A Consumer's Guide to LOGIST and BILOG

Robert J. Mislevy and Martha L. Stocking
Educational Testing Service

Since its release in 1976, Wingersky, Barton, and Lord's (1982) LOGIST has been the most widely used computer program for estimating the parameters of the three-parameter logistic item response model. An alternative program, Mislevy and Bock's (1983) BILOG, has recently become available. This paper compares the approaches taken by the two programs and offers some guidelines for choosing between the two programs for particular applications. *Index terms: Bayesian estimation, BILOG, IRT estimation procedures, LOGIST, marginal maximum likelihood, maximum likelihood, three-parameter logistic model estimation procedures.*

The theoretical advantages of item response theory (IRT) psychometric models over classical test theory are by now well known and appreciated in the educational and psychological measurement communities. To obtain these benefits over a broad range of practical applications requires access to flexible and economical computer programs to estimate IRT parameters for items, examinees, and populations of examinees. The most widely used computer program for estimating item and person parameters under the three-parameter logistic item response model has been LOGIST (Wingersky, 1983; Wingersky, Barton, & Lord, 1982), based on the joint maximum likelihood (JML) approach suggested by Birnbaum (1968). More recently, the marginal maximum likelihood (MML) solution proposed by Bock and Aitkin (1981) and the Bayes marginal modal solution described by Mislevy (1986) have been implemented in the BILOG computer program (Mislevy & Bock, 1983).

This paper compares the two programs' theoretical approaches and attendant practical consequences, outlines some problems that any estimation algorithm must address, describes the character of the solutions offered by LOGIST and BILOG, and offers examples to illustrate some important differences and similarities.

The Three-Parameter Logistic Item Response Model

At the heart of IRT is a mathematical expression for the probability, denoted by P or $P(\theta)$, that a particular examinee with ability (or trait or skill) denoted by θ will respond correctly to a particular test item. Under the three-parameter logistic (3PL) model for test items that are scored either correct or incorrect

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 13, No. 1, March 1989, pp. 57–75

0146-6216/89/010057-19$2.20

(Birnbaum, 1968), this expression takes the following form:

$$P \equiv P(\theta) = c + \frac{1-c}{1+\exp[-1.7a(\theta-b)]} \quad , \tag{1}$$

where a is the item discrimination, b is the item difficulty, and c is the guessing parameter. Note that a linear indeterminacy exists in the 3PL model: If $\theta^* = A\theta + B$, $b^* = Ab + B$, and $a^* = a/A$, then $P(\theta^*;a^*,b^*,c) = P(\theta;a,b,c)$. Constraints must therefore be imposed on a set of parameter estimates in order to set the origin and unit-size of the θ scale.

Two other item response models in common use, namely the two-parameter logistic (2PL) model (Lord, 1952) and the one-parameter logistic (1PL) model (Rasch, 1960/1980), can be written as special cases of the 3PL model. (See Andersen, 1973; Fischer, 1974; and Rasch, 1960/1980, 1968, for independent derivations of the 1PL model and discussions of its special properties.) Most of the same estimation problems arise under all three models. This paper focuses on the 3PL because the 1PL and 2PL models can be expressed as special cases of the 3PL model; therefore, any solution to the problems of the 3PL model applies to the simpler models as well (although some solutions for the 1PL model do not generalize to the 2PL or the 3PL models).

The Theory of Parameter Estimation

Capitalizing on the advantages of IRT would be a simple matter if true item and true person parameters were known. A practical strategy is estimating item and person parameters and using the estimates as if they were true values. LOGIST and BILOG face identical statistical estimation problems but solve them in different ways. Insights into these estimation problems are important in understanding the fundamental philosophical differences between the two procedures.

In theory, the likelihood function for the model parameters contains all the information that the observed data convey about the values of these model parameters. This function gives the probability of the observed data for any permissible combination of parameter values. A common statistical procedure is to take as parameter estimates values of the model parameters that maximize the probability of the observed data. Parameter estimates obtained in this fashion are referred to as *maximum likelihood estimates* (MLEs). To find MLE parameter estimates for complex likelihood functions for which explicit solutions are unavailable, numerical methods are typically used to search the parameter space for locations where the first partial derivatives of the likelihood function are 0 and where the matrix of second partial derivatives is negative definite. At such locations the likelihood function attains at least a local maximum.

However, the uniqueness of examinee/item interactions carries IRT outside the purview of standard asymptotic statistical theory, which deals with the behavior of estimates of a fixed set of parameters as the number of observations increases. Such asymptotic theory would be applicable, for example, for the estimation of item parameters, if examinees' true θs were known. The response of each additional examinee would provide additional information about a fixed number of item parameters, the values of which could be estimated as precisely as desired by simply gathering enough responses and finding the maximum of the likelihood function

$$\mathbf{L}(\mathbf{a},\mathbf{b},\mathbf{c}|\boldsymbol{\theta},\mathbf{U}) = \prod_{j=1}^{N}\prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} Q_i(\theta_j)^{1-u_{ij}} \quad , \tag{2}$$

where $i = 1, \ldots, n$ is the item index;
$j = 1, \ldots, N$ is the examinee index;
$P_i(\theta_j)$ is the probability of a correct response to item i by examinee j, obtained from Equation 1;

$Q_i(\theta_j)$ is $1 - P_i(\theta_j)$;

u_{ij} is the observed response to item i by person j, coded 0 if the response is incorrect and 1 if correct;

$\boldsymbol{\theta}$ is the vector of known examinee abilities, one for each of N examinees;

$\mathbf{U}$ is the matrix of observed item responses of all examinees to all items; and

$\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are vectors of item parameters, one (a, b, and c) triple for each of the n items.

However, true θs are not known. Each additional examinee introduces an additional parameter into the likelihood function shown in Equation 2; therefore, standard asymptotic results for MLE estimation need not hold (Neyman & Scott, 1948). LOGIST and BILOG handle this problem differently; LOGIST uses a JML approach, and BILOG uses a MML approach.

The Joint Maximum Likelihood Approach

The JML approach to estimating parameters in the 3PL model originates with Birnbaum (1968). It is described in detail in Lord (1980), and in Wood, Wingersky, and Lord (1976). Using JML, LOGIST finds the values of item and examinee parameters that simultaneously maximize the joint likelihood function

$$L_J(\boldsymbol{\theta};\mathbf{a},\mathbf{b},\mathbf{c}|\mathbf{U}) = \prod_{j=1}^{N}\prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} Q_i(\theta_j)^{1-u_{ij}} \quad , \tag{3}$$

where the quantities have the same meanings as in Equation 2, except for θ which is now a vector of unknown abilities to be estimated along with the item parameters.

Although this straightforward approach is logically appealing, a price is paid. Except under some simplified circumstances, it is difficult—if not impossible—to prove that parameter estimates obtained using JML are statistically consistent with increases in the number of examinees (N) and/or the number of items (n). Andersen (1973), for example, showed that JML estimates of item parameters in the Rasch model are not consistent for increasing N if n is held constant at 2. If both N and n are increased appropriately, however, consistency can be proven for the Rasch model (Haberman, 1977). Simulations conducted by Swaminathan and Gifford (1983) suggest that under the latter circumstances, consistency may hold for the 3PL model as well.

