs, the set of linear combinations must contain accurate approximations of the true OCCs.

To ensure that the estimated OCCs converge to the true OCCs as sample size increases, it is necessary that the CS contains functions with a wide diversity of shapes. Moreover, the set of linear combinations of selected CS functions must allow representations of OCCs with substantially different shapes.

An Orthonormal Basis for the Canonical Space

From the above discussion, it is clear that a major step in an MFS analysis is the selection of the small number of functions from the CS. The selected functions form an *orthonormal basis* for the CS.

Selecting an orthonormal basis for the CS is analogous to finding the principal components of a set of variables. In a principal components analysis, the basic idea is to create a new set of variables, the principal components, so that each of the original variables can be written as a linear combination of a few principal components plus a small residual. The principal components are linear combinations of the original variables; analogously, the orthonormal basis functions are linear combinations of functions in the CS. Levine (1985, pp. 54–58) described a method for determining an orthonormal basis for the CS. (The orthonormal basis functions were called "coordinate functions" in Levine's earlier papers.)

A principal components analysis is valuable when there is a large number of original variables and when linear combinations of the first few principal components accurately approximate the original variables (i.e., "explain almost all of their variance"). In the same way, an orthonormal basis is useful when the functions in the CS can be accurately approximated by linear combinations of the first few orthonormal basis functions. For example, the ICC for the *i*th item can be written

$$P_i(t) = \sum_{k=1}^{K} \alpha_k h_k(t) \quad , \qquad (6)$$

where K functions, denoted $h_1(t), \dots, h_K(t)$, are used in the orthonormal basis and the α_k are the weights used in the linear combination. Just as in principal components, this representation is exact if K is sufficiently large. If only the first J functions are used, instead of all K functions (where $J < K$), then there is some error. However, the residual

$$P_i(t) - \sum_{k=1}^{J} \alpha_k h_k(t) = \sum_{J+1}^{K} \alpha_k h_k(t) \qquad (7)$$

will be small if the α_k are small for values of k larger than J. In fact, the area under the squared residual is exactly $\alpha^2_{J+1} + \alpha^2_{J+2} + \dots + \alpha^2_K$. This situation is again analogous to principal components analysis:

The first few principal components provide an accurate summary of the original variables when the eigenvalues of the *excluded* principal components are all small.

In each MFS analysis a parsimonious representation of one or another collection of functions in the CS is important. Techniques are available (Levine, 1984, 1989a) that yield basis functions that give small α_k for large k, at least for the collection of functions being analyzed. Most MFS density analyses require 6 to 8 basis functions for an adequate representation. To ensure enough flexibility to model unexpected OCC shapes, 10 basis functions were used in this study.

To recapitulate, the analysis begins by estimating ICCs from the dichotomously scored item responses. Widely available programs such as LOGIST (Wingersky, Barton, & Lord, 1982) and BILOG (Mislevy & Bock, 1983) can be used to this end. The estimated ICCs (and the assumption of local independence) are subsequently used to define the CS. Then a small number of orthonormal basis functions are selected so that the functions in the CS are well approximated by linear combinations of the orthonormal basis functions.

The Likelihood Function for OCC Estimation

The next step of the MFS analysis is to use the orthonormal basis functions to represent the OCCs. For technical reasons (see below), orthonormal basis function weights for *conditional option characteristic curves* (COCCs) are first estimated. A COCC gives the probability of an option choice given that an examinee does not select the correct option. A COCC equals its associated OCC divided by $1 - P_i(\theta)$. Hence the COCCs for an item sum to 1 for all θ values, whereas the OCCs sum to $1 - P_i(\theta)$, which becomes very small for large θ values.

Each OCC is then represented as the product of two linear combinations of the h_js, namely the representation of $1 - P_i$ and a COCC. At this point the OCC can be represented by a single set of weights. This is done by calculating weights b_j such that $\Sigma b_j h_j(\cdot)$ is approximately equal to $(1 - P_i)$ times the COCC value. (An exact representation is generally impossible because a product of two functions in the CS is not necessarily in the CS. In practice, however, the approximation is very accurate.)

