ED 303 473 TM 012 461

AUTHOR	Mislevy, Robert J.; Stocking, Martha L.
TITLE	A Consumer's Guide to LOGIST and BILOG.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-87-43
PUB DATE	Nov 87
NOTE	76p.
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Bayesian Statistics; Comparative Analysis; *Computer Software; *Estimation (Mathematics); *Latent Trait Theory; *Maximum Likelihood Statistics; Research Methodology
IDENTIFIERS	*BILOG Computer Program; Item Parameters; *LOGIST Estimation Procedures; Three Parameter Model; User Guides

ABSTRACT

Since its release in 1976, LOGIST has been the most widely used computer program for estimating the parameters of the three-parameter logistic item response model developed by A. Birnbaum. An alternative program, BILOG, developed by R. J. Mislevy and R. D. Bock (1983), has recently become available. This paper compares the approaches taken by the two programs and offers some initial guidelines for choosing between the two programs for particular applications An application of the two programs to two simple simulated data sets is illustrated, wherein responses are assessed from simulated examinees to a 45-item artificial test comprised of three replications of 15 four-choice items. It is recommended that: (1) the user with short tests and/or small examinee samples consider using BILOG due to its incorporation of formal Bayesian procedures, while LOGIST performs better with longer tests and larger samples; (2) LOGIST can produce joint maximum likelihood estimates, while BILOG can produce marginal maximum likelihood estimates; and (3) neither program alone can produce finite and reasonable parameter estimates. A list of references, three data tables, and 19 graphs are provided. (TJH)

RR-87-43

RESEARCH REPORT

A Consumer s Guide to LOGIST and BILOG

Robert J. Mislevy
and
Martha L. Stocking

Educational Testing Service
Princeton, New Jersey
November 1987

A Consumer's Guide to LOGIST and BILOG[1]

Robert J. Mislevy[2]

and

Martha L. Stocking

Educational Testing Service

Princeton, New Jersey

November 1987

[1] This study was supported oy Educational Testing Service through Program Research Planning Council funding.

[2] Authors' names appear in alphabetical order.

Abstract

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 3-parameter logistic (Birnbaum) 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 initial guidelines for choosing between the two programs for particular applications.

Introduction

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. Among these are convenient ways to tailor tests to individual examinees, to link tests without expensive population equating studies, and to interpret scores in terms of predicted behavior on specific test items (Lord, 1980). To enjoy these benefits over a broad range of practical applications, one must have access to sufficiently flexible and economical computer programs to estimate IRT parameters -- for items, for examinees, for populations of examinees -- as an application requires. 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).

The purpose of the present paper is to compare the two programs, with respect to their theoretical approaches and attendant practical consequences. We cannot hope to run comparisons on the vast number of real and simulated datasets required to provide specific advice across the wide variety of situations, applications, and criteria found in practice. We can, however, outline some of the problems which any estimation algorithm must face, describe the character of the solutions offered by LOGIST and BILOG, and offer a few examples to illustrate some important differences and similarities.

The Three-Parameter Logistic Item Response Model

At the heart of item response theory 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 model for test items that are scored either right or wrong (Birnbaum, 1968), abbreviated hereafter by 3PL, this expression takes the following form:

$$P = P(\theta) = c + \frac{1 - c}{1 + e^{-1.7a(\theta - b)}} \qquad (1)$$

were a, b, and c are parameters characterizing the item and e is the mathematical constant. These item parameters have specific interpretations. The c parameter is the probability that a person completely lacking in ability will answer the item correctly. It is called the _guessing_ parameter. The b parameter is a location parameter. It determines the position of the curve along the ability scale. This parameter characterizes the _difficulty_ of an item, in that if a and c are held constant, higher values of b imply lower probabilities of correct response from all examinees. The logistic curve has its inflection point at $\theta = b$. The parameter a is proportional to the slope of the curve at the inflection point. This parameter characterizes the _discrimination_ of the item, in that probabilities of correct response to items with high a values are more sensitive to changes in θ in the neighborhood of the item difficulty.

Considering Equation 1 as a function of θ for fixed values of a, b, and c, yields the trace line or response function of an item (see Figure 1 for example). Assuming the veracity of the model, the importance of the item response function lies in its validity for each examinee regardless of any particular population with which he or she may be associated, and regardless of any other items he or she may be administered.

Insert Figure 1 about here

Note that a linear indeterminacy exists in the 3PL: 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. All c parameters fixed at zerc gives the 2PL, and all a parameters additionally fixed at 1/1.7 gives the 1PL. (But see Andersen, 1973, Rasch, 1960/1980, 1968, and Fischer, 1974, for independent derivations of the 1PL model and discussions of its special properties.)

Most of the same estimation problems arise under all three models. We focus our attention on the 3PL because, since the 1PL and 2PL can be expressed as special cases of the 3PL, any solution to the problems of the 3PL applies to the simpler models as well (although some solutions for the 1PL do not generalize to the 2PL or the 3PL). Our purpose is not to recommend the use of the 3PL over the 2PL or 1PL, or for that matter, over any other model for item responses.

The Theory of Parameter Estimation

Capitalizing on the advantages of Item Response Theory would be a simple matter if true item and true person parameters were known. A practical expedient is to estimate item and person parameters, and proceed to use 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 those 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' (MLE's). To find these parameter estimates for complex likelihood functions in which explicit solutions are unavailable, numerical methods are typically employed to search the parameter space for locations where the first partial derivatives of the likelihood function are zero and where the matrix of second partial derivatives is negative definite. At such locations the likelihood function attains at least local maxima.

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 abilities were known. The response of each additional examinee would provide additional information about a fixed number of item parameters, whose values could be estimated as precisely as desired by simply gathering enough responses and maximizing the likelihood function

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

where

- i indexes items and ranges from 1 to the number of items, n;
- j indexes examinees and ranges from 1 to the number of examinees, N;
- $P_i(\theta_j)$ 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;

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

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

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

However, true abilities are not known. Each additional examinee would introduce 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 approach the solution to this problem differently; LOGIST with a joint maximum likelihood (JML) approach, and BILOG with a marginal maximum likelihoou (MML) approach.

The Joint Maximum Likelihood Approach

The JML approach to estimating parameters in the 3PL 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

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

where the quantities have the same meanings as in Equation 2 except for $\underset{\sim}{\theta}$, which is now a vector of unknown abilities to be estimated along with the item parameters.

While 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, shows that JML estimates of item parameters in the Rasch model are not statistically consistent for increasing N if n is held constant at 2. If both N and n are increased appropriately, however, statistical consistency can be proved for the Rasch model (Haberman, 1977). Simulations conducted by Swaminathan and Gifford (1983) suggest that consistency may also hold for the 3PL as well under the latter circumstances.

The Marginal Maximum Likelihood Approach

The application of MML estimation in IRT originates with Bock and Lieberman (1970). The modern computing algorithms employed in BILOG were developed by Bock and Aitkin (1981). The focus of the

BILOG approach is to remove examinee parameters from the estimation problem entirely and estimate 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 correcc 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

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

The parameters to be estimated by maximizing this marginal likelihood function are the item parameters and, if desired, the parameters that describe the distribution of ability G. 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 statistically consistent estimates of item parameters--conditional, of course, on the veracity of the IRT model.

Once item parameter estimates have been obtained using this marginal maximum likelihood approach, abilities may be estimated. The estimated item parameters are assumed to be equal to their true values, and BILOG can then produce MLE's of ability using Equation 2, or Bayes estimates of ability with an assumed or estimated population distribution (Bock & Aitkin, 1981).

While the MML approach may be more appealing than the JML approach because of its formal statistical properties, there is a price to be paid here, too. In particular, a structure must be assumed for the distribution of abilities in the population of examinees. If _either_ the IRT model of the probabiʟity of a correct response given examinee ability _or_ the assumed model for the distribution of ability in the population are incorrect, the attractive statistical properties fail to hold.

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 cannot be counted upon to yield finite and reasonable item and person parameter estimates. This is so for two reasons discussed in detail in a later section. First, the unthinking application of numerical methods to find maxima of extremely complex functions of many variables is rarely successful.

Second, the likelihood surfaces over which maxima are sought can be very flat, with true maxima at unreasonable values of item parameters. However, regions of the likelihood surfaces only slightly lower than the true maxima may occur at reasonable values of th⸤͜e parameters. LOGIST and BILOG depart from the original JML and MML approaches, and they usually do provide finite and reasonable parameter estimates. They are able to do so by employing not only the assumed IRT model and the observed response data but also prior information (or perhaps, prior beliefs or even prior wishes) about how parameter estimates should look.