The Marginal Maximum Likelihood Approach

The application of MML estimation in IRT originates with Bock and Lieberman (1970). The improved computing algorithms employed in BILOG were developed by Bock and Aitkin (1981). BILOG focuses on removing examinee parameters from the estimation problem entirely and estimates only item parameters.

The probability of a correct response to an item for an examinee with ability θ is given in Equation 1. The marginal probability, or the probability of a correct response for an examinee who has been randomly selected from a population with distribution of ability $G(\theta)$, is $\int P(\theta)dG(\theta)$. If a sample of N examinees is selected, the corresponding marginal likelihood function for the observed data is

$$L_M(\mathbf{a},\mathbf{b},\mathbf{c}|\mathbf{U}) = \prod_{j=1}^{N}\int\prod_{i=1}^{n} P_i(\theta)^{u_{ij}} Q_i(\theta)^{1-u_{ij}}\, dG(\theta) \quad . \tag{4}$$

The item parameters are estimated by finding the maximum of L_M; if desired, parameters describing the distribution G of θ may also be estimated. As the number of examinees increases, the number of parameters does not. Standard statistical theory can thus be brought to bear on estimation problems in this marginal framework. Even for short tests, this approach yields consistent estimates of item parameters—conditional, of course, on the veracity of the IRT model.

Once item parameter estimates have been obtained using the MML approach, θs may be estimated. The estimated item parameters are treated as if they were equal to their true values, and BILOG then produces MLEs of θ using Equation 2, or Bayes estimates of θ with an assumed or estimated population distribution (Bock & Aitkin, 1981).

Although MML may be more appealing than JML because of its formal statistical properties, a price is paid here too. In particular, MML assumes a structure for the distribution of θ in the population of examinees. If either the IRT model of the probability of a correct response given θ or the assumed model for the distribution of θ in the population are incorrect, the attractive statistical properties may fail to hold (Mislevy & Sheehan, in press).

Item Parameter Estimation Using Response Data Alone

The straightforward application of either Birnbaum's JML approach to parameter estimation or Bock and Lieberman's or Bock and Aitkin's approach does not necessarily yield finite and reasonable item and person parameter estimates. LOGIST and BILOG depart from the original JML and MML approaches to provide finite and reasonable parameter estimates. To better understand the nature and various features of these departures, it is first necessary to explore the straightforward application of either JML or MML.

The LOGIST Approach

If requested to estimate item parameters from item response data alone, LOGIST estimates parameters for each item and for each examinee that maximize Equation 3. At this location, $3n$ equations for the partial derivatives of Equation 3 with respect to the item parameters, and N equations for the partial derivatives of Equation 3 with respect to the θs, are 0. In addition, the second derivative for each θ is negative, and the 3×3 matrices for the parameters of each item are negative definite. The equations for the first and second partial derivatives for the 3PL model are presented in Lord (1980, chap. 12).

In principle, the solution could be found by Newton-Raphson iterations involving all item and examinee parameters at once. However, considerations of cost and accuracy usually render it impractical to invert the required matrix of second derivatives. By default, LOGIST instead arranges the estimation procedure into a series of four subproblems or steps of the form summarized in Table 1. (The additional item parameter COMC appearing in this table is a MLE of a common c parameter for all items that contain little information about their lower asymptotes; more about this below.) This arrangement improves the stability and computational efficiency of the procedure by ensuring that the subproblem solved in each step is reasonably well determined. Details of the procedure can be found in the *LOGIST User's Guide* (Wingersky et al., 1982). LOGIST resolves the linear indeterminacy of the 3PL model by standardizing the estimated θs between the range of -3 to $+3$, so that the θs within that range have a mean of 0 and a standard deviation of 1.

Table 1
Parameter Status in LOGIST Estimation Steps

	Parameter				
Step	θ	a	b	c	COMC
1	estimated	fixed	estimated	fixed	not used
2	fixed	estimated	estimated	estimated	estimated
3	estimated	fixed	estimated	fixed	fixed
4	fixed	estimated	estimated	estimated	fixed

Maximizing values of Equation 3 are the JML estimates, three for each item and one for each examinee. LOGIST estimates are approximations to JML estimates because the four-step procedure does not give complete convergence to JML estimates, and subsequent repetitions rarely provide sufficient improvement to justify the cost.

Neither JML estimates nor the LOGIST approximations have been proven to be consistent, but some simulation studies (Swaminathan & Gifford, 1983) suggest that the JML estimates for the 3PL model behave better as both test length and sample size increase. Better behavior for increased test length is not surprising when the nature of the LOGIST estimation cycles is considered. In the steps in which the item parameters are estimated, examinee parameters are treated as known, whereas they are actually only estimates. The fewer responses used to estimate an examinee θ, the more likely it is to depart from the true value. This discrepancy is likely to be worse for very high- or very low-scoring examinees, yet all estimates are treated equally. Theory and common sense thus agree that JML is less satisfactory when examinees respond to few items. The authors of LOGIST advise its use be restricted to data with at least 20 items per person and at least 800 to 1,000 examinees responding to each item.

Given that JML estimates do not meet the conditions necessary for standard maximum likelihood results, rigorous theoretical bases are not currently available for either tests of model fit or large-sample standard errors. Nevertheless, it is possible to compute, under standard MLE procedures, the matrix of second derivatives from which the variation of estimates around their true values is forecast. In this situation, it becomes an empirical question as to whether these forecasts of variation of estimates around their true values are practically useful. Wingersky and Lord's (1984) simulation study found that empirical standard errors were in good accord with those predicted by standard maximum likelihood results.

The BILOG Approach

If requested to estimate item parameters from response data alone, BILOG finds values of the item parameters that maximize Equation 4. In principle, this maximum can be found by a series of Newton-Raphson steps that involve the vector of first derivatives and the matrix of second derivatives for all item parameters. This straightforward solution was first presented in Bock and Lieberman (1970). It becomes impractical, however, for more than about 20 items.

Bock and Aitkin's (1981) re-expression of the required first derivatives led to a more practical computing algorithm. In the Bock-Aitkin development, the population θ density G in Equation 4 is approximated by a step function with jumps at a finite number of points. Adopting the vocabulary of numerical integration methods, these points are referred to as *quadrature points*.

Estimation proceeds under the simplifying assumption that the only values examinee θs can take are those represented by the quadrature points. Although the value associated with a particular examinee is not known, the probabilities of the possible values can be calculated using Bayes theorem from the examinee's response vector, the item parameters, and G. This set of probabilities is called a posterior distribution of an examinee's θ. Having done this, the expected value of the log of Equation 2 can be calculated and maximized with respect to the item parameters.