Because OCCs and COCCs were not included in the set of functions used to define the CS, the mathematical question of how best to approximate the OCCs and COCCs with basis functions must be considered, as well as the substantive question of whether the basis functions *can* adequately approximate OCCs and COCCs. The analysis developed by Levine and Williams (1989) proceeds item-by-item with the weights for all the options (including omit as an option) to each item simultaneously estimated by marginal maximum likelihood. The log likelihood that is maximized with respect to the weights α_k is

$$L = \sum_{j=1}^{N} \log P(\mathbf{u}_j^*, v_{ij}^*) \quad , \tag{8}$$

where $\mathbf{u}_j^*$ is a vector containing the dichotomously scored item responses of the jth examinee and v_{ij}^* indicates the particular option on item i selected by examinee j. For a four-option multiple-choice item, $v_{ij}^* = 1$ if option A is selected, ..., $v_{ij}^* = 4$ if option D is selected, and $v_{ij}^* = 5$ if no response is made. Suppose all the items are recoded so that option A is always the correct response. Then Equation 8 can be rewritten as

$$L = \sum_{\substack{j=1 \\ v_{ij}^*=1}}^{N} \log P(\mathbf{u}_j^*) + \sum_{\substack{j=1 \\ v_{ij}^*\neq 1}}^{N} \log \int P(\mathbf{u}_j^*|t) P(v_{ij}^*|t,\, u_{ij} = 0) f(t)\, dt \tag{9}$$

where

$$P(\mathbf{u}_j^*|t) = \prod_{i=1}^{n} P_i(t)^{u_{ij}} [1 - P_i(t)]^{1-u_{ij}} \quad , \tag{10}$$

$$P(v_{ij}^{*}|t,\ u_{ij} = 0) = \sum_{k=1}^{J} \alpha_k h_k(t) \quad , \tag{11}$$

and $f(t)$ is the θ density. Notice that Equation 10 is the likelihood function for the three-parameter logistic model [i.e., Lord's (1980) Equation 4-20 and Hulin, Drasgow, & Parsons' (1983) Equation 2.6.2]. It is the α_ks in Equation 11 that must be estimated. (Actually, each option has its own set of J α_ks, but to avoid notational complexity another subscript to the α_ks was not added.)

It is important to observe that local independence is not used to derive Equation 9 from Equation 8; only the definition of conditional probability is used. Thus, even when the choice of particular incorrect options (e.g., omits) fails to obey the assumption of local independence, accurate estimates of COCCs can be obtained if the pattern of correct and incorrect answers satisfies local independence.

An Algorithm for Estimating the α_k

A quadratic programming algorithm was developed by Levine and Williams (1989) to obtain maximum likelihood estimates of orthonormal basis function weights for the COCCs in Equation 11. The weights α_k for the COCCs are easier to estimate than the weights for OCCs because the OCCs for easy items and OCCs for rarely chosen options are close to 0, which causes the α_k to become indeterminate; COCCs are not usually close to 0. Because the OCC at $\theta = t$ is equal to the COCC times $1 - P_i(t)$, the OCCs are available after the COCCs have been obtained. Moreover, because $1 - P_i(t) < 1$, accurately estimated COCCs imply accurate estimation of OCCs. The COCCs are intrinsically interesting as well as mathematically tractable because their shapes can be used to study the properties of effective distractors.

The quadratic programming methods used by Levine and Williams (1989) are convenient because they allow plausible constraints to be placed on the COCCs. One constraint is *positivity*: COCC estimates are not allowed to become negative. In the present analyses all COCCs were required to equal or exceed .001. A second constraint placed on COCCs is *smoothness*: The COCCs are not allowed to oscillate widely. The smoothness constraint can be implemented by restricting the third derivative of the COCCs to be less than .005. This condition can be thought of as requiring each small piece of the graph of the COCC to have a very accurate quadratic approximation. (A restriction on the second derivative would force the COCC to be locally linear and a first-derivative constraint would force the COCC to be locally constant.)

Conceptually, the constraints used when estimating COCCs serve the same function as Bayesian prior distributions. Specifically, they prevent the estimated COCCs from assuming implausible values (e.g., negative values) or implausible shapes (e.g., wild oscillations) that can result from maximum likelihood estimation when there is insufficient information. Weaker constraints are sufficient for larger sample sizes, but a systematic exploration of the types of constraints that are needed at different sample sizes has not yet been conducted.