In this section we discuss the LOGIST and BILOG approaches to the estimation of item parameters alone using only the observed response data. In subsequent sections we will discuss item parameter estimation that uses information in addition to that provided by the observed response data, and the approaches to ability estimation incorporated in the two programs.

The LOGIST Approach

If requested to estimate item parameters from item response data alone, LOGIST would find estimates of parameters for each item and for each examinee that maximize Equation 3. At this location, 3n equations for the partial derivatives of Lquation 3 with respect to the item parameters, and N equations for the partial derivatives of Equation 3 with respect to the abilities, are zero. In addition, the second derivative for each θ is negative, and the 3-by-3 matrices for the parameters of each item are negative definite. The formulas for the first and second partial derivatives for the 3PL are presented in Lord (1980, Chapter 12).

In principle, the solution could be found by Newton-Raphson iterations involving all item and examinee parameters at once. However, based on considerations of cost and accuracy, it is usually 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 maximum likelihood estimate of a common c parameters for all items that contain little information about their lower asymptotes; more about this in a later section.) This arrangement improves the overall stability and computational efficiency of the procedure by insuring that the subproblem solved in each step is reasonably well determined. A brief summary of the procedure follows: details can be found in the LOGIST User's Guide (Wingersky, et al., 1982).

Insert Table 1 about here

In Step 1, examinee ability and the most well-determined item parameter, b, are estimated while the a and c parameters are held fixed. Within this step, stages alternate between estimation of b for all items and of thetas for ali examinees until a loose criterion (default 200%) for the increase in the likelihood function between stages is met. Examinees with no correct or all correct responses on all items reached by that examinee are excluded from the estimation procedure. Examinees for whom the maximization process yields an infinite ability estimate are assigned floor and ceiling values. Extrapolation procedures are employed for item difficulties from blocks of items to speed convergence when the observed data matrix is structured so that not all examinees take all items.

In Step 2, the estimated abilities are held fixed at the values obtained in Step 1. Item parameters, including COMC, are estimated within stages of this Step until a slightly tighter criterion (default 20%) for the increase in the likelihood function is met. By waiting until Step 2 to estimate the a and c parameters, which are more difficult to estimate, the iteration process is more stable.

Step 3 is a repetition of Step 1 with a tighter convergence criterion. Step 4 is a repetition of Step 2 with an even tighter convergence criterion, with the exception that COMC is not reestimated. LOGIST resolves the linear indeterminacy of the 3PL by standardizing the estimated abilities between the range of -3 to +3, so that the abilities within that range have a mean of zero and a standard deviation of one.

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 to them have been proven to be statistically consistent, but some simulation studies (Swaminathan & Gifford, 1983) suggest that the JML estimates for the 3PL do appear to behave better as both test length and examinee sample sizes increase. Better behavior for increased test length is not surprising when one considers the nature of LOGIST estimation cycles. In the steps in which the item parameters are estimated, examinee parameters are treated as known, whereas they are in fact only estimates. The fewer responses used to estimate an examinee ability, 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 the user to restrict its use to data with at least 20 items per person and at least 800 to 1000 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 presently available for either tests of model fit or large-sample standard errors. Nevertheless, the matrix of second derivatives from which, under standard MLE procedures, the variation of estimates around their true values is forecast, can be computed. It becomes an empirical question as to whether, in this situation, these forecasts of variation of estimates around their true values is practically useful. Wingersky and Lord (1984) investigated this question and demonstrated 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 would find those values of the item parameters that maximize Equation 4. In principle, this maximum can be found by proceeding in a series of Newton-Raphson steps that involve the vector of first derivatives and the matrix of second derivatives for all item parameters. Such a straightforward solution was first presented in Bock and Lieberman (1970). However, this solution becomes impractical for more than about 20 items.

Bock and Aitkin (1981) reexpressed the required first derivatives in a way that led to a more practical computing algorithm. In the Bock-Aitkin development, the population ability 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 abilities can take are those represented by the quadrature points. Although the value associated with a particular examinee is not known, the probabilities that it takes each of the possible values can be calculated via 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 ability. Having done this, one can maximize the <u>expected value</u> of the log of Equation 2.

To obtain the expected value of the log of Equation 2, however, means that we must know the item parameters and G. The point of the exercise, of course, is that we do not know the item parameters, and may not know G either. In iterative cycles, however, one can recompute the desired expected value with updated estimates of the values of item parameter estimates and, if desired, G, that maximized the preceding expectation. These are exactly the steps of the EM algorithm (Dempster, Laird, & Rubin, 1977), in the special case of 'missing multinomial indicators' since abilities are assumed to take on only a finite number of values.

Since the progress of the EM-algorithm can be very slow, convergence is hastened by Ramsay's (1976) acceleration method, applied to each estimated a and b separately. After a set number of EM cycles (10 default) or sufficiently small changes in estimates of a, and the product of a and b (called the 'intercept') a final Newton-Raphson step is taken to find the maximum of Equation 4 and provide standard errors for item parameter estimates.

The key idea in the MML solution is that the uncertainty associated with each examinee's unknown ability is accounted for by effectively spreading his or her ability across potential values it might take, in accordance with the probabilities given by the posterior distribution. To do this successfully, one should have enough quadrature points to ensure that the typical examinee's posterior distribution is nontrivial for at least 3 or 4 points.

The greater the number of items taken by an examinee, the more concentrated the resulting posterior distribution, so more quadrature points are needed as the number of items taken by an examinee increases. BILOG defaults to using 10 quadrature points for 50 or fewer items per person and 20 quadrature points for more than 50 items per person.

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 means of constraints upon the estimated densities at the quadrature points. If G is assumed normal or specified by the user, these densities are specified as fixed from the start of the estimation procedure. If G is estimated concurrently with the item parameters, the quadrature points and item parameters are readjusted by a linear transformation that standardizes the estimated examinee population distribution.

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

Item Parameter Estimation Using Information External to the Response Data

Under the 3PL (and also the 2PL), 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 upon the data and the model, but at least partly upon the method and the strength with which prior beliefs about how item parameter estimates 'ought to' look. In this section we discuss first the nature of problems encountered by the 3PL estimation procedures described in the previous section that use only the observed response data. We tnen describe the approaches used by LCGIST and BILOG in incorporating information in addition to the observed response data to handle these problems.

Problem 1: 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 zero 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, 1985). The 2PL model with an assumed normal distribution for examinee ability is nearly identical to a one-factor model for dichotomous variables, with the item discrimination inversely related to a corresponding unique variance. The same relationship holds between the a parameters in the 3PL and uniquenesses in a factor analytic model with a lower asymptote. As the unique variance for an item approaches zero, as in a Heywood case, a becomes infinite. After fifty years' experience with factor analysis, it comes as no surprise to find that the maximizing values for item discriminations under the 2PL or the 3PL are sometimes infinite.

It has been speculated that without constraints upon their values, at least one a will become infinite in the attempt to fit the 2PL or the 3PL to any set of response data (Wright, 1977). The reader is invited to verify that not all datasets lead to infinite

ML solutions by fitting the 2PL to the data in Table 2 by JML. The fact that some datasets do not yield infinite estimates offers little comfort, however, as long as 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.

Insert Table 2 about here

Problem #2: Multicollinearity

Even when constrained item parameter estimates under the 3PL are finite, they need not be reasonable. It is easy to see how this can occur. While an item response function traces the probability of a correct response across the entire range of ability, data are available in only a limited region: the neighborhood in which the abilities of the sample of examinees lie. Even if the true abilities were known, only an approximation of the response curve would be observed, 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,c) triples can capture the data nearly equally well but disagree about what happens where there are no data. Figures 2a and 2b show two items with data fit well within the neighborhood of (-1,+1) by two (a,b,c) that would lead one to different conclusions about the nature of the items.

Insert Figures 2a and 2b about here

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 is changing very gradually and therefore a local maxim.m is difficult to find. In extreme cases, the surface is not changing at all, in which case there is no local maximum and the solution fails entirely.

