DOCUMENT RESUME

ED 264 267 TM 850 752

AUTHOR Mislevy, Robert J.; Bock, R. Darrell
TITLE Implementation of the EM Algorithm in the Estimation of Item Parameters: The BILOG Computer Program.
PUB DATE Jul 82
NOTE 15p.; In: Item Response Theory and Computerized Adaptive Testing Conference Proceedings (Wayzata, MN, July 27-30, 1982) (TM 850 744).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Entrance Examinations; *Computer Oriented Programs; Computer Simulation; Computer Software; *Estimation (Mathematics); *Item Analysis; *Latent Trait Theory; *Maximum Likelihood Statistics; Postsecondary Education; Psychometrics; Test Items
IDENTIFIERS BILOG Computer Program; *EM Algorithm; *Item Parameters; Law School Admission Test

ABSTRACT

This paper reviews the basic elements of the EM approach to estimating item parameters and illustrates its use with one simulated and one real data set. In order to illustrate the use of the BILOG computer program, runs for 1-, 2-, and 3-parameter models are presented for the two sets of data. First is a set of responses from 1,000 persons to five items of the Law School Admissions Test. Second is a set of simulated data of 1,000 persons to 18 items. The examples bring into focus the degree to which item parameters in the 3-parameter model can be recovered. The review discusses an EM Algorithm for estimating item parameters; solution for item parameters when person ability values are known; early computer program approaches; and the key elements of the Bock-Aitkin approach. Further described are extensions of the Bock-Aitkin approach, which include: (1) extension of the 3-parameter model; (2) prior distributions on item parameters; (3) estimation of the latent distribution; and, (4) different patterns of item attempts for different persons. (PN)

Implementation of the EM Algorithm in the Estimation of Item Parameters: the BILOG Computer Program

Robert J. Mislevy
International Educational Services

R. Darrell Bock
University of Chicago

Marginal maximum likelihood equations for estimating the item parameters in the 1- and 2-parameter normal ogive item response models were introduced by Bock and Aitkin (1981). The iterative solution of these equations bears strong resemblence to the EM algorithm of Dempster, Laird, and Rubin (1977). Over the past year, similar procedures have been implemented in the BILOG computer program (Bock & Mislevy, 1982) for estimating item parameters in the 1-, 2-, and 3-parameter logistic ogive models. Extensions of the original Bock and Aitkin solution include the simultaneous characterization of the latent population distribution and the incorporation of Bayes priors on item parameters, so that Bayes modal rather than maximum likelihood estimates may be obtained.

The purpose of this paper is to review the basic elements of the EM approach to estimating item parameters and to illustrate its use with one simulated and one real data set. The examples bring into focus a topic of occasional discussion in psychometric circles, namely, the degree to which item parameters in the 3-parameter model can be recovered.

An EM Algorithm for Estimating Item Parameters

The 3-parameter logistic ogive item response model for dichotomous test items, of which the 1- and 2-parameter models may be considered special cases, expresses the probability that person i will respond correctly to item j as

$$\begin{aligned} P_{ij} &= \text{Prob}(x_{ij}=1) \\ &= G_j + (1-G_j)\,\Psi[A_j(\theta_i - B_j)] \\ &= G_j + (1-G_j)\,\Psi[A_j\theta_i + C_j]\,, \end{aligned} \qquad [1]$$

where

x_{ij}, the item response, is 1 if correct and 0 if incorrect;
$\Psi(x)$ is the cumulative logistic function; $1/[1 + \exp(-x)]$;

G_j is the lower asymptote, often called the guessing parameter of item j, identically zero in the 1- and 2-parameter models;

A_j is the slope of item j, a constant over items in the 1-parameter model;

B_j is the threshold of item j;

C_j, equal to $-A_jB_j$, is the item intercept, introduced because estimation equations for the intercepts are simpler than those for item thresholds; and

θ_i is the ability of person i.

Given observed responses x_{ij} from N persons to n items, item parameters may be estimated. The main problem arising in this endeavor is that except in the 1-parameter model, the person parameters cannot be eliminated from the maximum likelihood estimation equations of the item parameters. In the presence of the so-called "nuisance" parameters, the standard results of maximum likelihood theory (e.g., consistency) do not apply.

A Solution When Ability Is Known

Estimation of item parameters would be straightforward if person ability values were known rather than implied by item responses. This is essentially the case that obtains in the bioassay setting, where the researcher controls the level of treatment dosage to each experimental unit, observes the proportion of units exhibiting the targeted response at each dosage level, then estimates an hypothesized underlying logistic or normal response function. In anticipation of the EM solution for item parameters, likelihood equations are presented for a logit regression problem that parallels the psychometric problem.