To obtain the expected value of the log of Equation 2, however, requires knowing the item parameters and G. Of course, the item parameters and G are unknown. In iterative cycles, however, it is possible to recompute the desired expected value and G with updated estimates of the values of item parameter estimates that maximize the preceding expectation. These are exactly the steps of the EM algorithm (Dempster, Laird, & Rubin, 1977), in the special case of "missing multinomial indicators" because θs are limited to a finite number of values.

The shape of the population distribution G may be either (1) assumed normal, (2) fixed at values

specified by the user, or (3) estimated concurrently with the item parameters (an empirical prior). The linear indeterminacy of IRT models is resolved by constraints on the estimated densities at the quadrature points which effectively standardize G.

MML estimation of item parameters meets the conditions necessary for standard maximum likelihood results to apply. Thus tests of model fit and large-sample standard errors are available from BILOG. However, these depend on the assumption that the population distribution G is correctly specified and either known or consistently estimated with only a finite number of parameters. This assumption is usually better satisfied if G is estimated simultaneously with the item parameters. Initial evidence suggests that using the normal prior distribution, which leads to more rapid convergence, introduces little bias into item parameter estimates or large-sample standard errors (Bartholomew, 1988; Bock & Aitkin, 1981), but more study of this issue is required.

Item Parameter Estimation Using Information External to the Response Data

Under the 3PL model, the item parameter values that maximize the JML or MML likelihood function need not be either finite nor reasonable. If finite and reasonable estimates are required, then this requirement must be included in the estimation routine. Resulting estimates will depend not only on the data and the model, but at least partly on the method and the strength of prior beliefs about how item parameter estimates "ought to" look.

Infinite Item Parameter Estimates

As early as 1931, Heywood pointed out that some correlation matrices, in accordance with a linear factor analysis model, lead to 0 or negative values for unique variances. Occasional "Heywood cases" are a familiar—if unwelcome—feature of maximum likelihood factor analysis, both of measured variables and of dichotomous variables in the IRT extension of the Thurstonian paradigm (Bock, Gibbons, & Muraki, 1988). After 50 years of experience with factor analysis, it is not surprising that the maximizing values for item discriminations under the 2PL or 3PL model are sometimes infinite.

It has been speculated that without constraints on their values, at least one a will become infinite in the attempt to fit the 2PL or 3PL model to any set of response data (Wright, 1977). The authors would be happy to supply interested readers with a dataset which, when fit by the 2PL model with JML, does not produce infinite a estimates. The fact that some datasets do not yield infinite estimates offers little comfort, however, if others do. Infinite parameter estimates are neither plausible nor useful. Additional information or structure is required to obtain estimates that may be less likely (i.e., do not maximize the likelihood function), but more satisfactory.

Multicollinearity

Even when constrained item parameter estimates under the 3PL model are finite, they need not be reasonable. It is easy to see how this can occur. Although an item response function traces the probability of a correct response across the entire range of θ, data are available in only a limited region: the neighborhood in which the θs of the sample of examinees lie. Even if the true θs were known, only an approximation of the response curve would be available, and only in this neighborhood. The data have nothing to say about probabilities of correct response elsewhere. JML and MML procedures find the item parameter estimates that best describe proportions of correct response in this neighborhood, and can make statements about probabilities outside the neighborhood only because the resulting curve is required to

be 3PL. When the neighborhood is small, or when the item is relatively easy or difficult for the sample of examinees, a variety of apparently discrepant (a, b, and c) triples can capture the data nearly equally well but disagree about what happens where no data exist.

This phenomenon is reflected numerically by a poorly conditioned matrix of second derivatives, which must be inverted in the Newton-Raphson steps taken by both LOGIST and BILOG. This matrix describes the surface of the likelihood function being maximized with respect to the three parameters of a given item. Near singularity implies that this surface changes very gradually and therefore a local maximum is difficult to find. In extreme cases, the surface does not change at all, in which case no local maximum exists and the solution fails entirely.

Methods of Incorporating External Information

A Bayesian solution to item parameter estimation incorporates external information by imposing prior distributions on item parameter estimates. A prior distribution itself can have "higher-level" parameters, either specified a priori or estimated from the data at hand.

The posterior probability distribution of the item parameters is the product of the likelihood function (either JML or MML) and the prior distribution for the item parameters. Bayesian modal estimates of item parameters are the values that maximize the posterior probability. Bayesian modal estimates have been developed for JML by Swaminathan and Gifford (1986) and for MML by Mislevy (1986). The large-sample properties of modal estimates are equivalent to the large-sample properties of the likelihood functions (JML or MML) used to obtain them. Thus large-sample indices of fit and standard errors formally hold for the Bayesian extension of MML but not of JML (see Lewis, 1980).

Unless previous analyses provide concrete information about the values of item parameter estimates, it is reasonable to enforce fairly unobtrusive prior distributions. Parameters estimable from the observed data alone would then receive Bayes estimates similar to their MLEs. Infinite and extreme estimates would become finite and reasonable. Similar effects can also be achieved informally through constraints on a maximum likelihood procedure. The practical problem under both formal and informal Bayesian solutions is specifying prior distributions or procedures that produce the desired outcome, that is, an appropriate balance between external information and information from the observed response data itself.

The LOGIST Approach

LOGIST approaches the problem informally, partly by using simple constraints to handle extreme item parameter estimates. Upper and lower limits are specified for the values of estimates of a and c, a procedure equivalent to specifying uniform prior distributions on the allowable intervals. If neither a nor c for an item exceeds a boundary in a given cycle with provisionally fixed values of θ estimates, then none of the item's parameter estimates will be affected by this prior specification. If one or more estimates do exceed the boundary, they are assigned the boundary values and the remaining estimates for that item are values that maximize the likelihood function with these values fixed at the boundaries. Boundary values affect the next cycle's θ estimates, so that the parameter estimates for all other items and all examinees are affected, though probably minimally, whenever even a single parameter estimate for any item takes on a boundary value.

Although LOGIST provides default boundary value settings, the manual shows how to estimate more appropriate boundary values for a given set of data by using a partial run of the program. For item discriminations, for example, the user is advised to examine a frequency distribution of estimates from the partial run. If many more estimates equal to the provisional upper limit exist than estimates slightly

less than that limit, the manual suggests raising the upper limit before continuing the run to completion. If many estimates are equal to the limit and this value is substantially above the next lowest estimate, the manual suggests lowering the limit before continuing.

Such simple procedures informally incorporate the user's beliefs about reasonable values for item parameter estimates and reasonable distributions of these values. These procedures, in which individual values are estimated using a population distribution estimated simultaneously from the same data, have been called hierarchical Bayesian models when a formal Bayesian framework is used to estimate the population distribution (Lindley & Smith, 1972) and empirical Bayes models when it is not. Note that when these ideas are employed to obtain reasonable estimates, expected estimates for a given item can depend on the other items in the test.