Methods of Incorporating External Information

A Bayesian solution to item parameter estimation incorporates external information through the imposition of prior distributions on item parameter estimates. Such priors can reflect such beliefs as "the c parameters for items that cannot be estimated from the data are probably similar to those that can be" and "1000 is not a reasonable value of an a parameter." 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 given by 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 those values that maximize the posterior probability. Bayesian modal estimates have been developed for the JML by Swaminathan and Gifford (1983) and for the MML by Mislevy (1986). The large-sample properties of modal estimates are determined by the large-sample properties of the likelihood functions, either JML or MML, used to obtain them. Thus indices of fıt and large-sample standard errors formally hold for the Bayesian extension of MML but are not formally supported for 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 distribution: Parameters estimable from the observed data alone would then receive Bayes estimates that were similar to their maximum likelihood estimates. Infinite and extreme estimates would be pulled in to finite and reasonable values. Similar effects can also be achieved informally through constraints upon the maximum likelihood procedure. The practical problem under both formal and informal Bayesian solutions is to specify priors or procedures that give rise to 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, in part by employing simple constraints to handle extreme item parameter estimates. Floors and ceilings are specified for the values that estimates of a and c can take; .01 and AMAX are floor and ceiling values for the a parameter, 0 and .5 are floor and ceiling values for the c parameter. This is equivalent to specifying a uniform prior distribution on the intervals (.01,AMAX) for a, and (0,.5) for c. If neither a nor c for an item exceeds a boundary in a given cycle with provisionally fixed values of ability estimates, then none of the item's parameter estimates will be affected by this prior. 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 ability estimates, so that the parameter estimates for all other items and all examinees are affected, though probably minimally, whenever a single parameter estimate for any item takes on a boundary value.

While LOGIST provides default boundary value settings, the manual shows how to estimate boundary values that can be more appropriate 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 there are many more estimates equal to the provisional AMAX than there are slightly less than the provisional AMAX, the suggestion is made to raise AMAX before continuing the run to completion. If there are estimates equal to AMAX and this value is substantially above the next lowest estimate, the suggestion is made to lower AMAX before continuing the run to completion. Such simple procedures informally incorporate the user's beliefs not only about reasonable _values_ for item parameter estimates, but also about reasonable _distributions_ of these values. Such procedures, in which individual values are estimated with respect to a population distribution that is 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 upon the other items in the test.

Another LOGIST constraint upon estimates of c can also be thought of in an empirical Bayes framework: the MLE estimate 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 c's 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 c's were not so restricted, severe multicollinearity would result and the poorly determined c would have a large (and undue) influence upon 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 having 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. Since 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 towards 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. The normal distribution is used for the b's, the lognormal for the a's, and the beta for the c's. Priors 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 towards 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 and this contribution increases with sample size. The s:cond term is the contribution from the prior and remains constant with respect to sample size. Shrinkage towards the (possibly estimated) mean of parameters of this parameter type therefore decreases as sample size increases.

Default specifications for the strength of the priors are a standard deviation of 2 for b's, a standard deviation of .5 for log a's, and a weight of 20 responses from low ability examinees for c's (see Swaminathan & Gifford, 1983, for details on using the beta prior in this way). Early experience with BILOG indicates that the effects of these default priors are minimal for b parameters. These are the item parameters most easily determined from the observed data. Even when the sample of examinees is small, say under 500, the prior has little influence on the estimation of item difficulty. Stronger shrinkage towards their respective means is observed for the a and c parameters as sample size decreases. With extremely small samples, say, under 100 examinees, a model is approached in which all the a parameters and all the c parameters have nearly equal common values. This gives, in effect, not a 3PL model, but a "(n + 2)/n" PL model.

The cost of obtaining more reasonable estimates by incorporating formal prior distributions (some of the features of which may be estimated from the data at hand) is twofold. First, as with the more informal LOGIST constraints, the expected estimates for a given item depend upon 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. The prior on c is a case in point. BILOG assumes a common distribution for c parameters; its mean is determined primarily by well-estimated c's. All c's shrink towards this value. If the true values of poorly determined c's depart from this common mean, their estimates will be biased accordingly. Such biases are reduced with 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.

Ability 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 provisions for estimating individual abilities, either in the same run as item parameters are estimated, or with respect to previously-estimated item parameters. In either case, the item parameters are treated as known (see Lewis, 1985, and Tsutakawa, 1986, on taking into account the uncertainty associated with item parameter estimates).

We have seen that maximum likelihood estimates of ability are integral to LOGIST's JML inspired item parameter estimation. Point estimates of abilities 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 ability 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. MLE's are one BILOG option; Bayes mean estimates, more sympathetic to the MML approach to item parameter estimation, are another. We now describe briefly the procedures by which LOGIST and BILOG produc꜀ point estimates of ability for a given examinee with a particular response pattern, assuming the item parameters are known.

Maximum Likelihood Estimates

Both LOGIST and BILOG are able to produce maximum likelihood estimates of ability. For a given set of item parameters and response patterns, LOGIST and BILOG estimates will differ from each other only insofar as the details of the two numerical procedures are different. Lord (1980, Page 54, Equation 4-20) is the likelihood equation both programs solve for ability. Both use Newton-Raphson iterations from a starting value based on a standardized percent correct adjusted for guessing. If a provisional value of estimated ability is far from the maximizing value, Newton-Raphson steps can diverge. Both programs reduce this possibility by limiting stepsize and forcing steps to be in the direction that increases the likelihood. If the number of items to which an examinee has responded is large, the MLE estimated ability for an examinee is approximately normally distributed, with mean estimated ability equal to the true ability and large sample variance given by Lord (1980, Page 71, Equation 5-5).

A unique finite maximum exists under the 3PL for most response patterns above chance level, although multiple maxima are occasionally encountered. This is more likely to happen 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 do so.

For response patterns yielding infinite MLE's, BILOG provides floor and ceiling values of -4 and +4, flagging such values with a 'dummy' standard error of 999. LOGIST implements a more complex procedure. If abilities are being estimated from previously estimated item parameters, examinees with zero and perfect scores or examinees that answered only a few items may be included in the estimation. Infinite ability estimates will then be set at default or user specified boundary values. In the more typical LOGIST run, where item and ability parameters are estimated simultaneously, more constraints are imposed since ability estimates will be used in the next cycle of item parameter estimation. In this case, by default, examinees with zero and perfect scores and examinees who answered fewer than 1/3 of the items presented to them are assigned the 'dummy' abilities of -999999, +999999, and -333333 respectively, and excluded from the estimation of item parameters. Examinees who do not fall into these three categories but whose estimated ability tends to become infinite are given default boundary values of -7 and +4. These default values, which may be changed by the user, were chosen because abilities below -7 have almost no effect on the value of the likelihood function, and +4 is higher than ability estimates typically obtained.

Bayes Estimates

Bayes estimates of ability can be produced by BILOG. As a byproduct of BILOG item parameter estimation, one obtains expected values of the density of the examinee population at each of the quadrature points. The posterior probability that an examinee ability is equal to a particular quadrature point can then be obtained from this information. One can then summarize what is known about an examinee in terms of a Bayes mean estimate, i.e., the mean of this estimated posterior distribution, and its associated standard deviation. Bayes mean estimates are sometimes called 'expectation a posteriori' or EAP estimates.

Properties of EAP estimates are described by Bock and Mislevy (1982). By using a population distribution in the course of ability estimation, finite values are obtained for all response patterns, including those that yield infinite MLE's. The reasonableness of EAP estimates obtained for these response patterns depends upon 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 were in fact representative of a populatirn of interest, but less so if it were not. For example, reasonable estimates would result for point estimates of zero scores of third graders if an estimate of the third-grade population distribution were employed. However, overly high estimates would probably result if a fifth-grade distribution were used.

A Comment on Estimating Ability Distributions

Consider the problem of estimating the distribution G of ability in a population of interest, from the item responses of a sample of examinees. It is somewhat paradoxical that the distribution of ability estimates, each of which is in some sense optimal for the particular examinee, is not necessarily a good estimate of G. Maximum likelihood estimates tend to have too large a variance; Bayes 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 ability from either LOGIST or BILOG will not converge to the true distribution of ability 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 G. In order to have point estimates of ability for each examinee that yield a consistent estimate of G, one would have to sample a value at random from the posterior distribution of each examinee. This would provide 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 ability would require these 'noisy' estimates that are decidedly nonoptimal for each examinee considered individually.

Additional Considerations

The preceding sections have dealt with the foundations of the approaches by which LOGIST and BILOG produce estimates of item and person parameters. This section deals with some miscellaneous topics that may be of interest to the prospective user.

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 by reasons intended by the researcher and sometimes not. LOGIST and BILOG handle these situations in the same ways. Methods of handling three types of nonresponse are incorporated in both programs.