Suppose that, as in the bioassay setting, responses to each of n test items are observed from groups of persons at each of q specified points along the ability scale. Let N_{jk} be the number of responses to item j from persons with ability X_k and let R_{jk} be the number of these responses that are correct. Under the usual assumption of local independence, the total likelihood of a collection of observations of this type is as follows:

$$L = \prod_j \prod_k \frac{N_{jk}!}{(N_{jk}-R_{jk})!\ R_{jk}!} P_{jk}^{R_{jk}} (1-P_{jk})^{N_{jk}-R_{jk}} \qquad [2]$$

where

$$P_{jk} = G_j + (1-G_j)\ \Psi(A_jX_k + C_j)\ . \qquad [3]$$

The likelihood equations for the item parameters are the first derivatives of the log of Equation 2, equated to zero:

$$C_j: \quad 0 = \sum_k (R_{jk} - P_{jk}N_{jk})W_{jk} \qquad [4]$$

$$A_j: \quad 0 = \sum_k (R_{jk} - P_{jk}N_{jk})W_{jk}X_{jk} \qquad [5]$$

$$G_j: \quad 0 = (1-G_j)^{-1} \sum_k (R_{jk} - P_{jk}N_{jk})/P_{jk} \qquad [6]$$

where

$$W_{jk} = \frac{(1-G_j)P^*_{jk}(1-P^*_{jk})}{P_{jk}(1-P_{jk})} \qquad [7]$$

with

$$P_{jk}^* = \Psi(A_jX_k + C_j) \ . \qquad [8]$$

If the vector of zeros that solves these equations is unique and if the matrix of second derivatives of the log of Equation 2 is positive definite when evaluated at these values, then these values are the maximum likelihood estimates of the item parameters. The second derivatives are

$$C_j,C_j: \quad \sum_k P^*_{jk}(1-P^*_{jk})(G_jR_{jk}/P^2_{jk} - N_{jk}) \qquad [9]$$

$$C_j,A_j: \quad \sum_k P^*_{jk}(1-P^*_{jk})(G_jR_{jk}/P^2_{jk} - N_{jk})X_{jk} \qquad [10]$$

$$C_j,G_j: -\sum_k P^*_{jk}(1-P^*_{jk})R_{jk}/P^2_{jk} \qquad [11]$$

$$A_j,A_j: \quad \sum_k P^*_{jk}(1-P^*_{jk})(G_jR_{jk}/P^2_{jk} - N_{jk})X^2_{jk} \qquad [12]$$

$$A_j,G_j: -\sum_k P^*_{jk}(1-P^*_{jk})R_{jk}X_{jk}/P^2_{jk} \qquad [13]$$

$$G_j,G_j: \quad (1-G_j)^{-2} \sum_k [R_{jk}/P_{jk} - N_{jk} - R_{jk}(1-P_{jk})/P^2_{jk}] \qquad [14]$$

The solution of the likelihood equations may be accomplished by Newton-Raphson iterations, carried out item by item. The t + 1th iteration is

$$\begin{bmatrix} \hat{C}_j^{\,t+1} \\ \hat{A}_j^{\,t+1} \\ \hat{G}_j^{\,t+1} \end{bmatrix} = \begin{bmatrix} \hat{C}_j^{\,t} \\ \hat{A}_j^{\,t} \\ \hat{G}_j^{\,t} \end{bmatrix} - \begin{bmatrix} SDRV(C_jC_j) & SDRV(A_jC_j) & SDRV(C_jG_j) \\ SDRV(C_jA_j) & SDRV(A_jA_j) & SDRV(A_jG_j) \\ SDRV(C_jG_j) & SDRV(A_jG_j) & SDRV(G_jG_j) \end{bmatrix}^{-1} \begin{bmatrix} FDRV(C_j) \\ FDRV(A_j) \\ FDRV(G_j) \end{bmatrix} \qquad [15]$$

where all first and second derivatives are evaluated at the stage t estimates of the item parameters.

An Earlier Approach to the Problem

In the bioassay setting, where the criterion (dosage level) is known, the preceding solution is correct. One approach to the psychometric setting, where the criterion (ability) is not known, is to replace the unknown ability parameters with provisional estimates. This approach is employed by computer programs such as LOGOG (Kolakowski & Bock, 1973), LOGIST (Wood, Wingersky, & Lord, 1977), and BICAL (Wright & Mead, 1978). LOGOG, for example, employs for the 2-parameter model an algorithm similar to one outlined below:

1. Use persons' logits of percent correct as provisional ability estimates.
2. Standardize provisional ability estimates.
3. On the basis of provisional ability estimates, form groups of persons with apparently similar abilities.
4. Assuming all persons in a group have the same true ability--the mean of their provisional estimates--solve Equations 4 and 5 to estimate item parameters.
5. Using provisional item parameter estimates, re-estimate person abilities.
6. Return to Step 2.

Cycles of this type were repeated until convergence was attained--which, it was learned, became less likely as the number of items and/or persons decreased. A major problem is the unreliability of the estimates of person ability when the number of items was small; in such cases, person ability estimates were a poor substitute for the true values.