Summary

The MFS analysis begins by deriving orthogonal basis functions $h_k(t)$ from ICCs, which can be estimated by a program such as LOGIST or BILOG. COCCs are represented as linear combinations of the basis functions in Equation 11, and marginal maximum likelihood estimates of the weights α_k in this equation are obtained. OCC values can then be obtained by multiplying COCC values by $1 - P_i$.

Estimation and Information

Dataset

The dataset used in these analyses was a spaced sample of 2,978 examinees (i.e., every fourth examinee was selected from the total sample of 11,914 examinees). The larger dataset is fully described in the *Profile of American Youth* (1982). The examinees were administered the 30-item Arithmetic Reasoning (AR) subtest of the Armed Services Vocational Aptitude Battery (ASVAB). Each item on this test has four options.

ICC Estimation

The first step in the analysis was to estimate ICCs from the dichotomously scored item responses. To this end, the item responses of the ASVAB examinees were scored dichotomously. All unanswered items were scored as incorrect (because skipping and not reaching were treated as a separate—and incorrect—response option). Then the LOGIST (version 2B) computer program (Wood, Wingersky, & Lord, 1976) was used to estimate item parameters. (Current versions of MFS programs estimate ICCs nonparametrically.) Estimates of item discrimination parameters ranged from about .5 to 2.0 and estimates of item difficulties varied from about -3.0 to 1.4 (mean = .14, standard deviation = .99).

Density Estimation

The θ density f in Equation 9 was estimated by the nonparametric method developed by Levine and Williams (1989). The density was represented as a linear combination of basis functions, and the weights were estimated by maximum likelihood. The weight vectors were restricted to a convex set determined by hypotheses about the shape of the unknown density. This was done by selecting a grid of closely spaced points and controlling the sign of differences between function values at consecutive points and second-order differences (i.e., differences between differences). Current versions of the program use derivatives rather than differences.

After experimenting with various shape hypotheses, the following conditions were selected. The density was constrained to be non-negative, to have non-negative second-order differences for θs between -4.8 and -3.1, to have nonpositive second-order differences for θs between $-.3$ and 1.0, to be monotonically increasing for θs between 3.1 and $-.3$, and to be monotonically decreasing for θs between 1.0 and 3.5. These conditions imply that the density will be unimodal between -3.1 and 3.5, that the mode will occur between $-.3$ and 1.0, and that the density will either decrease to a lower asymptote as θ decreases to -5 or will have a second mode in the left tail if such is indicated by the data. It was decided to allow a second maximum at very low θs because the data seemed somewhat better fit when bimodality was permitted. A substantive interpretation of bimodality is noted below.

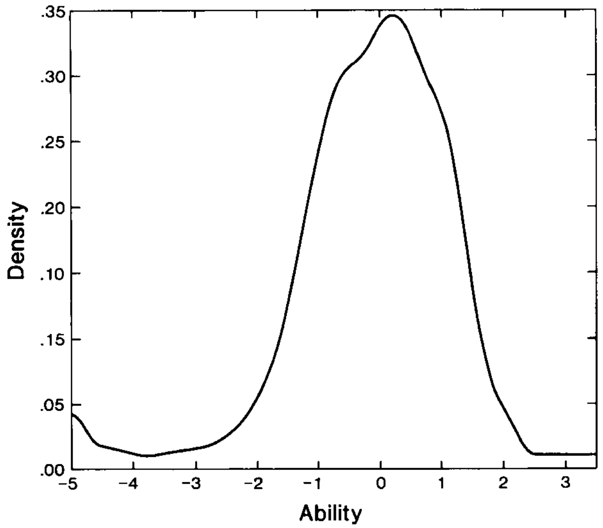
After some preliminary analyses, it was decided to remove examinees who answered less than half of the items. There were 87 such examinees, leaving 2,891 examinees for the density and COCC estimation.