Another LOGIST constraint on estimates of c can also be thought of in an empirical Bayes framework: the MLE of a single common value (COMC) for the c parameters of all items whose provisional estimate of the quantity $b - 2/a$ falls below a specified criterion. In this way, limited information for individual cs is pooled to provide a single, better-determined, common estimate. (The index $b - 2/a$ is heuristically justified by the observation that less information is available in the response data for estimating c for an item that is easy and not very discriminating. The default criterion value of the index is -2.5.) Because the c values in question are poorly determined by the response data, restricting them to a common value estimated from the data decreases the likelihood only modestly. If poorly determined cs were not so restricted, severe multicollinearity would result and the poorly determined cs would have a large and undue influence on the estimates of the a and b parameters of the items involved. By reducing multicollinearities among a, b, and a poorly determined individual c in this manner, it is likely that better (lower mean squared error) estimates of a and b will be obtained.

A final LOGIST procedure with Bayes-like effects is the imposition of the structure of estimation steps described earlier. With each step, estimates generally depart further from their starting values in the direction of the JML solution. Terminating early gives an informally weighted average of starting values and JML estimates. Because within-cycle constraints tend to restrain step sizes in cases of near-collinearity or extreme values, limiting the number of steps tends to weight the JML estimates less heavily for items with less information than items with more information. The failure to attain complete convergence to JML estimates, then, may in fact prove advantageous, informally shrinking poorly determined estimates toward their apparently reasonable starting values.

The BILOG Approach

BILOG incorporates information external to the observed response data by using a formal Bayesian framework. By default, BILOG implements prior distributions on all item parameters under the 3PL model. The normal distribution is used for the bs, the log-normal for the as, and the beta for the cs. Specification of prior distributions may be omitted for some or all types of parameters if desired, and the parameters of the prior distributions may either be specified by the user or partially estimated from the data. This latter approach, termed "floating priors", is the BILOG default. The effect of using floating priors is that all parameters of a given type shrink toward the mean of that type with a predetermined strength, while that mean is estimated from the data (see Mislevy, 1986, for details).

In this formal approach to incorporating external information, the estimation equation for an individual item parameter is the sum of two terms. The first term is the contribution from the likelihood, which increases with sample size. The second term is the contribution from the prior, which remains constant with respect to sample size. Shrinkage toward the (possibly estimated) mean of parameters therefore decreases as sample size increases.

The cost of obtaining more reasonable estimates by incorporating formal prior distributions is twofold.

First, as with the more informal LOGIST constraints, the expected estimates for a given item depend on the characteristics of other items in the test. Second, the prior information may not be appropriate for the data, biasing estimates of poorly determined item parameters. Such biases are reduced when higher-level parameters are estimated from the data, arguing for "floating" rather than "fixed" prior distributions in item parameters, unless strong prior information truly exists.

θ Estimation

Most applications of IRT aim to make statements about the abilities of individual examinees for the purpose of classification, selection, or placement. Both LOGIST and BILOG offer ways to estimate θs, either in the same run as item parameters are estimated or by using previously estimated item parameters. In either case, the item parameters are treated as known. (See Lewis, 1985, and Tsutakawa & Soltys, 1988, on incorporating the uncertainty associated with item parameter estimates.)

Clearly, MLEs of θ are integral to LOGIST's JML item parameter estimation. Point estimates of θs and item parameters are jointly obtained that (approximately) maximize the fit of the specified model to the data, as gauged by the joint likelihood function. Point estimates of θ do not arise during the course of BILOG's item parameter estimation; they are calculated, if requested, in a separate program phase, after any item calibration that may be performed. MLEs of θ are one BILOG option; Bayes mean estimates, more consistent with the MML approach to item parameter estimation, are another.

Maximum Likelihood Estimates

Both LOGIST and BILOG can produce MLEs of θ. For a given set of item parameters and response patterns, LOGIST and BILOG estimates will differ negligibly from each other insofar as the details of the two numerical procedures are different. Lord (1980, p. 54, Equation 4.20) provides the likelihood equation that both programs solve to estimate θ.

Both programs use Newton-Raphson iterations from a starting value based on a standardized percent correct that is adjusted for guessing. If a provisional value of estimated θ is far from the maximizing value, Newton-Raphson steps can diverge. Both programs reduce this possibility by limiting step size and forcing steps in the direction that increases the likelihood. If the number of items to which an examinee has responded is large, the MLE estimated θ for an examinee is approximately normally distributed, with mean equal to the true θ and large sample variance given by Lord (1980, p. 71, Equation 5.5).

A unique finite maximum exists under the 3PL model for most response patterns above chance level. Although multiple maxima are occasionally encountered, this occurs more often with short tests and is often associated with response patterns in poor accord with the model. A more extensive and time-consuming grid search would be required to find the global maximum in such cases; neither LOGIST nor BILOG currently does this.

For response patterns yielding infinite MLEs, BILOG provides floor and ceiling values. In the typical LOGIST run, where item and θ parameters are estimated simultaneously, more constraints are imposed because θ estimates will be used in the next cycle of item parameter estimation. By default, examinees with 0 and perfect scores and examinees who answered fewer than one-third of the items presented to them are excluded from the estimation of item parameters. Examinees who do not fall into these categories but whose estimated θ tends to become infinite are given default boundary values.

Bayes Estimates

BILOG can also produce Bayes estimates of θ. As a by-product of BILOG item parameter estimation,

expected values of the density of the examinee population are obtained at each of the quadrature points. The posterior probability that an examinee θ equals a particular quadrature point can then be obtained from this information. Then a summary can be developed explaining what is known about an examinee in terms of a Bayesian mean estimate (i.e., the mean of this estimated posterior distribution, and its associated standard deviation). These Bayesian mean estimates are sometimes called "expectation a posteriori" (EAP) estimates.

Bock and Mislevy (1982) described properties of EAP estimates. By using a population distribution in the course of θ estimation, finite values are obtained for all response patterns, including those that yield infinite MLEs. The reasonableness of EAP estimates obtained for these response patterns depends on the reasonableness of the population distribution that is employed. An empirical estimate of the examinee distribution accumulated in the course of item parameter estimation would be quite appropriate for this purpose if the calibration sample represented a population of interest, but less so if it did not.

A Comment on Estimating θ Distributions

Consider the problem of estimating, in a population of interest, the distribution G of θ from the item responses of a sample of examinees. Paradoxically, the distribution of θ estimates, each of which is in some sense optimal for the particular examinee, is not necessarily a good estimate of G. MLEs tend to have too large a variance; Bayesian estimates have too small a variance. Increasing test length decreases the discrepancies, but for any test of fixed length, the distribution of point estimates of θ from either LOGIST or BILOG will not converge to the true distribution of θ as the number of examinees increases without bound.

Methods of estimating G directly are described by Andersen and Madsen (1977), Mislevy (1984), and Sanathanan and Blumenthal (1978). Mislevy's histogram solution for G is approximated in BILOG, although the solution is run to effective convergence of item parameters, not of G. Sampling a value at random from the posterior distribution of each examinee could provide point estimates of θ for each examinee that yield a consistent estimate of G. This would extract a crude monte carlo approximation of the integral equations employed in the direct solutions of G mentioned above. Hence the paradox: A consistent estimate of G from point estimates for each θ would require these "noisy" estimates that are decidedly non-optimal for each examinee considered individually.