Most easily dealt with are potential examinee/item combinations that are missing by design. Different examinees may take different forms of overlappir tests, for example, so that they have no opportunity to provide responses to items not presented to them. It is intuitively clear that these occurrences of nonresponse 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 ignorability may continue to hold when patterns of nonresponse might be related to ability or item parameters. If these patterns of nonresponse are determined wholly by previous observable responses, as in adaptive testing, then they remain ignorable (Mislevy, 1985). 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 realized in the sample. This feature proves convenient for linking tests through common items.

Less clear cut 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 phenomenor. 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 obse.vations. 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 'wrong' 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 has obviously encountered and decided to omit. Encoding these observations as wrong is palatable for free response items, but less so for multiple choice items. Had the examinee guessed at random, as others with equally little knowledge have undoubtedly done, 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 guess at random when confronting an item for which they feel no preference among response alternatives. The best one car: do in LOGIST and BILOG is to treat such observations as partially correct, with the 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). If examinees omit only when their probability of responding correctly is the chance value, modifying the JML and MML likelihood functions in this manner gives the expectations of the corresponding functions that would obtain had there been no omits, conditional on the observed pattern of omissions (Mislevy & Wu, 1987).

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 $\hat{\theta}$ 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, i.e., the two BILOG scales will differ and the two LOGIST scales will differ, as a function of differences in the averages and dispersions of ability in the two samples, as well as in the sampling variation generally associated with any estimation procedures. A linear transformation, found by a procedure such as Stocking and Lord's (1983), puts the two sets of LOGIST estimates on approximately the same scale; or the two sets of BILOG estimates on approximately the same scale;

remaining differences 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 parameter of the distribution of ability 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 an individual ability for each examinee; increasing the number of examinees increases the number of estimated abilities thereby increasing the accuracy of the distribution of estimated ability. But the relationship between an estimated distribution of ability and an estimated distribution of estimated ability is nonlinear in a way that depends 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 items are appropriate for the examinee sample, this nonlinearity becomes negligible only if (1) the test is long enough so that estimated abilities are indistinguishable from true abilities and (2) examinee sample sizes are large enough so that the distribution of these estimated abilities can be accurately obtained.

Incorporating the Results of Previous Runs

Both programs save files of parameters that can be input to subsequent runs, either for all parameters or reduced parameter sets. In addition, steps can be taken to constrain the estimates of selected parameters without constraining those of other parameters, as would be done, for example, when calibrating new items onto an existing scale.

LOGIST accomplishes this by allowing the user to hold selected parameters for any item fixed at user provided values. It is also possible to estimate item parameters conditional on fixed ability estimates from a previous run. BILOG allows the user to place different priors on the parameters of different items. If previous analyses were available for a subset of items, one could employ priors with means based on the previous estimates and dispersions based on the standard errors of those estimates. Weak priors would then be imposed on the new items.

Diagnostic Information

Both LOGIST and BILOG provide information on the progress of the numerical procedures invoked. This type of information is vital if the user is to monitor the successful completion of the program.

Strictly speaking, of course, it is a foregone conclusion that 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.

After each iteration and for the final solution, BILOG provides the value of -2 times the log of the likelihood factor of the criterion function. The criterion function is the product of the prior distribution for item parameters if there is one, and the appropriate marginal likelihood function--i.e., if omits are not given partial credit, the actual MML likelihood; if they are, the expectation of the complete-data MML likelihood, conditional on the observed pattern of omissions. As the number of examinees increases, the behavior of the criterion depends increasingly on just the likelihood term. In large samples, differences between values obtained under the 1PL, 2PL, and 3PL are approximately chi-square under the assumption that the more restrictive of two models being compared is correct. The degrees of freedom is the number of additional parameters estimated in the less restrictive model. For tests of ten items or fewer, when all examinees have taken all items without omits, a chi-square test against the general multinomial alternative is also given. The limiting distributions for both indices are approached less rapidly if priors are employed for item parameters. While the corresponding chi-square distributions generally provide a guideline to gauge the degree to which estimates maximize the likelihood term, the strict interpretation of associated probability levels may not be justified for small samples of examinees (particularly if priors are employed for item parameters).

We strongly recommend that every calibration of item and person parameters be examined by means of plots of observed verses 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 fitting problems. Through this mechanism the user can detect unsatisfactory limits on values of parameter estimates for LOGIST or unsatisfactory priors placed on some items for BILOG. These are conditions that 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 one. These problems are not correctable since 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.

It may also be useful to examine pseudo-chi-squares (Yen, 1981). These statistics do not actually follow a chi-square distribution, but may be useful as a rough guide in interpreting the severity of model departures. BILOG provides line printer plots and pseudo-chi-squares of this type. Their usefulness appears limited with short tests (fewer than, say, 15 or 20 items) because they require treating point estimates of abilities as known quantities.

Both programs lack two other potentially useful diagnostic tools. One would be the residuals from item/ability plots with different symbols to distinguish different subgroups of examinees. These plots could demonstrate differential item performance with respect to gender or ethnicity groups, educational treatments, or points in time. A second would be residuals computed from the matrices of observed interitem correlations and those predicted by the IRT model (McDonald, 1980). An examination of such a matrix could suggest additional factors or lack of conditional independence among subsets of items.

Ease of Use (or Lack Thereof)

Neither LOGIST nor BILOG is particularly easy to learn to use. It is an irrefutable fact that in order 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. The reason for this is that the assumed model and the observed response data are not sufficient to guarantee 'reasonable' results under the 3PL. Both programs offer default settings that get the novice started, 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

We have noted that 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. What we can do is to illustrate an application of the programs to two simple simulated datasets, and examine costs and recovery of generating parameters. The results pertain to the program versions publicly available at the time of this writing, LOGIST 5 and BILOG 2.2.

The Data

We analyzed responses from simulated examinees to an artificial test containing 45 items comprised of three replications of 15 four-choice items. Generating values of item parameters and of abilities for 1500 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. (We postpone until later the question of whether this method of generating data--or, for that matter, any other method--produces a 'fair' comparison of the programs.) Item response data were then generated by first computing the model probability of a correct response to each item/examinee combination, 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. Abilities were restricted to the range (-7,+4).
3. Individual c's were estimated only for items with b - 2/a > -3.
4. The default 4-step estimation procedure was used.

Item and examinee parameter estimates are 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 ability distribution was estimated jointly with the item parameters.
2. Ten quadrature points were used.
3. Default specifications of prior distributions were employed for item parameters of each type, so that the locations were estimated from data and the dispersions were fixed at the program defaults.
4. Default values controlled the number of cycles and the convergence criterion.
5. To facilitate cost comparisons for different types of data, two different data storage methods were used in each problem. One uses a faster algorithm that is applicable only to data for which all examinees take all items without omits; the second, slower, algorithm must be applied when omits and/or not-presented items can occur.
6. Bayes (EAP) ability estimates were produced for each examinee. While these are not required to obtain item

parameter estimates, the typical user would probably ask for them.

Results

Comparisons of the LOGIST and BILOG item parameter estimates for the 45-item test with the true values are shown in Figures 3 through 5. Both procedures appear to recover the true parameters equally well. Examinee parameter estimates are shown in Figure 6. As might be expected, BILOG's Bayes estimates shrink modestly toward the population mean, while LOGIST's MLEs are slightly more dispersed than the true values, with a few outliers for near-chance-level patterns.

Insert Figures 3, 4, 5, and 6 about here

Item parameter estimates from the two programs for the 15-item test, plotted against the true values appear as Figures 7 through 9. BILOG appears to recover the true values better. Ability estimates are shown in Figure 10. The shrinkage of Bayes estimates and the dispersion of MLEs noted for the 45-item test have been accentuated. The authors of LOGIST do not recommend its use for tests as short as our 15-item trial. These results serve to confirm the prudence of the authors' guidelines.

Insert Figures 7. 8, 9, and 10 about here

Execution times of the two programs are shown in Table 3. 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 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 to each other. Given that both programs ended up at essentially the same place, one might prefer LOGIST because it got there faster, or BILOG because of the statistical properties associated with the procedures by which it traveled. These statistical properties, however, are only strictly applicable if omits are not given partial credit, since asymptotic results are not available for the 'pseudo' marginal likelihood that then dominates the BILOG criterion.