Figure 1 shows the obtained density. It can be seen that the density is roughly bell-shaped with a mode near 0. The left tail turns up at low θs, suggesting a relatively large number of examinees with very low θs. One substantive interpretation of this turned-up left tail is that even among examinees who answered more than half of the items, there may have been some who were poorly motivated and did not make a serious attempt to do their best. In fact, examinees were all paid the same amount regardless of their scores, and consequently some of them may not have been adequately motivated to do their best. The test information function at $\theta = -5$ is very low; consequently, bimodality cannot be established unequivocally without a larger sample.

COCC Estimation

Four COCCs were estimated for each item—the three incorrect response curves and an omit curve. Omits included both skipped and not-reached items. Ten orthonormal basis functions were used in the

Figure 1
Ability Density for the Profile of American Youth Sample for the Arithmetic Reasoning Test



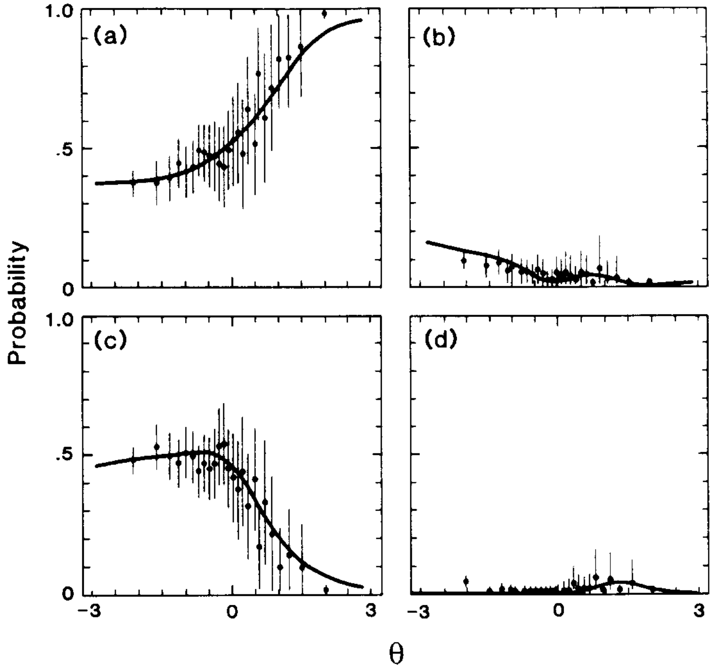
analysis. Thus, 30 weights (10 weights for each of three coccs) were estimated for each item. Because the sum of the coccs equals 1, the weights for the fourth cocc turn out to be a known linear combination of the weights for the other three.

Figures 2 and 3 contain plots of the four coccs estimated for each of two AR items. (Plots for all items are presented by Drasgow, Levine, Williams, McLaughlin, & Candell, 1987.) The solid curves indicate the estimated coccs.

An indication of the goodness of fit of the estimated coccs can be obtained by examining the vertical lines in each panel. These lines were obtained by computing three-parameter logistic θ estimates for all 11,914 examinees in the *American Youth* dataset, forming 25 strata on the basis of estimated θs by using the 4th, 8th, ..., 96th percentile points of the standard normal distribution as cutting scores, and then computing the proportion of examinees selecting each option among the subset of examinees who answered the item incorrectly. The centers of the vertical lines correspond to the observed proportions, and they are plotted above the category medians (the 2nd, 6th, ..., 98th percentile points of the standard normal distribution). The vertical lines represent approximate 95% confidence intervals for the observed proportions (± 2 standard errors, where the observed proportion is used to compute the standard error). Observed proportions of 0 and 1 are offset slightly from their true locations so that they will be visible.

For the most part, the AR items seem to be ordered by difficulty. Consequently, the 95% confidence intervals for the first few items on the test (including Item 6 shown in Figure 2) are very wide because these items were easy and few examinees chose incorrect options. Confidence intervals for later items were much narrower and provided a severe test for cocc estimates. Item 27, which is displayed in Figure 3, shows that the cocc estimates provided a very good description of option choice. Note that the cocc for the omit category (Figure 3d) lies below most observed proportions. This occurred because examinees with high omitting rates were excluded from the sample used to estimate coccs, but were included in the total sample used to compute sample proportions.