Additional Considerations

Handling Missing Responses

For convenience of presentation, the preceding discussions have assumed that all examinees responded to every item under consideration. This situation is frequently not realized in practice, sometimes for reasons intended by the researcher and sometimes not. LOGIST and BILOG handle these situations in the same ways by incorporating methods for three types of nonresponse. (See Mislevy & Wu, 1988, for a rigorous treatment of missing responses in IRT.)

Most easily dealt with are potential examinee/item combinations that are missing by design. Different examinees may take different forms of overlapping tests, for example, so that they have no opportunity to provide responses to items not presented to them. It is intuitively clear that these nonresponses can be ignored for the purpose of maximum likelihood and Bayesian estimation of item and examinee parameters.

It is less obvious—but true nonetheless—that patterns of nonresponse that might be related to θ or item parameters can also be ignored. If these patterns of nonresponse are determined wholly by previous

observable responses, as in adaptive testing, then they may be ignored in Bayesian and direct likelihood inference. Both LOGIST and BILOG allow the user to encode a "not presented" indicator for a given examinee on a given item. All calculations are then carried out with respect to only those item/examinee combinations included in the sample. This feature is convenient for linking tests through common items.

Less clear is how to handle responses to items an examinee was presented, but did not reach due to time limitations. A fully satisfactory treatment of this phenomenon would require an extended model with an ability parameter and a speed parameter for each examinee. All models allowed by both programs assume nonspeeded testing conditions, so arbitrary decisions must be made about how to handle these observations. The options available to the user are to code such item/examinee combinations as "not presented," so that they will be treated as if they were missing by design; as "incorrect" because they have not been answered correctly; or as "partially correct" (see below). The first option is most usual; some empirical evidence suggesting its reasonableness has been provided by van den Wollenberg (1979).

Finally—and most troublesome—are the items an examinee obviously encountered and decided to omit. Coding these observations as incorrect is palatable for free-response items, but less so for multiple-choice items. Had the examinee guessed at random, a positive probability of a correct response would have resulted. Lord (1983) has suggested for such data a model with two examinee parameters, one for ability and one for a propensity to omit rather than to guess at random when confronting an item for which they have no preference among response alternatives. The best strategy for using LOGIST and BILOG is to treat such observations as partially correct, with a weight of the reciprocal of the number of alternatives to the item. This leads in expectation to the same results as replacing each omit by a randomly assigned response (Lord, 1974).

Scaling Issues

By default, both LOGIST and BILOG resolve the indeterminacy in the 3PL's θ scale by standardizing estimates with respect to the calibration sample of examinees—LOGIST using estimated θs between -3 and $+3$, BILOG using the estimated θ distribution. If a single test is calibrated twice by either program using two different samples of examinees, the resulting scales will differ. Thus, the two BILOG scales will differ and the two LOGIST scales will differ as a function of differences in the means and dispersions of θ in the two samples, as well as in the sampling variation generally associated with any estimation procedures. A linear transformation, such as Stocking and Lord's (1983), puts the two sets of LOGIST estimates or BILOG estimates on approximately the same scale. Remaining differences between sets of estimates from a given program can be attributed to estimation errors of various types.

A second scaling issue of practical importance arises from a subtle but fundamental difference between the JML procedure used by LOGIST and the MML procedure used by BILOG. BILOG estimates the parameters of the distribution of θ from which the sample of examinees was drawn; increasing the number of examinees increases the accuracy of the estimates of this population distribution. LOGIST estimates θ for each examinee; increasing the number of examinees increases the number of estimated θs, thereby increasing the accuracy of the distribution of estimated θ. But the relationship between an estimated distribution of θ and an estimated distribution of estimated θ is nonlinear, depending on test length and item parameters. Even after applying the transformation described above, nonlinearity remains between the scales from BILOG and LOGIST runs, or between LOGIST runs with appreciably discrepant tests or examinee samples. Assuming that the items are appropriate for the examinee sample, this nonlinearity becomes negligible as (1) the test is lengthened so that estimated θs are indistinguishable from true θs, and (2) examinee sample sizes are increased so that the distribution of the estimated θs can be accurately obtained.

Diagnostic Information

Both LOGIST and BILOG provide information on the progress of the numerical procedures invoked. This information is vital to the monitoring of the successful completion of the program. Both programs also provide values of the criterion functions that are maximized. Such information is useful in comparing the appropriateness of alternative models, such as the 3PL versus the 1PL model.

Strictly speaking, however, the IRT models that LOGIST and BILOG use will never fit data exactly. More aid to the user about the nature of lack of model fit would be welcome in both programs. This area deserves greater emphasis in IRT more generally.

Every calibration of item and person parameters should be examined using plots of observed versus predicted item/ability regressions, as described in Kingston and Dorans (1985). No other check on model fit provides such satisfactory guidance in the detection of (possibly) correctable fit problems. Through this mechanism unsatisfactory limits on values of parameter estimates for LOGIST or unsatisfactory priors placed on some items for BILOG can be detected. These conditions are potentially correctable by rerunning either program with new settings. Such plots can also be useful in identifying items for which the observed proportions correct are nonmonotonic or have an upper asymptote other than 1. These problems are not correctable because these items cannot be well fit by the logistic item response model. The user may wish to eliminate these items from a second run of the data. BILOG provides line printer plots of this type.

Ease of Use (or Lack Thereof)

Neither LOGIST nor BILOG is particularly easy to learn to use. To obtain consistently satisfactory results with either program, the user must possess a fairly high degree of knowledge about what the program is trying to do and how it goes about trying to do it—a level at least equal to that of the present article. Both programs offer default settings for the novice, but knowledgeable application of the model requires informed troubleshooting skills and, as often as not, a second or even a third run to improve the solution.

A Numerical Example

It is not possible within the scope of this paper to compare the behavior of LOGIST and BILOG with a wide variety of item and examinee parameter combinations, nor to hunt out possibly subtle effects on applications such as equating and adaptive testing. Therefore, an illustrative application to two simple simulated datasets is provided, and costs and recovery of generating parameters are examined. The results pertain to the mainframe program versions publicly available at the time of this writing, LOGIST 5 and BILOG 2.2.

The Data

Responses from simulated examinees to an artificial test containing 45 items comprised of three replications of 15 four-choice items were generated. Values of item parameters and θs for 1,500 examinees were obtained by applying LOGIST to a typical form of the Test of English as a Foreign Language (TOEFL) and using the LOGIST estimates as generating "true" parameters for the simulation. Item response data were then generated by computing the model probability of a correct response to each item/examinee combination and then assigning it a correct response if a random number selected from the unit interval did not exceed this probability. Two simulated tests were analyzed: a 15-item test consisting of one

replication of the generating item parameter set and a 45-item test consisting of all three replications.

Both LOGIST runs had the following specifications:

1. The maximum for the a parameters was set to 1.5.
2. θs were restricted to the range $-7, +4$.
3. Individual cs were estimated only for items with $b - 2/a > -3$.
4. The default four-step estimation procedure was used. Item and examinee parameter estimates were produced automatically. To compare the resulting estimates with the generating values, the results of both LOGIST runs were transformed to the scale of the generating values using the Stocking and Lord (1983) procedure to optimize the congruence of the true and estimated test characteristic curves.