However, the similarity of the results for LOGIST and BILOG on the 45-item test does not necessarily imply that the user will be indifferent as to choices in more demanding applications such as long equating chains or several cycles of item pool refreshment in adaptive testing. It is quite possible in these circumstances that potential subtle differences between the two programs might have ramifications that lead to an obvious choice. Our 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.

We would also like to comment on the difficulty of constructing 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 (1) realistic and (2) known. But our notions of what 'realistic' means are determined by what available programs provide, and we cannot count on any program to tell us the true parameters for any dataset of reasonable size dataset. 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 (private communication, 1984) has pointed out that our simulation may favor LOGIST somewhat by using previous LOGIST estimates as generating values. From one perspective, the generating values used in this example can be viewed as representing fewer than 3 parameters for each item. This is so because some items have identical lower asymptotes arising from a common c value estimated for poorly determined c's 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 c's might have a tendency towards 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 we can see from this potential for circular reasoning is to accumulate experience over a broad range of problems. One path that future research should follow has been lead by Yen (1985), who compared estimates of the two programs over a broader range of generating values. A second path would not focus 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.

Conclusion

The joint maximum likelihood approach to estimating parameters in the 3PL originates with Birnbaum (1968), and the theory of marginal maximum likelihood estimation originates with Bock and Lieberman (1970). By setting appropriate switches, the user can ask LOGIST to produce JML estimates and BILOG to produce MML estimates. This user will soon find that the unadulterated version of either approach cannot be counted upon to produce finite and reasonable parameter estimates. LOGIST and BILOG depart from the original approaches by employing prior information about how the parameter estimates should look. The spirit is obviously Bayesian; the details of LOGIST are less formally so than those of BILOG.

From what we have seen so far, 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, so 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.

We would recommend that the user with short tests and/or small examinee samples consider using BILOG. In these situations, BILOG's more formal Bayesian procedures are likely to provide reasonable results, although 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 that look less like the 3PL and more like a model with individual b's but common a and c estimates. If Bayesian ability estimates are requested f^r short tests, they will be shrunk noticeably towards the center of the estimated examinee distribution. In these situations, then, the reasonableness of the results depends upon the reasonableness of the prior structure.

References

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

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and 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: An application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D., Gibbons, R. D., & Muraki, E. (1985). Full information item factor analysis (MRC Report No. 85-1). Chicago: National Opinion Research Center.

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). EAP estimation of ability in a microcomputer enviro.ment. 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). Eifuhrung 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 of 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, June). 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 Monograpb No. 7. Psychometric Society.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 37, 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.

McDonald, R. P. (1980). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 33, 205-233.

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

Mislevy, R. J. (1985, October). Stochastic test designs. Paper presented at the Invitational Workshop on Item Response Theory, CITO, The Netherlands.

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., & Wu, P. K. (1987, in progress). Inferring examinee ability when some item responses are missing.

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

Ramsay, J. O. (1976). Solving implicit equations in psychometric data analysis. Psychometrika, 40, 361-372.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research. Chicago IL: University of Chicago Press

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. (1983). Estimation of parameters in the three-parameter latent trait model. In D. Weiss (Ed.), New horizons in testing. New York: Academic Press.

Tsutakawa, R. K. (1986). Approximations for Bayesian ability estimation. Paper presented at the Office of Naval Research Contractors meeting on model-based measurement, Gatlinburg, TN.

van den Wollenberg, A. L. (1979). The Rasch model and time-limit tests: An application and some theoretical contributions. 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. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.

Yen, W.M. (1985). A comparison of the efficiency and accuracy of BILOG and LOGIST. Paper presented at the annual meeting of the Psychometric Society, Nashville, TN.

Table 1

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

*COMC is the MLE estimate of a single common value for the c parameters of items for which insufficient data is available to estimate individual c's.

Table 2

Item Response Data That Yields Finite 2PL Estimates When Fit by JML

(wrong answers coded as zeros; right answers coded as ones)

	Item Number						
Examinee	1	2	3	4	5	6	7
1	0	0	1	0	0	0	1
2	0	0	1	0	0	1	0
3	0	0	1	0	1	0	0
4	0	0	1	1	0	0	0
5	0	1	0	0	0	0	1
6	0	1	0	0	0	1	0
7	0	1	0	0	1	0	0
8	0	1	0	1	0	0	0
9	1	0	0	0	0	0	1
10	1	0	0	0	0	1	0
11	1	0	0	0	1	0	0
12	1	0	0	1	0	0	0
13	0	1	1	1	1	1	0
14	0	1	1	1	1	0	1
15	0	1	1	1	0	1	1
16	0	1	1	0	1		1
17	1	0	1	1	1	1	0
18	1	0	1	1	1	0	1
19	1	0	1	1	0	1	1
20	1	0	1	0	1	1	1
21	1	1	0	1	1	1	0
22	1	1	0	1	1	0	1
23	1	1	0	1	0	1	1
24	1	1	0	0	1	1	1

Table 3

Execution Times of LOGIST and BILOG on Two Simulated Datasets

15-Item Test	CPU Seconds
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

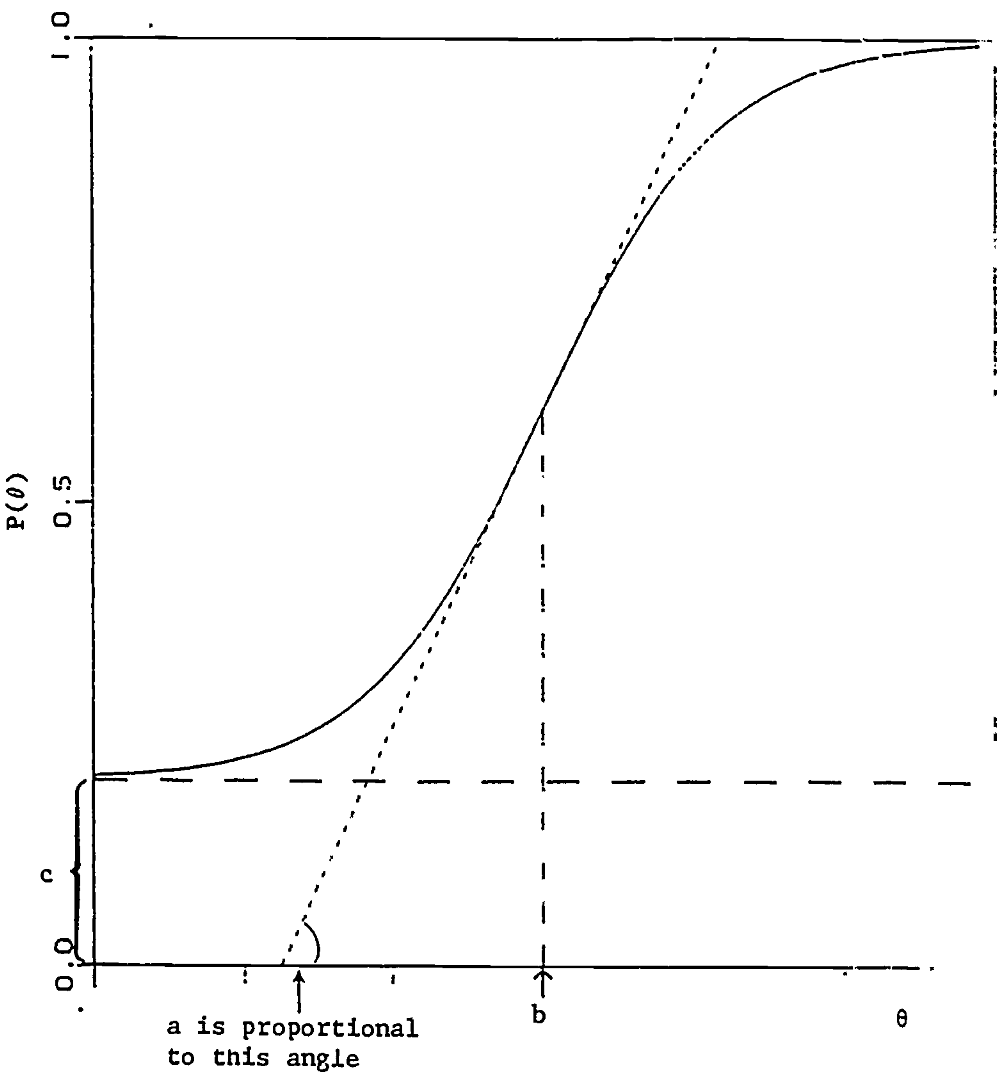


Figure 1. A typical item response function (solid line), with discrimination, difficulty, and guessing parameters denoted by a, b, and c.