Figure 2
COCCs and Empirical Proportions for Item 6
(a) First Incorrect Option
(b) Second Incorrect Option
(c) Third Incorrect Option
(d) Omits



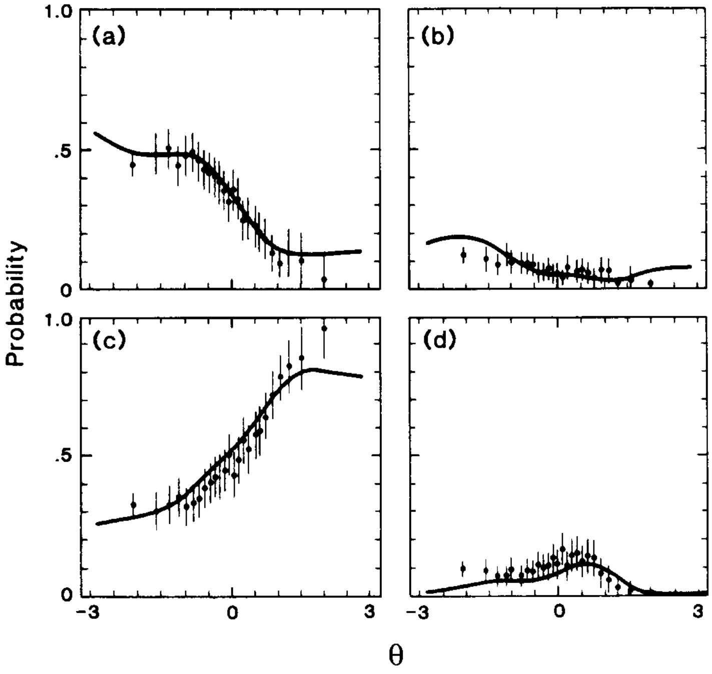
COCC Estimation Verification

Figures 2 and 3 show that MFS estimates of COCCs closely follow the actual patterns of item responses. This is also indicated in the figures for all 30 items in Drasgow, Levine, Williams, et al. (1987). It is difficult, however, to determine the accuracy of COCC estimates from these figures because the true COCCs are not known. To gain further insight into the properties of MFS estimates of COCCs, a simulation dataset of 3,000 response patterns was generated. Simulated θs were sampled from the standard normal distribution, probabilities of correct and incorrect responses were determined from the ICCs obtained by the LOGIST run previously described, and probabilities of option selections (for responses simulated to be incorrect) were computed using the MFS-estimated COCCs. Thus the assumptions used to estimate COCCs corresponded exactly to the way in which the dataset was generated.

COCCs were reestimated from the simulation dataset. The true θ density (the standard normal) was used in Equation 9 and the true ICC values were used to compute probabilities of correct and incorrect responses. The true θ density and ICC values were used in order to determine the errors of COCC estimates in a way that was not confounded with inaccuracies in density estimates and ICC estimates.

The results of the simulation study for two items are shown in Figures 4 and 5. (The reestimated COCCs for all 30 items are shown in Drasgow, Levine, Williams, et al., 1987.) Heavy lines represent

Figure 3
COCCs and Empirical Proportions for Item 27
(a) First Incorrect Option
(b) Second Incorrect Option
(c) Third Incorrect Option
(d) Omits