Both BILOG runs had the following specifications:

1. A standardized θ distribution was estimated jointly with the item parameters using 10 quadrature points.
2. Default specifications of prior distributions were employed for item parameters of each type, as were default values for the number of cycles and the convergence criterion.
3. Two different data storage methods were used in each problem: a faster algorithm applicable only to data for which all examinees take all items, and a slower algorithm applicable when omits and/or not-presented items can occur.
4. Bayesian EAP θ estimates were produced for each examinee.

Results

The residuals (estimated minus true) of the LOGIST and BILOG item parameter estimates for the 45-item test are plotted against the true values in Figure 1; Figure 2 shows the item parameter residuals for the 15-item test. Both procedures appear to recover the true parameters equally well for the 45-item test. Residuals for examinee parameter estimates for this test are shown in the top row of Figure 3. As might be expected, BILOG's Bayesian estimates (Figure 3a) shrink modestly toward the population mean, while LOGIST's MLEs (Figure 3b) are slightly more dispersed than the true values, with a few outliers for near-chance-level patterns. For 10 low-θ examinees (true θ between -2.81 and $-.64$) LOGIST produced extreme θ estimates with residuals falling below the bottom of the range plotted.

For the 15-item test (Figure 2), three easy items (true difficulty of -1.83, -1.38, and $-.35$) were substantially misestimated by LOGIST (residuals of -2.60, $-.81$, and $+.86$ respectively) and are not plotted. BILOG appears to recover the true values better in this test. θ estimates for the 15-item test are shown in the bottom row of Figure 3. The shrinkage of Bayesian estimates and the dispersion of MLEs noted for the 45-item test have been accentuated. Over 100 low-θ examinees were sufficiently misestimated by LOGIST for their residuals not to be plotted. The authors of LOGIST do not recommend using it for tests as short as the 15-item test used here. These results serve to confirm the prudence of the authors' guidelines.

Execution times of the two programs are shown in Table 2. Obviously CPU seconds are machine dependent, but relative values should be more broadly meaningful. Times are comparable under the case of no omits and no not-presented items, but for the more general model the default settings for LOGIST exhibit an advantage over BILOG's default settings. The advantage is about 2:1 for the short test and 1.5:1 for the long test.

Comments on the Example

BILOG appeared to recover generating item parameters better than LOGIST for the 15-item test, due

Figure 1
Residuals (Estimated Minus True) for Item Parameter Estimates From BILOG and LOGIST for the 45-Item Test

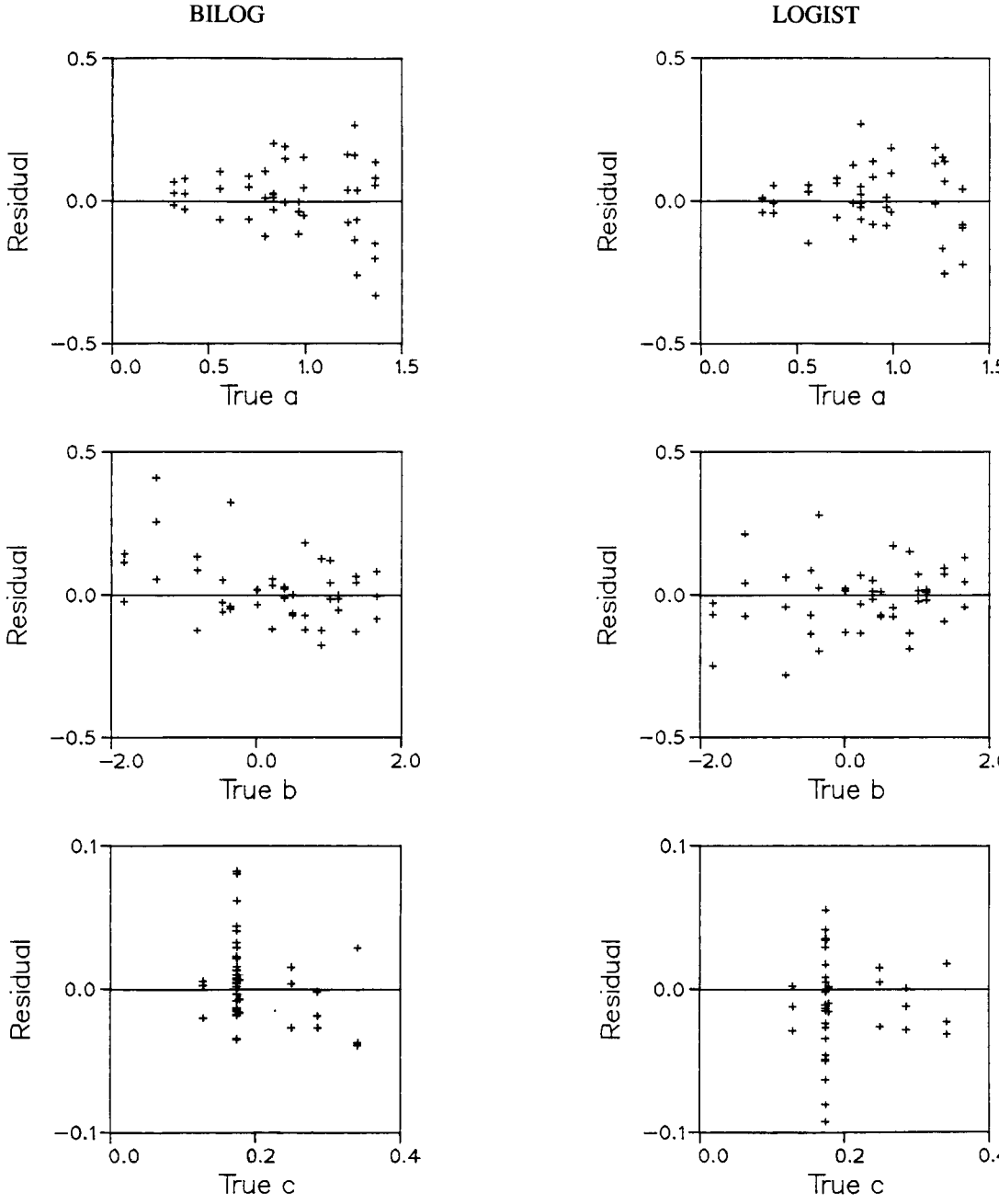


Figure 2
Residuals (Estimated Minus True) for Item Parameter Estimates
From BILOG and LOGIST for the 15-Item Test