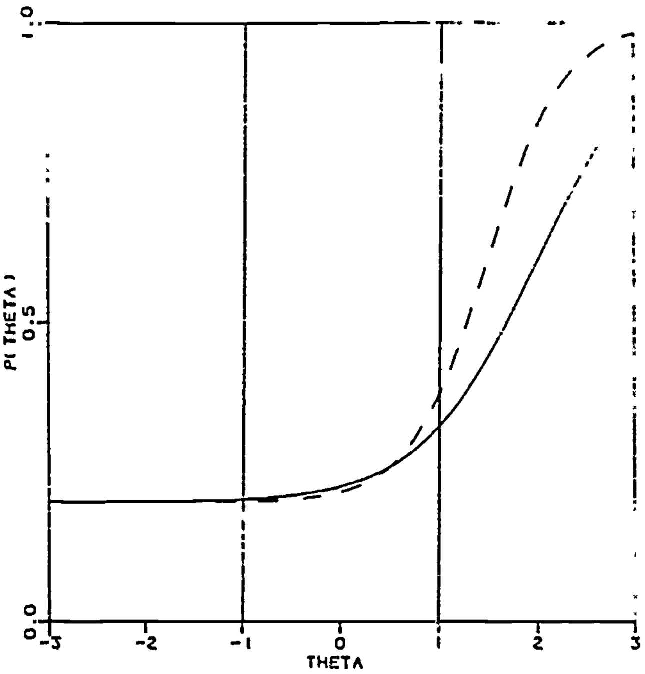


Figure 2a. Two estimated item response functions for a hard item.

——— $\hat{a}$ = 1.0, $\hat{b}$ = 2.0, $\hat{c}$ = .2

- - - $\hat{a}$ = 1.5, $\hat{b}$ = 1.5, $\hat{c}$ = .2

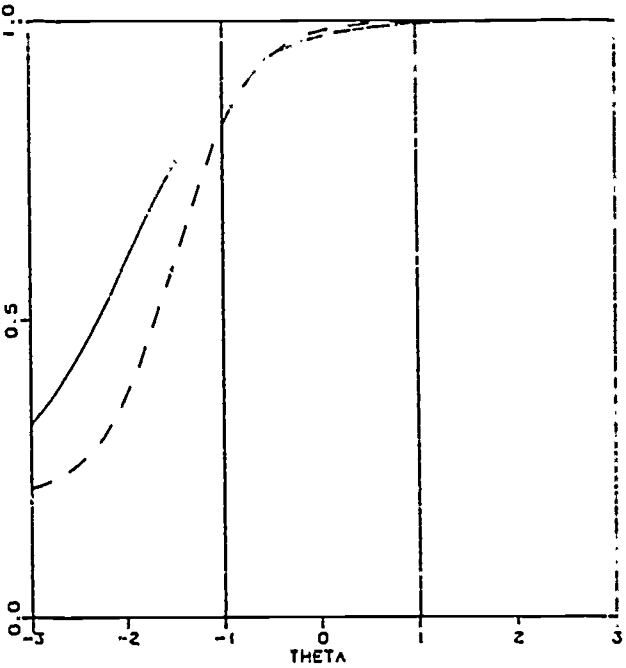


Figure 2b. Two estimated item response functions for an easy item.

——— $\hat{a}$ = 1.0, $\hat{b}$ = -2.0, $\hat{c}$ = .2

- - - $\hat{a}$ = 1.5, $\hat{b}$ = -1.5, $\hat{c}$ = .2

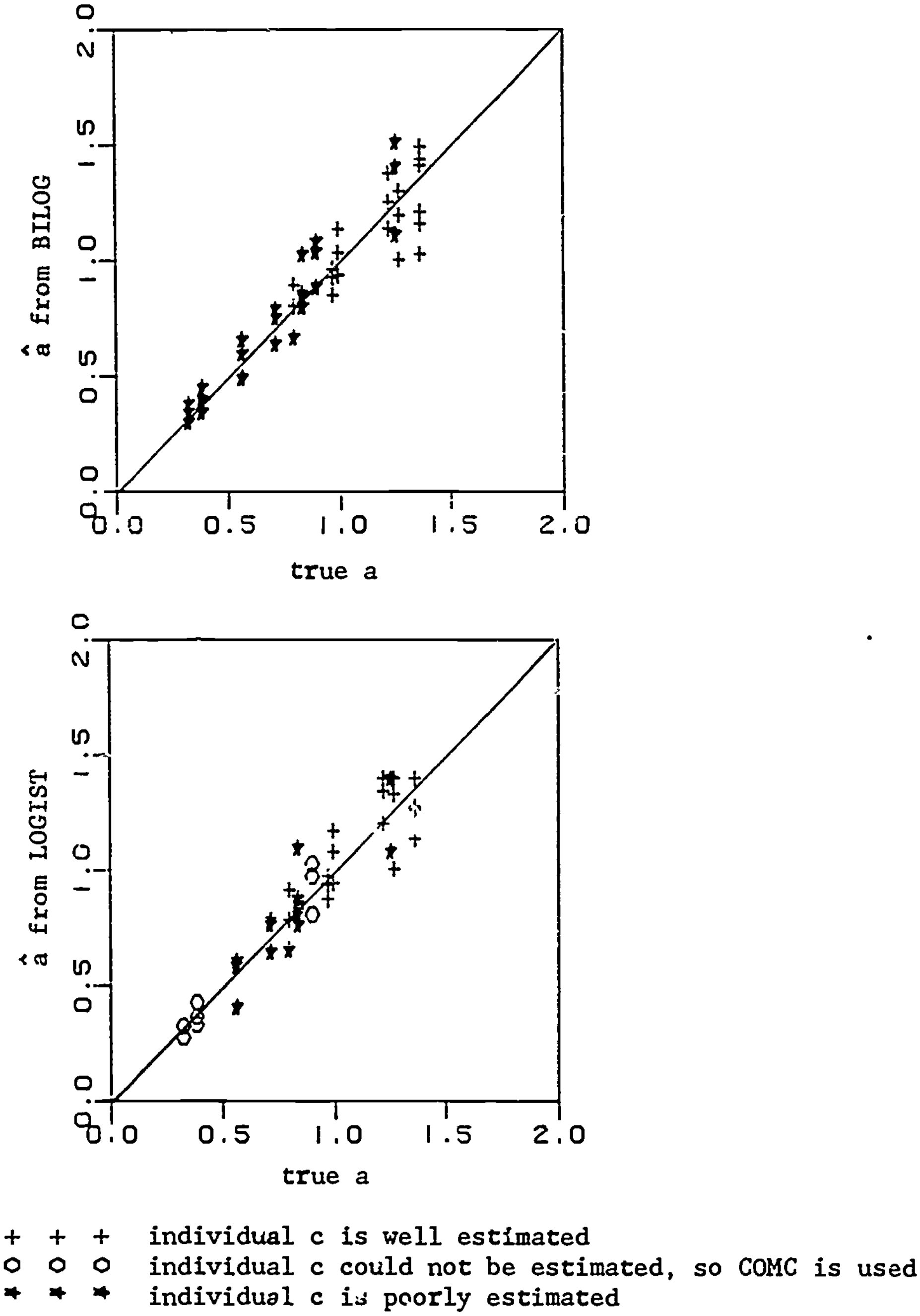


Figure 3. Estimated a's for BILOG (top) and LOGIST (bottom) compared to true a's, n = 45.