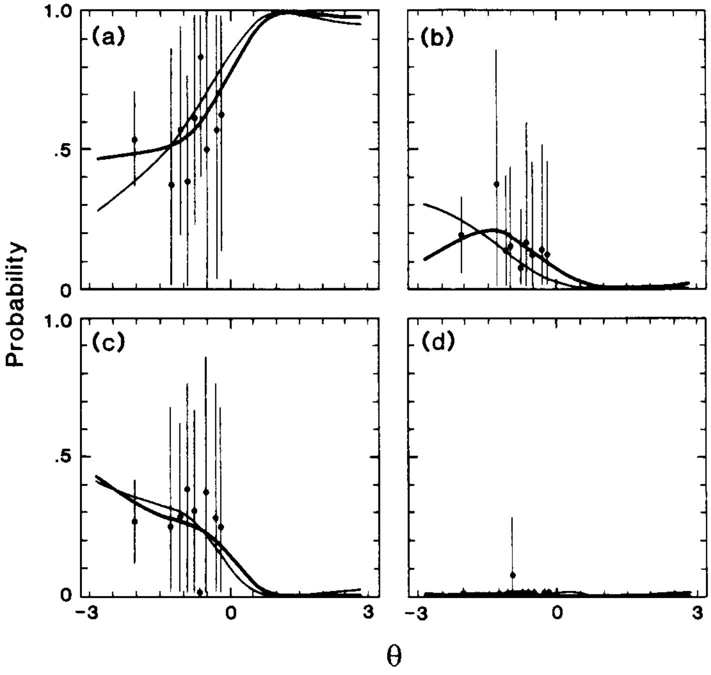


the reestimated COCCs and thin lines represent the true COCCs. Observed proportions and their approximate 95% confidence intervals are shown for the simulation sample of N = 3,000. The observed proportions were not plotted if five or fewer incorrect responses were made in a θ stratum.

Figure 4 displays the results for Item 2. It shows estimated COCCs that are very close to the true COCCs for all θ levels. This is remarkable because there were almost no incorrect responses made by simulated examinees with above-average θ. Results for other items show that well-estimated COCCs cannot always be expected when there are no data available: Some large differences between true and estimated COCCs did occur at high θ levels. The COCCs were, however, accurately estimated in θ ranges for which there were enough incorrect responses. More than a handful of examinees from each θ stratum answered Item 28 incorrectly; hence the COCCs for Item 28, shown in Figure 5, were accurately estimated.

Inspection of the results for all of the items in the simulation indicates that COCC values were accurately estimated when there were six or more incorrect responses in adjacent θ strata. Sometimes COCC values were well estimated when fewer incorrect responses were available, but this seemed to be a matter of chance. Also, note that in Figures 4 and 5 COCCs for the omit option were not underestimated as they were in the analysis of the real AR data. In the present analysis, all response vectors were used; there was no restriction on omitting as in the previous analysis. (In the simulation study, data were unidimensional in the sense that the probability of omitting depended only on θ, although it was permitted

Figure 4
Simulation COCCs, Estimated COCCs, and Empirical Proportions for Item 2
(a) First Incorrect Option
(b) Second Incorrect Option
(c) Third Incorrect Option
(d) Omits



to vary from item to item. It would have been more realistic to use a two-dimensional simulation model with examinees varying in both θ and tendency to omit.)

Information

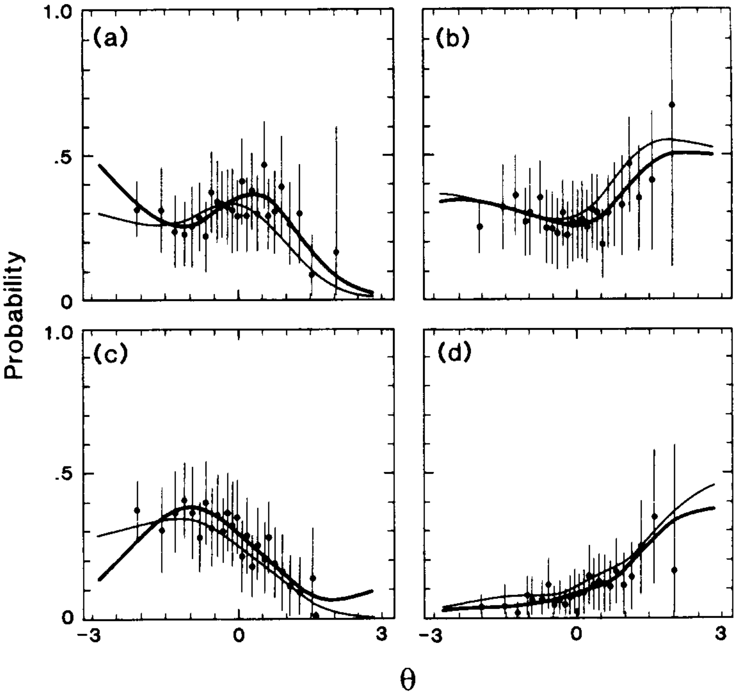
As noted by Bock (1972), Thissen (1976), Sympson (1983, 1986), and others, an important benefit of polychotomous item response theory models lies in their increased information about θ due to polychotomous scoring of item responses. Here the term "information" is used in its statistical sense to mean the expected squared derivative of the logarithm of the likelihood function. Because the asymptotic standard error of the maximum likelihood estimate of θ equals the square root of the reciprocal of the information function at θ, an increase in information due to polychotomous scoring is readily translated into percent test length reduction made possible by polychotomous scoring.