Key Elements of the Bock-Aitkin Approach

An alternative does exist, however--an alternative that derives from long-standing procedures in the statistical literature ir general and from an honorable tradition in psychometrics in particular (e.g., Kelley's paradox). The idea is this: Suppose that persons can be thought of as a random sample from a population in which ability is distributed in accordance with a distribution $g(\theta)$. Although each person's response vector x_i may not contain very much information about that person, it contains information about g. Taken together, the data of all persons may be sufficient to produce a fairly good characterization of g, which, in turn, may be used to condition and improve the inference about any individual person.

Now if g is a smooth distribution with finite moments, it may be approximated to any desired degree of accuracy by a discrete distribution over a finite number of points, i.e., a histogram. Let X_k, for $k = 1, \ldots, q$, be the points and let $A(X_k)$ be the densities at those points. By Bayes theorem, the posterior density of θ, given the response vector of person i is obtained as

$$P(X_k|\underset{\sim}{x}_i) = \frac{P(\underset{\sim}{x}_i|\theta_i=X_k)\ A(X_k)}{\sum_s P(\underset{\sim}{x}_i|\theta_i=X_s)\ A(X_s)} \qquad k=1, \ldots,q\ . \qquad [16]$$

Application to the estimation of item parameters is accomplished in the algorithm outlined below:

1. Using provisional estimates of item parameters, compute via Equation 1 the likelihood of each person's response pattern at each of the points, namely, $P(\underset{\sim}{x}_i|X_k)$.
2. Using given values (Bock & Aitkin, 1981) or provisional estimates (see below) of the densities $A(X_k)$ at each of the points, compute via Equation 16 the posterior probability that the ability of person $\underline{i}$ is X_k.
3. (E-Step) Pseudo-counts of numbers of items attempted and number of items correct are then obtained by effectively distributing the data from each person over the points in proportion to the likelihood of his/her being there as follows:

$$N_{jk} = \sum_i d_{ij} \, P(X_k|\underset{\sim}{x}_i)$$

$$= \sum_i d_{ij} \frac{P(\underset{\sim}{x}_i|X_k) \, A(X_k)}{\sum_s P(\underset{\sim}{x}_i|X_s) \, A(X_s)} \qquad [17]$$

and

$$R_{jk} = \sum_i d_{ij} \, x_{ij} \, P(X_k|\underset{\sim}{x}_i)$$

$$= \sum_i d_{ij} \, x_{ij} \frac{P(\underset{\sim}{x}_i|X_k) \, A(X_k)}{\sum_s P(\underset{\sim}{x}_i|X_s) \, A(X_s)} \qquad [18]$$

where d_{ij} is 1 if person $\underline{i}$ was presented item $\underline{j}$ and 0 if not.

4. (M-step) The maximum likelihood equations for the item parameters, Equations 4 through 6, are then solved with respect to the pseudo-counts.
5. Unless item parameters are unchanged from the previous cycle, return to Step 1.

Bock and Aitkin (1981) showed that for given $\underline{g}$, this procedure provides item parameter estimates that solve the marginal maximum likelihood equation

$$P(\text{data}|\text{item parameters}) = \prod_i P(\underset{\sim}{x}_i)$$

$$= \int_\Theta \prod_i P(\underset{\sim}{x}_i|\theta) \, g(\theta) \, d\theta \qquad [19]$$

The problem with the "nuisance" ability parameters has been solved by integrating over their range, rather than by replacing them with estimates as in LOGOG or conditioning them away as is possible with the 1-parameter model only.

As a result, the unreliability in the ability estimate for a person has been ameliorated. Rather than basing the estimation of item parameters on a larger number of unreliable person ability estimates, they have been based on the much more stable estimates of population densities at various points along the ability scale and expected proportions of correct response at those points.

Extensions of the Bock-Aitkin Approach

The basic approach to estimating item parameters outlined above was shown by Bock and Aitkin to be a maximum likelihood solution under the conditions of (1) the 1- and 2-parameter normal ogive model, (2) all persons being administered the same set of items, and (3) the weights $A(X_k)$ remaining fixed throughout the solution, i.e., persons were in effect assumed to be a random sample from a known distribution. (By comparing item parameter estimates obtained with different priors on ability, this latter assumption was shown to be relatively unimportant; the item parameters varied little in the examples shown.) Since the publication of the article, progress has continued in the investigation of this approach. A number of extensions have been incorporated into the BILOG program.

Extension to the 3-parameter model. Along with the change to the logistic rather than to the normal ogive response curve, the provision for obtaining item parameter estimates in the 3-parameter model has been included. It is known that item parameter estimation in this model has been problematic. Certain improvement is achieved in the EM approach by the use of the estimation of provisional densities and probabilities at selected points rather than of person abilities, since proper estimates always exist for the former but not necessarily for the latter in the 3-parameter model. Difficulties remain, however, from another source: The matrix of second derivatives of the log likelihood function is often poorly conditioned in the 3-parameter model. The inversion of this matrix, required in the Newton-Raphson solution of Equations 4 through 6, can become unstable. This practical problem at least partly motivates the extension discussed immediately below.