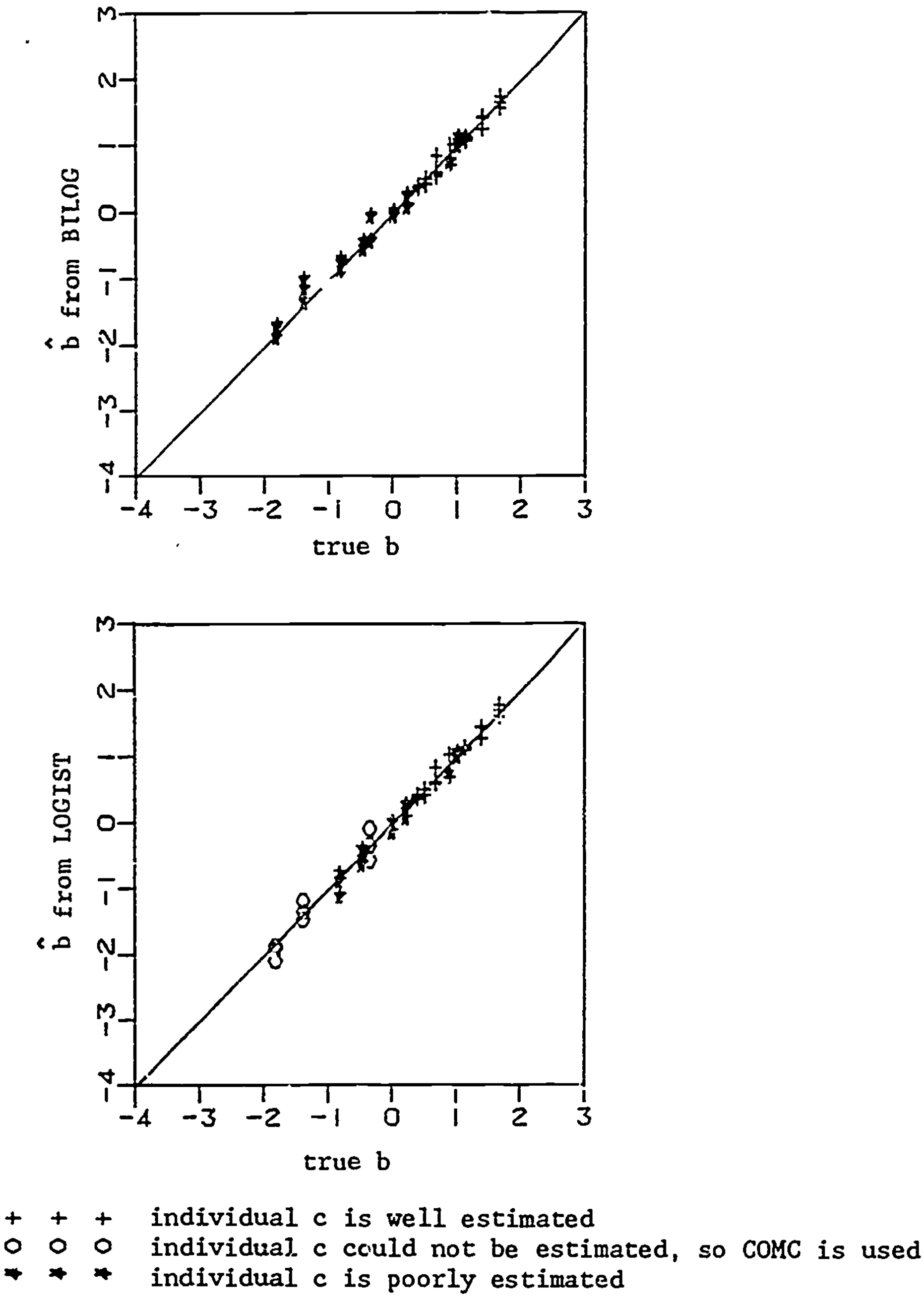


Figure 4. Estimated b's for BILOG (top) and LOGIST (bottom) compared to true b's, n = 45.

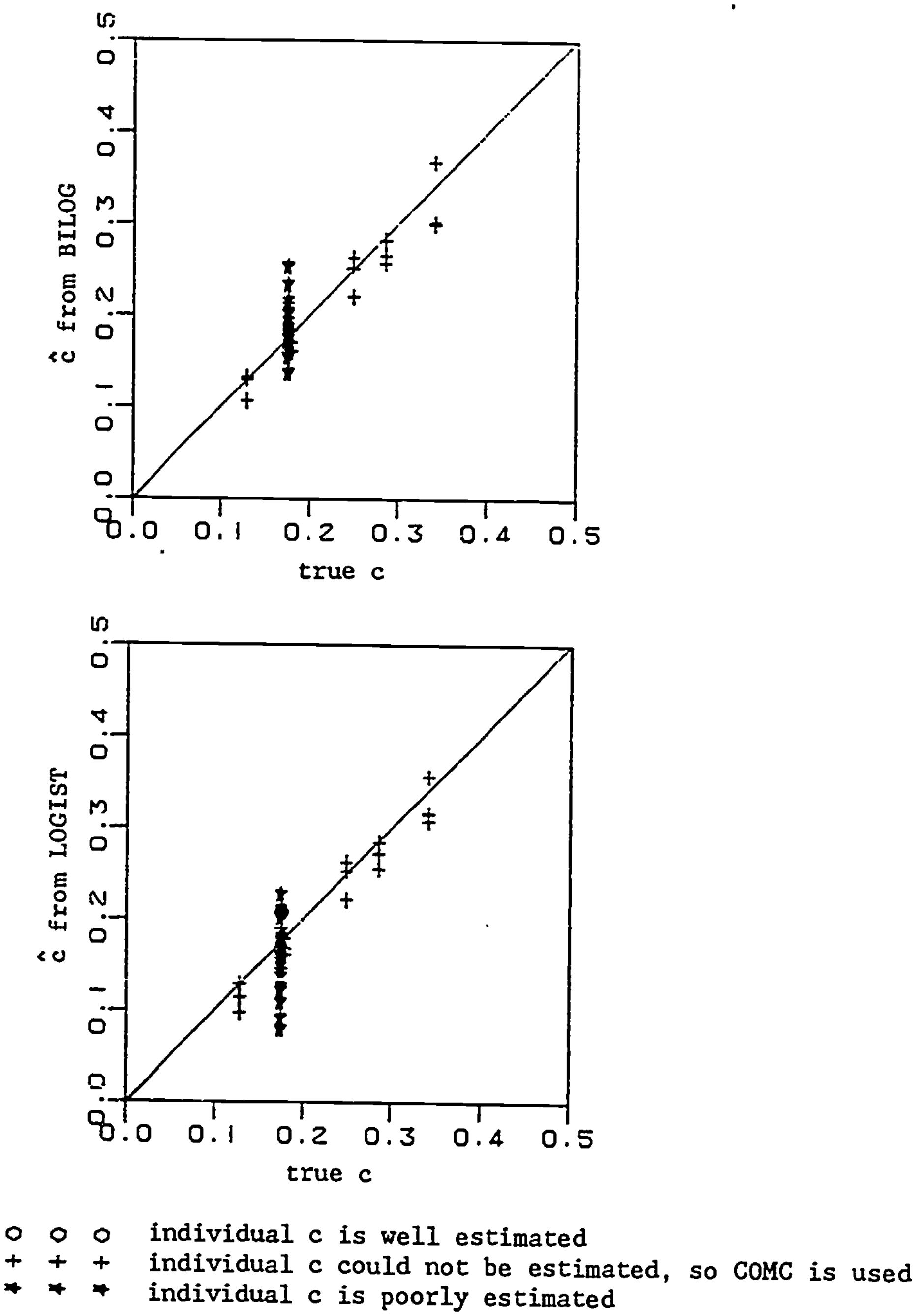


Figure 5. Estimated c's for BILOG (top) and LOGIST (bottom) compared to true c's, n = 45.