The information function of the three-parameter logistic model can be expressed as

$$\text{Information at } t = \sum_i \frac{[P_i'(t)]^2}{P_i(t)} + \sum_i \frac{[Q_i'(t)]^2}{Q_i(t)} \quad , \tag{12}$$

where $Q_i = 1 - P_i$ and P_i' and Q_i' are the first derivatives of P_i and Q_i. The information function of the

Figure 5
Simulation COCCs, Estimated COCCs, and Empirical Proportions for Item 28
(a) First Incorrect Option
(b) Second Incorrect Option
(c) Third Incorrect Option
(d) Omits



polychotomous model is

$$\text{Information at } t = \sum_i \frac{[P_i'(t)]^2}{P_i(t)} + \sum_i \sum_{j=2}^{J} \frac{[P_{ij}'(t)]^2}{P_{ij}(t)} \quad , \tag{13}$$

where P_{ij} is the OCC for option j on item i and P_{ij}' is its first derivative. The correct option makes the same contribution to information for both the dichotomous and polychotomous scorings, namely, the first term on the right sides of Equations 12 and 13. Thus any differences in information are entirely due to the treatment of incorrect responses. Samejima (1969) and Park (1983) have shown that

$$\sum_{j=2}^{J} \frac{[P_{ij}'(t)]^2}{P_{ij}(t)} \geqslant \frac{[Q_i'(t)]^2}{Q_i(t)} \quad , \tag{14}$$

and therefore any increase in information is entirely due to polychotomous scoring.

Information functions for the dichotomous and polychotomous modelings of the AR test are shown in Figure 6, which shows that there were moderate gains in information due to polychotomous scoring of the AR items for low to moderately high θs. Little or no information was gained for high-ability examinees; this latter finding is not surprising because high-ability examinees are expected to answer nearly all the items correctly.

Because the dichotomous model used here was a nested submodel of the polychotomous model, and both models used exactly the same ICCs, the information function for the polychotomous model cannot lie below the information function of the dichotomous model. Figure 6 illustrates this mathematical result. This relation is not necessarily true for other pairs of polychotomous and dichotomous models that are not similarly nested. Bock (1972, p. 50), Thissen (1976), and Thissen and Steinberg (1984, p. 515) presented information functions that showed more information for their *dichotomous* models for some ranges of θ values. Due to the nested relationship of the dichotomous and polychotomous models studied here, it appears that Figure 6 provides a better characterization of the difference in information functions across the two models than in earlier studies.

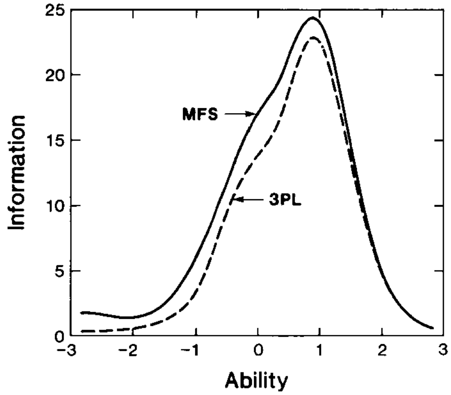
Note that the AR items were not written with polychotomous scoring in mind; hence the gains in information shown in Figure 6 are to some extent accidental. Larger gains might be realized if item writers knew the attributes of incorrect options that typically led to substantial increases in information.

Discussion

The test information function of the polychotomous model was found to be moderately larger than the three-parameter logistic information function for low to moderately high θ levels. Because there *is* information in incorrect options, it seems prudent to use it if items are expensive to write, the number of items that can be administered is severely limited, or very accurate θ estimates are required. Furthermore, the differences in items with informative incorrect options and items with essentially noninformative incorrect options can be studied systematically. It may be possible to identify different characteristics of these two types of items and thereby help item writers increase the information about ability provided by tests by writing items with highly informative incorrect options.