Prior distributions on item parameters. In order to provide for stable and "reasonable" item parameter estimates in the 3-parameter model and in all models for small samples of persons, provision has been made for the incorporation of prior distributions on item parameters. For lower asymptotes, beta priors are employed; for slopes, log-normal; for intercepts, normal. (Priors are rarely necessary for intercepts; provision is made to facilitate linking studies, since the prior distribution of a given parameter may be based on a previous estimate and its standard error). The program provides Bayes modal estimates rather than maximum likelihood estimates when priors are used. Uncorrelated priors are assumed, thereby effecting a modification of the first derivatives Equations 4 through 6 by a so-called "penalty" function and the addition to the second double derivatives Equations 9, 12, and 14 of an augmenting term. The terms added to the diagonal of the matrix of second derivatives improve conditioning of this matrix. Solutions may be obtained from any data set with the imposition of sufficiently strong priors on the item parameters, though judicious and thoughtful choice of priors is recommended.

Estimation of the latent distribution. The original Bock-Aitkin solution

assumes that persons are drawn from a specified distribution, normal or otherwise. The program now allows for the simultaneous estimation of the latent distribution if the user prefers. This is accomplished by revising the weights $A(X_k)$ at the beginning of each iteration as follows:

$$A^{(t+1)}(X_k) = (1/N) \sum_i P^{(t)}(X_k|\underset{\sim}{x}_i)$$

$$= \frac{1}{N} \sum_i \frac{P(\underset{\sim}{x}_i|X_k)\, A^{(t)}(X_k)}{\sum_s P(\underset{\sim}{x}_i|X_s)\, A^{(t)}(X_s)} \qquad [20]$$

The distribution is then restandardized to set the scale and location of the latent ability variable. Under this convention, a common slope parameter is estimated in the 1-parameter model while the standard deviation of the latent distribution is fixed at one; this is equivalent to the more typical practice of fixing all slopes at one but not restricting the ability parameters.

Different patterns of item attempts for different persons. As seen in Equations 17 and 18, there is no necessity of assuming that all persons are presented the same items. This feature is of particular value in the assessment setting because item parameters may be estimated from data gathered in highly efficient multiple-matrix sampling designs where each person responds to only one to five items in a scale. Despite the sparsity of data for each person proscribing the estimation of his/her ability, it is no barrier to iteratively building up the estimates of population densities and item proportions correct at the points X_k. Persons with few responses are spread more broadly and persons with more responses are spread less broadly, each in accordance with the information conveyed by his/her response pattern.

Examples

In order to illustrate the use of the BILOG program, runs for 1-, 2-, and 3-parameter models are presented for two sets of data. First is a set of responses from 1,000 persons to five items of the Law School Admissions Test (LSAT), a data set which has been analyzed in the past by Bock and Lieberman (1970), Bock and Aitkin (1981), Andersen (1973), Andersen and Madsen (1977), and Thissen (1982). These data have been found to be well fit by a 1-parameter logistic item response model and a normal distribution of ability. Second is a set of simulated data of 1,000 persons to 18 items. The known parameters of the items, which include lower asymptotes, may be compared with the estimated values.

Example 1: LSAT

The five items of the LSAT analyzed by Bock and Lieberman in 1970 and others since were, on the whole, rather easy for the persons in the sample; about 30% of the examinees answered all five items correctly. It has been found by Andersen (1973) that the data are well fit by a 1-parameter logistic ogive model and an underlying normal distribution of ability. These data were subjected to

item analysis via the 1-, 2-, and 3-parameter logistic models with BILOG, all under the assumption of an underlying normal distribution.

Table 1 presents the resulting item parameter estimates and, for the 1- and 2-parameter solutions, a likelihood ratio test of fit against a general multinomial alternative (see Bock & Aitkin, 1981). A straight maximum likelihood solution could not be obtained for the 3-parameter model, so the solution shown incorporates weak prior distributions on both slopes and asymptotes. The slopes had log normal prior distributions with means of zero (i.e., slopes of one) and standard deviations of two (slope values corresponding to a range of two standard deviations would be .018 and 54.598); asymptotes had a beta prior with parameters 1.25 and 5.75 (roughly comparable to saying with the weight of five observations that the asymptotes were .05). The formula for the likelihood ratio test was applied to the 3-parameter solution, but it must be noted that its distribution is not chi-square because the parameter estimates are modes of posteriors, not maximums of the likelihood function; its value, gauged in comparison with the degrees of freedom appropriate to a true maximum likelihood solution for the 3-parameter model, may be considered a somewhat more conservative index of fit.

Table 1
LSAT Item Parameter Estimates

Model	Chi-Square	df	Item	Threshold	Slope	Asymptote
1-P	9.90	19	1	-3.482	.788	.000
			2	-1.270	.788	.000
			3	-0.305	.788	.000
			4	-1.659	.788	.000
			5	-2.664	.788	.000
2-P	7.74	12	1	-3.318	.836	.000
			2	-1.356	.731	.000
			3	-0.279	.891	.000
			4	-1.845	.697	.000
			5	-3.074	.669	.000
3-P	9.27	7	1	-3.217	.831	.049
			2	-1.176	.752	.048
			3	-0.127	1.207	.029
			4	-1.704	.694	.048
			5	-3.114	.624	.050

It is no surprise to see that the 1-parameter model fits the data well and that the 2-parameter model fits even better but not sufficiently better to justify the additional parameters estimated. As noted by Thissen (1982), the 1-parameter solution agrees (after rescaling) with Andersen's conditional maximum likelihood solution (Andersen, 1973).