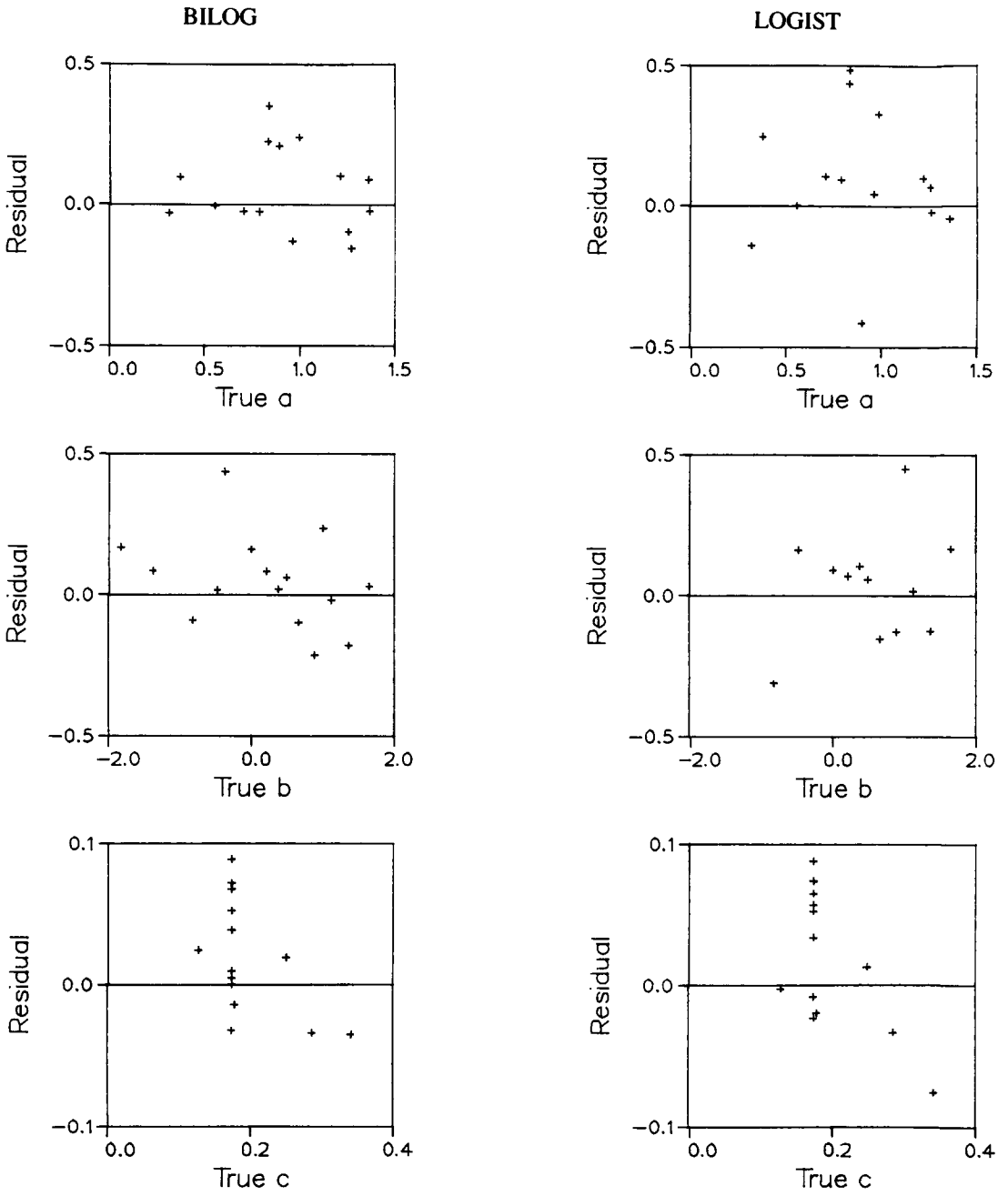
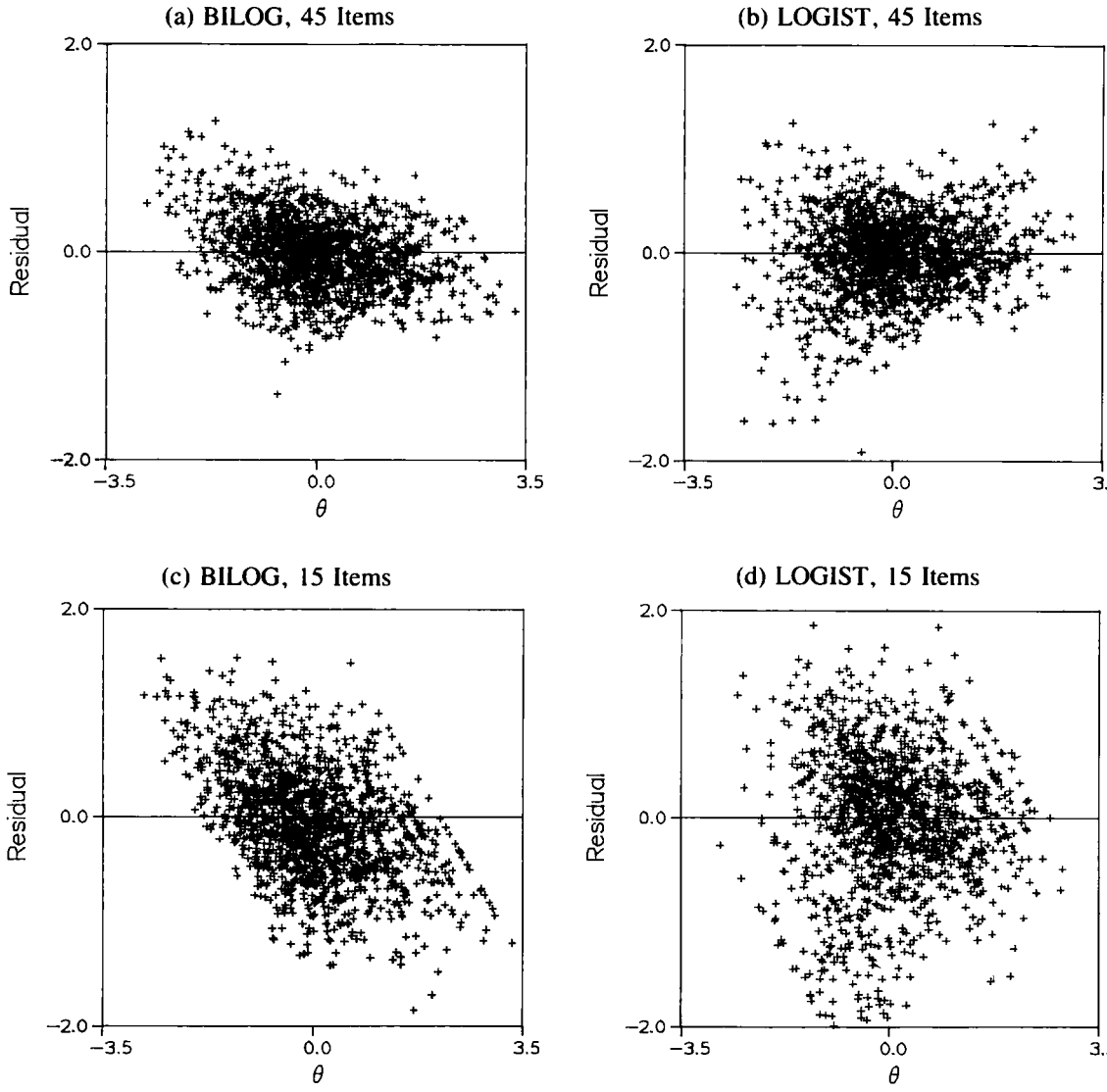


Figure 3
Residuals (Estimated Minus True) for EAP θ Estimates From BILOG and MLE θ Estimates From LOGIST



in large part to the shrinkage of *c* parameters toward their estimated mean. The *b* parameters for the 15-item test were recovered fairly well by both programs. For the 45-item test, the results from the two programs were very similar. Given these results, LOGIST might be preferred for the longer test because of its faster execution time, or BILOG because of its statistical properties.