COCC estimation provides opportunities to improve testing by increasing the accuracy of θ estimates, discovering the shapes of OCCs, and improving the theory and practice of item writing. Applications in areas such as item and test bias (Lord, 1980, chap. 14), appropriateness measurement (Drasgow, Levine, & Williams, 1985; Levine & Drasgow, 1988), and adaptive testing (Lord, 1980, chap. 10) may also be fruitful. Hence, because there *is* useful information in incorrect responses, polychotomous item response models in general—and the MFS polychotomous model in particular—can make important contributions to psychological testing.

Figure 6
Information Functions for Dichotomous and Polychotomous Scorings of the Arithmetic Reasoning Test



References

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

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11,* 59–79.

Drasgow, F., Levine, M. V., McLaughlin, M. E., & Earles, J. A. (1987). *Appropriateness measurement* (AFHRL-TP-87-6). Brooks Air Force Base TX: Air Force Human Resources Laboratory.

Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M. E., & Candell, G. (1987). *Modeling incorrect responses to multiple-choice items with multilinear formula score theory* (Measurement Series 87-1). Champaign IL: University of Illinois, Department of Educational Psychology, Model-Based Measurement Laboratory.

Drasgow, F., Levine, M. V., & Williams, E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–86.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood IL: Dow Jones-Irwin.

Levine, M. V. (1984). *An introduction to multilinear formula score theory* (Measurement Series 84-5). Champaign IL: University of Illinois, Department of Educational Psychology, Model-Based Measurement Laboratory.

Levine, M. V. (1985). The trait in latent trait theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory/Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Levine, M. V. (1989a). *Classifying and representing ability distributions* (Measurement Series 89-1). Champaign IL: University of Illinois, Department of Educational Psychology, Model-Based Measurement Laboratory.

Levine, M. V. (1989b). *Parameterizing patterns* (Measurement Series 89-2). Champaign IL: University of Illinois, Department of Educational Psychology, Model-Based Measurement Laboratory.

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53,* 161–176.

Levine, M. V., & Williams, B. (1989). *Methods for estimating ability densities*. Manuscript in preparation.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

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

Mislevy, R. J., & Bock, R. D. (1983). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory/Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Park, R. K. (1983). *Application of a graded response model to the assessment of job satisfaction*. Unpublished doctoral dissertation, University of Illinois.

Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery (1982). Washington DC: Office of the Assistant Secretary of Defense (Manpower, Reserve Affairs, and Logistics).

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 34.

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.

Samejima, F. (1984). *Plausibility functions of Iowa Vocabulary Test items estimated by the simple sum procedure of the conditional P.D.F. approach* (Research Report No. 84-1). Knoxville: University of Tennessee, Department of Psychology.

Sympson, J. B. (1983). *A new item response theory model for calibrating multiple-choice items*. Paper presented at the meeting of the Psychometric Society, Los Angeles.

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

Sympson, J. B., & Haladyna, T. M. (1988). *An evaluation of "polyweighting" in domain-referenced testing*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

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

Thissen, D. (1987). *MULTILOG user's guide (version 5)*. Mooresville IN: Scientific Software, Inc.

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

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton NJ: Educational Testing Service.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST—A computer program for estimating examinee ability and item characteristic curve parameters* (Research Memorandum 76-6). Princeton NJ: Educational Testing Service.

Acknowledgments

This research was supported by Contract No. N00014-83K-0397, NR 150-518 and Contract No. N00014-86K-0482, NR 442-1546, from the Office of Naval Research (Michael V. Levine, Principal Investigator) and by Contract No. F41689-84-D-0002 from the U.S. Air Force Human Resources Laboratory to the Human Factors and Logistics Division of Universal Energy Systems, Inc., Dayton OH. The authors thank Malcolm Ree, Randy Park, James Earles, and J. Bradford Sympson for their help.

Author's Address

Send requests for reprints or further information to Fritz Drasgow, Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign IL 61820, U.S.A.