It is somewhat of a surprise to see that the 3-parameter solution appears to fit poorer than the 2-parameter solution, but this is because a maximum likelihood solution was not attained; the resulting parameter estimates depend not

only on the data but on the priors. Bock and Lieberman (1970), estimating intercepts and slopes for different fixed values of asymptotes, found that asymptotes of zero did indeed fit best. It may be seen from the estimates of asymptotes that the only item which shows much difference from the prior is that of Item 3--the only item sufficiently difficult to provide much information about an asymptote. For this item, the information pushes the asymptote value down in the direction of zero.

Example 2: Simulated Data

Responses were generated from a random sample of 1,000 simulated examinees from a standard normal distribution to an 18-item test, in accordance with a 3-parameter logistic ogive item response model. The generating item parameters are shown in Table 2. There are essentially two groups of nine items each. In the first group, all slopes are 2.0 and all lower asymptotes are .05; thresholds range from -2.0 to +2.0 in increments of .5. In the second group, all slopes are 2.0 and all asymptotes are .25; thresholds again range from -2.0 to +2.0 in increments of .5. The broad range of difficulty of the items is reflected in their resulting proportion-correct values, which ranged from .11 to .96 correct. Item-test biserials ranged from .4 to .8.

Table 2
Generating Values of Item Parameters for Simulated Data Example

Item	Threshold	Slope	Asymptote
1	-2.00	2.00	.05
2	-1.50	2.00	.05
3	-1.00	2.00	.05
4	-0.50	2.00	.05
5	0.00	2.00	.05
6	0.50	2.00	.05
7	1.00	2.00	.05
8	1.50	2.00	.05
9	2.00	2.00	.05
10	-2.00	2.00	.25
11	-1.50	2.00	.25
12	-1.00	2.00	.25
13	-0.50	2.00	.25
14	0.00	2.00	.25
15	0.50	2.00	.25
16	1.00	2.00	.25
17	1.50	2.00	.25
18	2.00	2.00	.25

BILOG solutions for the 1-, 2-, and 3-parameter models are shown in Table 3. The 1- and 2-parameter solutions are straight maximum likelihood solutions, with the normal distribution of persons assumed. The 3-parameter solution required priors on all item parameters, the specification of which is described in

Table 3
Item Parameter Estimates for Simulated Data
for the 1-, 2-, and 3-Parameter Models