However, the similarity of the results for LOGIST and BILOG on the 45-item test does not necessarily imply that the user should be indifferent as to choices in more demanding applications, such as long

Table 2
Execution Times of LOGIST and BILOG on Two Simulated Datasets

Test and Program	CPU Seconds
15-Item Test	
LOGIST	19.45
BILOG (assuming data contain no omitted or not-presented items)	20.14
BILOG (assuming data may contain omitted and not-presented items)	39.32
45-Item Test	
LOGIST	37.29
BILOG (assuming data contain no omitted or not-presented items)	34.26
BILOG (assuming data may contain omitted and not-presented items)	55.58

equating chains or several cycles of item pool refreshment in adaptive testing. Quite possibly, in these circumstances potential subtle differences between the two programs might have ramifications leading to an obvious choice. The first—and possibly most important—comment, then, is to reiterate a caveat: By no means does this example offer a comprehensive comparison of LOGIST and BILOG.

It is difficult to construct any single dataset from which a "fair" comparison of LOGIST and BILOG would result. An ideal comparison would employ data generated with parameters that are both realistic and known. But notions of what "realistic" means are determined by what available programs provide, and programs do not necessarily provide the true parameters for any dataset of reasonable size. Every program must make arbitrary choices about how to produce estimates of item parameters poorly supported by the data, and an artificial dataset generated from such results can spuriously favor one program over another by the configuration of poorly determined parameters.

In this connection, Thissen (personal communication, 1984) has pointed out that the above simulation may favor LOGIST because it uses previous LOGIST estimates as generating values. From one perspective, the generating values used in this example can be viewed as representing fewer than three parameters for each item. This occurs because some items have identical lower asymptotes arising from a common c value estimated for poorly determined cs in the original application of LOGIST to the TOEFL data. It would not be unreasonable to find that a procedure that permits such a reduced parameterization (LOGIST) is more efficient than a procedure that does not (BILOG). If, on the other hand, previous BILOG estimates had been used to generate data, the estimated cs might have a tendency toward a beta distribution, offering an equally fortuitous but spurious advantage to a subsequent BILOG run. Similar, although less obvious, influences may also come into play for values of the other parameters.

The only escape from this potential for circular reasoning is accumulating experience over a broad range of problems. One path that future research should follow has been led by Yen (1987), who compared the behavior of the two programs over a broader range of generating values. A second path would focus not on parameter estimates but on criteria relevant to specific applications. Examining recovery of the first test of a circle of linked tests in equating would be an example of such an experiment.

Conclusions

For applications for which LOGIST is recommended—with longer tests and larger samples, and when some items are omitted or not reached—the programs provide similar item parameter estimates; therefore,

LOGIST might be preferred on the basis of costs. With longer tests and larger samples in which all possible item/examinee interactions are observed, BILOG is competitive with LOGIST in terms of cost, and its formal statistical properties provide useful information about the large-sample properties of the resulting estimates, particularly if priors on the item parameters are weak.

Users with short tests and/or small examinee samples should consider using BILOG. In these situations, BILOG's more formal Bayesian procedures are likely to provide reasonable results. However, for small samples of examinees, particularly if not all possible item/examinee interactions are observed, the statistical indices based on large-sample MML theory may be less useful. Assuming that the examinee distribution and item parameter means are estimated from the data, the effect of the prior in small samples of examinees will be to produce item parameter estimates less like the 3PL model and more like a model with individual *b*s but common *a* and *c* estimates. If Bayesian θ estimates are requested for short tests, they will be shrunk noticeably toward the center of the estimated examinee distribution. In these situations, the reasonableness of the results depends on the reasonableness of the prior structure.

References

Andersen, E. B. (1973). *Conditional inference and models for measuring*. Copenhagen: Danish Institute for Mental Health.

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika, 42,* 357–374.

Bartholomew, D. J. (1988). The sensitivity of latent trait analysis to choice of prior distribution. *British Journal of Mathematical and Statistical Psychology, 41,* 101–107.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–280.

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

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

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

Fischer, G. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Haberman, S. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics, 5,* 815–841.

Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society, Series A, 134,* 486–501.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9,* 281–288.

Lewis, C. (1980). *Difficulties with Bayesian inference for random effects* (Research Bulletin 80-448-EX). Groningen, The Netherlands: Psychological Institute, University of Groningen.

Lewis, C. (1985). *Estimating individual abilities with imperfectly known item response functions*. Paper presented at the annual meeting of the Psychometric Society, Nashville TN.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B, 34,* 1–41.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, No. 7.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 29–51.

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

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

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software, Inc.

Mislevy, R. J., & Sheehan, K. M. (in press). The role

of collateral information about examinees in the estimation of item parameters. *Psychometrika*.

Mislevy, R. J., & Wu, P. K. (1988). *Inferring examinee ability when some item responses are missing* (ETS Research Report 88-48-ONR). Princeton NJ: Educational Testing Service.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika, 16*, 1–32.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. [Expanded edition, University of Chicago Press, 1980.]

Rasch, G. (1968). *A mathematical theory of objectivity and its consequences for model construction*. In Report from European Meeting on Statistics, Econometrics, and Management Sciences, Amsterdam.

Sanathanan, L., & Blumenthal, N. (1978). The logistic model and estimation of latent structure. *Journal of the American Statistical Association, 73*, 794–798.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 589–601.

Tsutakawa, R. K., & Soltys, M. J. (1988). Approximations for Bayesian ability estimation. *Journal of Educational Statistics, 13*, 117–130.

van den Wollenberg, A. L. (1979). *The Rasch model and time-limit tests: An application and some theoretical contributions*. Unpublished doctoral dissertation, University of Nijmegen.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver BC: Educational Research Institute of British Columbia.

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

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347–364.

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

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97–116.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, 52, 275–291.

Acknowledgments

The authors thank Maxine Kingston, Charles Lewis, Peter Pashley, Kathleen Sheehan, and Marilyn Wingersky for their comments on earlier drafts of this paper. An expanded version of this paper is available as ETS Research Report 87-43. This study was supported by Educational Testing Service through Program Research Planning Council funding. Authors' names appear in alphabetical order.

Author's Address

Send requests for reprints or further information to Martha L. Stocking, Educational Testing Service, Princeton NJ 08541, U.S.A.