Item	Inter-cept	SE	Slope	SE	Threah-old	SE	Diaper-sion	SE	Asymp-tote	SE	Chi-Square	df	Prob
1-Parameter Model													
1	-3.632	.141	1.197	.015	-3.035	.141	.836	.011	.0	.0	7.7	9	.5640
2	-3.324	.117	1.197	.015	-2.777	.117	.836	.011	.0	.0	26.3	9	.0019
3	-2.083	.089	1.197	.015	-1.741	.089	.836	.011	.0	.0	29.3	9	.0006
4	-1.415	.081	1.197	.015	-1.182	.082	.836	.011	.0	.0	46.1	9	.0000
5	-0.384	.074	1.197	.015	0.320	.075	.836	.011	.0	.0	20.3	9	.0161
6	0.391	.078	1.197	.015	0.327	.079	.836	.011	.0	.0	39.8	9	.0000
7	1.272	.084	1.197	.015	1.063	.085	.836	.011	.0	.0	25.7	9	.0024
8	1.885	.095	1.197	.015	1.575	.096	.836	.011	.0	.0	8.5	9	.4840
9	2.196	.105	1.197	.015	1.835	.105	.836	.011	.0	.0	29.9	9	.0005
10	-4.603	.170	1.197	.015	-3.847	.170	.836	.011	.0	.0	30.0	9	.0005
11	-2.867	.109	1.197	.015	-2.396	.110	.836	.011	.0	.0	4.2	9	.8961
12	-2.619	.100	1.197	.015	-2.188	.101	.836	.011	.0	.0	23.9	9	.0046
13	-1.616	.081	1.197	.015	-1.350	.082	.836	.011	.0	.0	19.7	9	.0196
14	-0.818	.072	1.197	.015	-0.684	.073	.836	.011	.0	.0	20.7	9	.0142
15	-0.301	.070	1.197	.015	-0.251	.070	.836	.011	.0	.0	26.5	9	.0018
16	0.275	.071	1.197	.015	0.230	.072	.836	.011	.0	.0	28.6	9	.0008
17	0.669	.071	1.197	.015	0.559	.072	.836	.011	.0	.0	57.1	9	.0000
18	0.837	.073	1.197	.015	0.700	.073	.836	.011	.0	.0	56.8	9	.0000
All Items											501.5	162	.0000
2-Parameter Model													
1	-3.587	.142	1.513	.116	-2.368	.149	.660	.050	.0	.0	3.7	8	.8821
2	-3.341	.121	2.008	.105	-1.664	.121	.498	.026	.0	.0	8.5	8	.3370
3	-1.982	.092	1.922	.105	-1.032	.095	.520	.029	.0	.0	13.5	8	.0958
4	-1.332	.087	2.168	.122	-0.614	.089	.461	.026	.0	.0	19.5	8	.0123
5	-0.154	.075	1.490	.103	-0.104	.088	.672	.046	.0	.0	5.6	8	.6933
6	0.695	.081	1.747	.113	0.398	.088	.573	.037	.0	.0	19.2	8	.0141
7	1.522	.085	1.368	.092	1.113	.097	.732	.049	.0	.0	22.7	8	.0038
8	2.111	.096	1.190	.089	1.774	.113	.840	.063	.0	.0	14.6	8	.0673
9	2.275	.103	0.735	.092	3.093	.199	.360	.170	.0	.0	13.4	8	.0995
10	-4.806	.175	2.235	.126	-2.149	.173	.447	.026	.0	.0	12.3	8	.1356
11	-2.657	.109	1.318	.101	-2.016	.122	.758	.058	.0	.0	4.5	8	.8095
12	-2.487	.102	1.562	.098	-1.593	.108	.640	.040	.0	.0	15.2	8	.0543
13	-1.418	.081	1.333	.090	-1.150	.099	.811	.060	.0	.0	12.0	8	.1510
14	-0.625	.072	1.034	.085	-0.604	.106	.967	.080	.0	.0	15.1	8	.0566
15	-0.117	.068	.885	.079	-0.133	.122	1.130	.101	.0	.0	10.9	8	.2067
16	0.446	.070	.944	.081	0.473	.114	1.059	.091	.0	.0	20.8	8	.0079
17	0.761	.069	.501	.071	1.521	.291	1.997	.283	.0	.0	15.2	8	.0558
18	0.918	.071	.466	.072	1.971	.338	2.148	.330	.0	.0	12.6	8	.1250
All Items											239.3	144	.0000
3-Parameter Model													
1	-3.363	.367	1.328	.167	-2.532	.280	.753	.094	.053	.032	7.6	7	.3729
2	-3.232	.407	1.956	.192	-1.652	.275	.512	.080	050	.030	6.8	7	.4545
3	-1.804	.332	1.795	.138	-1.005	.276	.557	.043	.052	.030	6.7	7	.4642
4	-1.131	.346	1.936	.138	-0.584	.303	.517	.037	.035	.021	19.3	7	.0074
5	-0.027	.348	1.463	.123	-0.018	.338	.683	.057	.040	.023	5.1	7	.6537
6	0.852	.454	1.803	.154	0.472	.450	.554	.047	.035	.016	9.3	7	.2278
7	1.991	.551	1.832	.167	1.087	.577	.546	.050	.038	.013	11.1	7	.1313
8	2.588	.643	1.542	.183	1.679	.692	.649	.077	.030	.011	5.8	7	.5592
9	3.089	.814	1.266	.215	2.439	.901	.790	.134	.039	.013	8.1	7	.3277
10	-4.325	.481	1.858	.233	-2.327	.301	.538	.067	.052	.032	14.7	7	.0400
11	-2.571	.308	1.283	.135	-2.004	.249	.779	.082	.051	.031	2.2	7	.9443
12	-2.217	.376	1.605	.157	-1.382	.316	.623	.061	.186	.051	7.0	7	.4317
13	-1.101	.404	1.442	.145	-0.764	.390	.693	.070	.196	.052	18.3	7	.0110
14	-0.125	.561	1.465	.183	-0.086	.573	.682	.086	.214	.048	14.1	7	0484
15	0.484	.545	1.432	.171	0.338	.565	.698	.083	.192	.037	6.7	7	.4570
16	1.274	.613	1.765	.188	0.722	.637	.567	.061	.156	.026	8.3	7	.3080
17	1.670	.680	1.009	.180	1.654	.764	.990	.176	.164	.033	18.6	7	.0096
18	2.093	.792	1.193	.210	1.756	.877	.839	.148	.165	.027	11.2	7	.1277
All Items											181.0	126	.0010

greater detail below. The indices of goodness of fit that accompany the estimates are not true likelihood chi-squares, but approximations based on combining persons into 10 homogeneous groups on the basis of their Bayes ability estimates. Counts of correct responses observed in each group were then compared with those expected under the assumption that all persons in a group have the same true ability.

The 1-parameter solution exhibits biases in both thresholds and slopes, as compared with the generating values. Although all items have the same generating slope of 2.0, the common value estimated is only 1.2, due to the attenuation caused by the nonzero lower asymptotes. There is a tendency for difficult items to fit more poorly than easy items, and for items of the second group (with higher asymptotes) to fit more poorly than items of the first group (with lower asymptotes).

The 2-parameter solution represents a marked improvement in fit. Many items, particularly easier items, are well explained by this solution. Serious biases are apparent, however, in the slope estimates. Again, because of the nonzero lower asymptotes, slopes are consistently underestimated to a degree that increases with difficulty and with the values of the asymptote itself.

The 3-parameter solution represents another, though less impressive, improvement in fit; the solution required prior distributions on intercepts, slopes, and asymptotes. Normal priors with mean zero and standard deviation two were placed on all intercepts. It may be seen that the data effectively dominated the prior in this case, as considerable information about intercepts is present in the data. Log-normal priors with mean .588 and standard deviation .500 were placed on slopes; this corresponds roughly to a prior mean of 1.8 and a standard deviation of 1.0 for the slopes themselves, suggesting a prior belief that slopes would probably range between about .5 and 5.0. Beta priors with parameters (3.5, 47.5) were placed on asymptotes for the first 11 items and with parameters (11, 41) for the last 7 items; this corresponds to saying with the weight of 50 observations that the asymptotes were .05 for the first 11 items and .20 for the last 7. These values were obtained by inspecting plots of the residuals from the 2-parameter solution, as illustrated by Figure 1.

Although the 3-parameter solution provides an adequate fit to the data, with a chi-square ratio less than one and a half, discrepancies remain between final parameter estimates and generating values. For the second group of items in particular, both thresholds and slopes tend to be too low. The apparent paradox of adequate fit but imperfect recovery of item parameters is resolved at least partially by an examination of estimated and observed response curves. Figure 1 plots data for Item 16 under the 2-parameter solution; Figure 2 plots the data for the same item under the 3-parameter solution. Despite nontrivial differences in estimates of item parameters (.5 vs. .7 for threshold, .9 vs. 1.8 for slope, 0 vs. .16 for asymptote), both curves are able to explain observed proportions of correct response in the region where the majority of persons are to be found. Despite the differences in their parameters, the 2- and 3-parameter curves are not very different with respect to the data at hand.

Figure 1
Observed and Expected 2-Parameter Logistic Response Curve for Item 16
(Smooth Line is Fitted Response Curve; "X" Represents Proportion Correct of a Group of Persons with Approximately Similar Abilities; Vertical Bars around Curve Represent Two Standard Errors around Expected Group Proportions Correct)

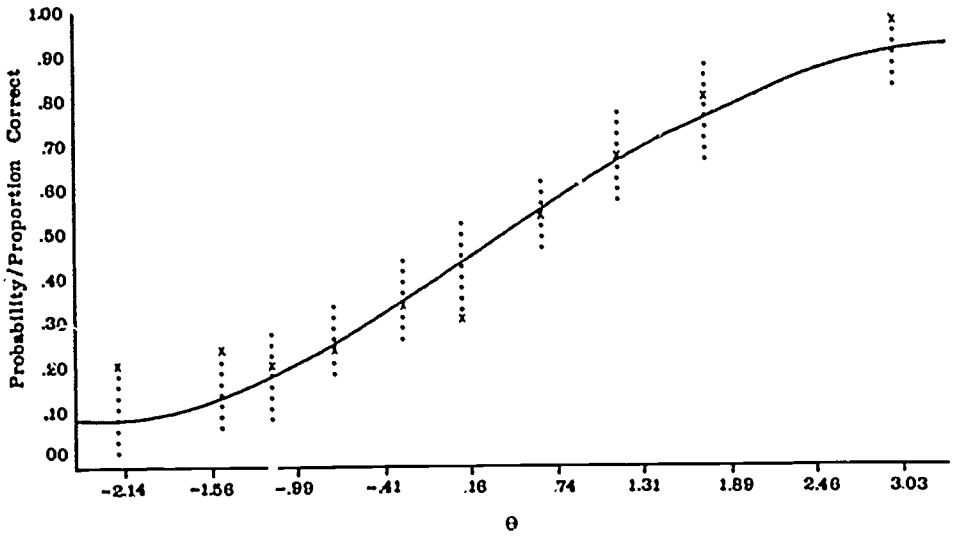
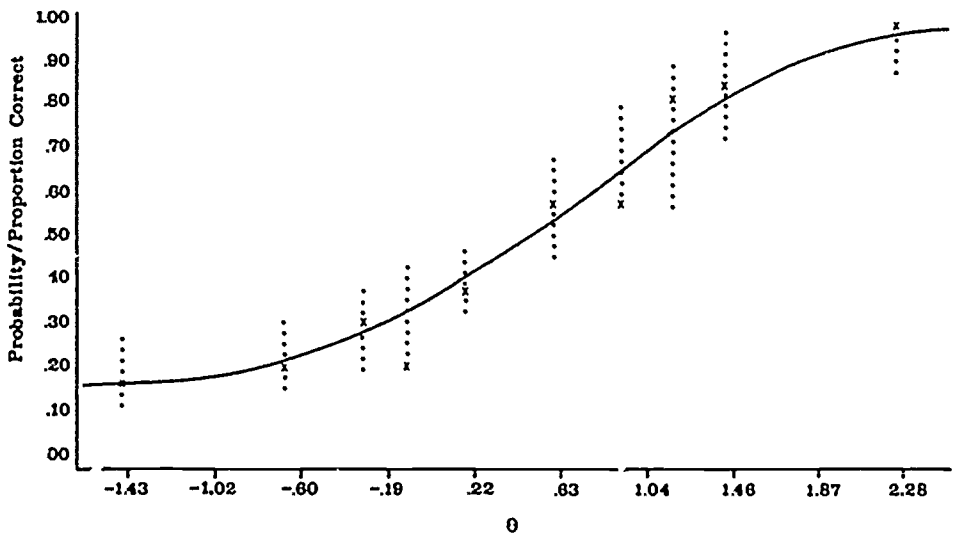


Figure 2
Observed and Expected 3-Parameter Logistic Response Curve for Item 16
(Smooth Line is Fitted Response Curve; "X" Represents Proportion Correct of a Group of Persons with Approximately Similar Abilities; Vertical Bars around Curve epresent Two Standard Errors around Expected Group Proportions Correct)



Discussion

With the use of marginal maximum likelihood estimation procedures and prior distributions on item parameters, it is now possible to estimate item response curves under the 1-, 2-, and 3-parameter logistic models from even very sparse data sets. It will be noted that the emphasis here is on the estimation of response curves rather than on item parameters. Simulation studies suggest that the recovery of generating item parameters is problematic, even with large numbers of items and persons, when the parameters of an item are not well identified by the calibration sample. These circumstances seem to obtain quite frequently with the 3-parameter model and, occasionally, with the 2-parameter model when the calibration sample does not span a sufficiently broad range of ability. Item response curves are estimated that do, on the other hand, explain the data satisfactorily.

The explanation of these findings is that for typical educational tests, data are well explained by a region of values in the parameter space. For an easy item, for example, data at hand may be well explained by either a 2- or a 3-parameter ogive; curves of each type can be found that are virtually identical in the region of the ability scale where the calibration examinees are to be found. The use of weak prior distributions will function in this situation to keep the resulting parameter estimates "reasonable," or in line with the values that the substantive interpretations of the item parameters would suggest (e.g., item slopes ranging between, say, 0 and 4) and asymptotes ranging between, say, 0 and .25).

The practical implication of this result is that the substantive interpretation of item parameters in the 3-parameter model (and, to a lesser extent, the 2-parameter model as well) may not always be justified. Maximum likelihood estimates for a given item may differ substantially from another set of values that reproduce the calibration data nearly as well. Discussion of item characteristics could be couched in terms of the item information function instead, since all sets of item parameter estimates in the "solution space" will yield similar information functions in the region where the data lies. Characteristics such as the point of maximum information and the value of the information function at that point can be expected to be much more stable than the item parameter estimates themselves.

Fortunately, most applications of IRT depend on the shape and location of response curves rather than the parameter values, particularly when applications are foreseen for examinees who are typical of the calibration sample. The estimation of an individual's ability from a given response pattern would typically be similar if computed from any item parameter values that produce similar response curves in the neighborhood of his/her ability. Discrepancies would be more likely for persons with abilities that are extreme.

One application that demands special attention, however, is vertical equating, or the linking of tests across broad ranges of ability--often across several grades or age groups. One approach to the equating problem is to calibrate tests separately in the low and high ability groups, say, and then to attempt to find the linear transformation that produces the closest match of item parameter

estimates for those items that were administered to both groups. Now, a linking item will tend to be comparatively easy for the high ability group and comparatively difficult for the low ability group. This means that the range of ability for which its response curve is well estimated in either group does not cover the region where the groups overlap, i.e., here the two estimated curves are supposed to be made to match. Poor linking may result as an artifact of the multicollinearity of item parameter estimates. The information needed for a proper link is found in not just the item parameter estimates and their standard errors, but in the matrix of correlations among the estimates as well. (This problem may be avoided by calibrating all items together with responses from all groups simultaneously, an option available in both BILOG and LOGIST.)

REFERENCES

Andersen, E. B. A goodness of fit test for the Rasch model. Psychometrika, 1973, 38, 123-140.

Andersen, E. B., & Madsen, M. Estimating the parameters of a latent population distribution. Psychometrika, 1977, 42, 357-374.

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

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

Bock, R. D., & Mislevy, R. J. BILOG: Maximum likelihood item analysis and test scoring; binary logistic models. Chicago: International Educational Services, 1982.

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

Kolakowski, D., & Bock R. D. LOGOG: Maximum likelihood test scoring and item analysis; logistic model for multiple item responses. Chicago: International Educational Services, 1973.

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

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

Wright, B. D., & Mean, R. BICAL: Calibrating items and scales with the Rasch model. Chicago: University of Chicago, Department of Education, Statistical Laboratory, 1978.