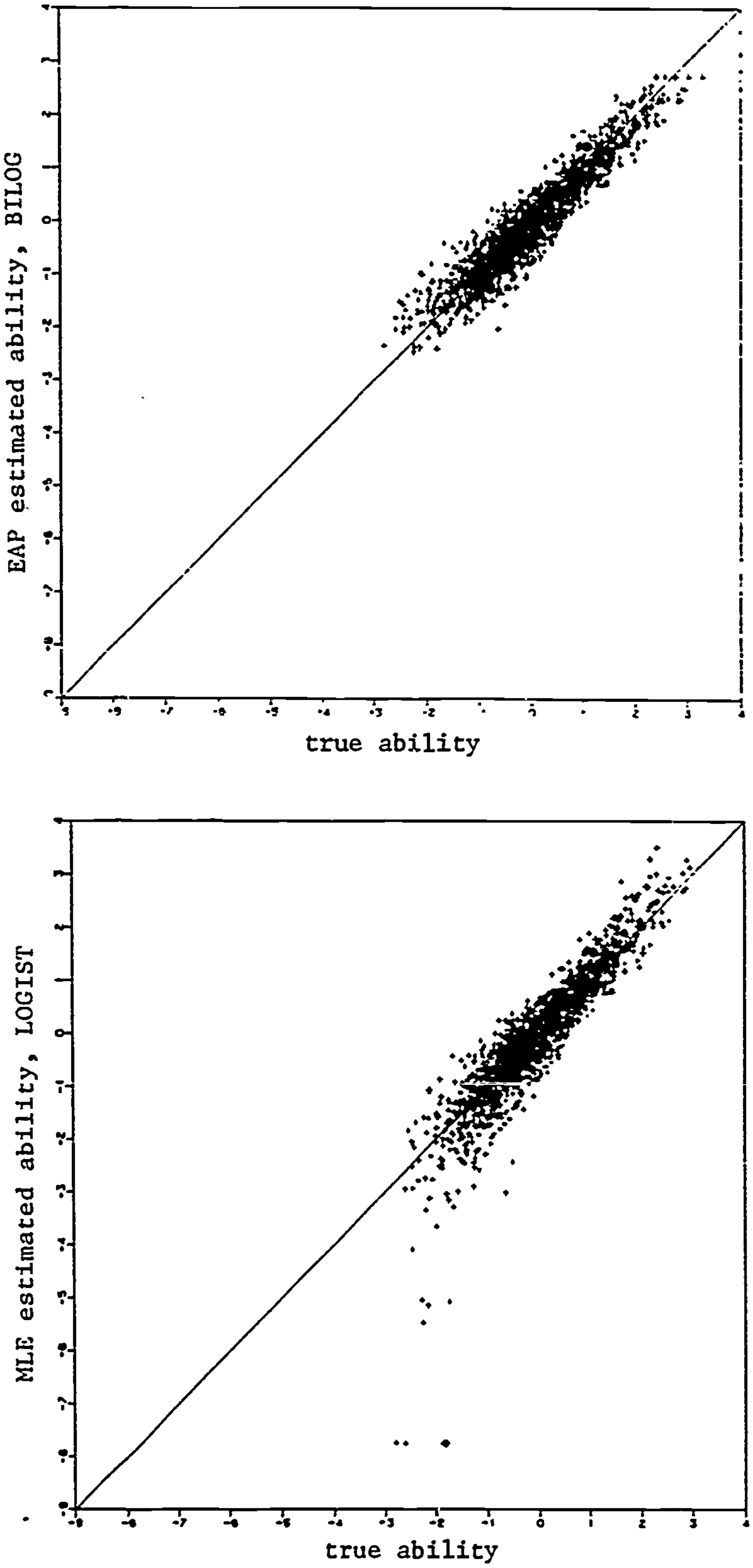


Figure 6. Estimated abilities from BILOG (top, EAP estimates) and LOGIST (bottom, MLE estimates) compared to true values, n = 45.

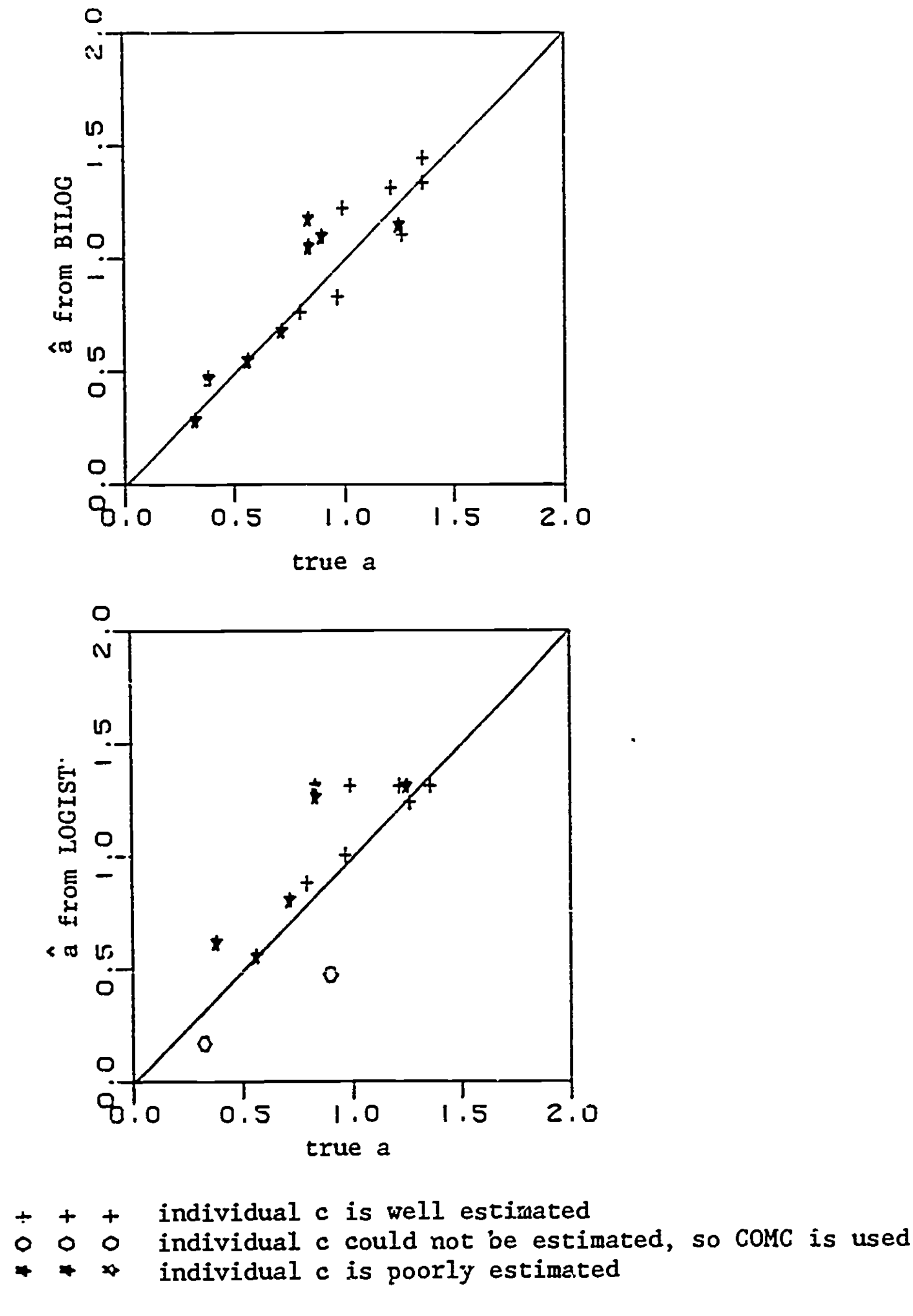
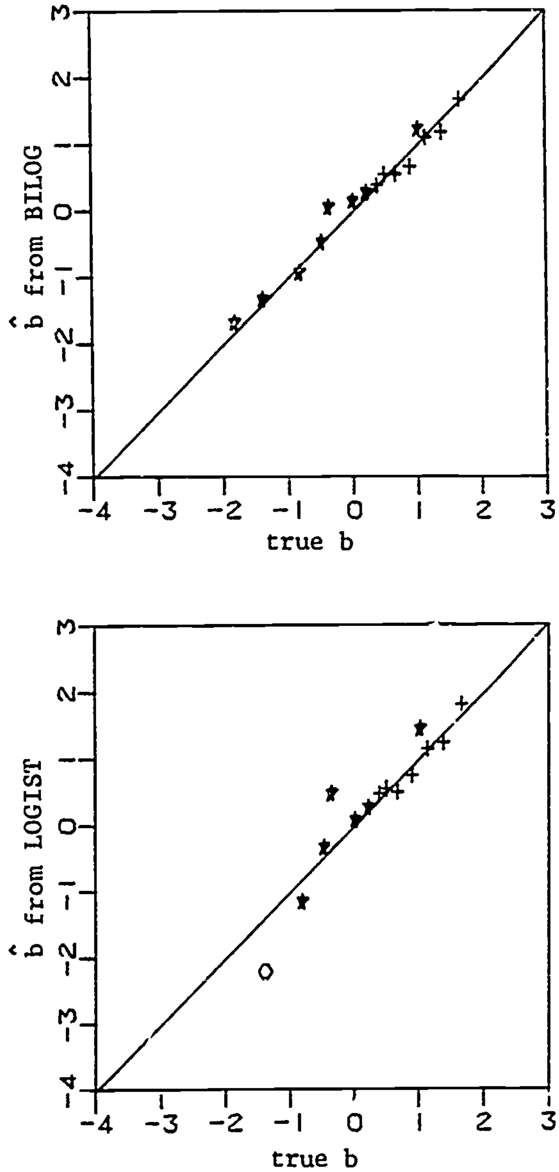


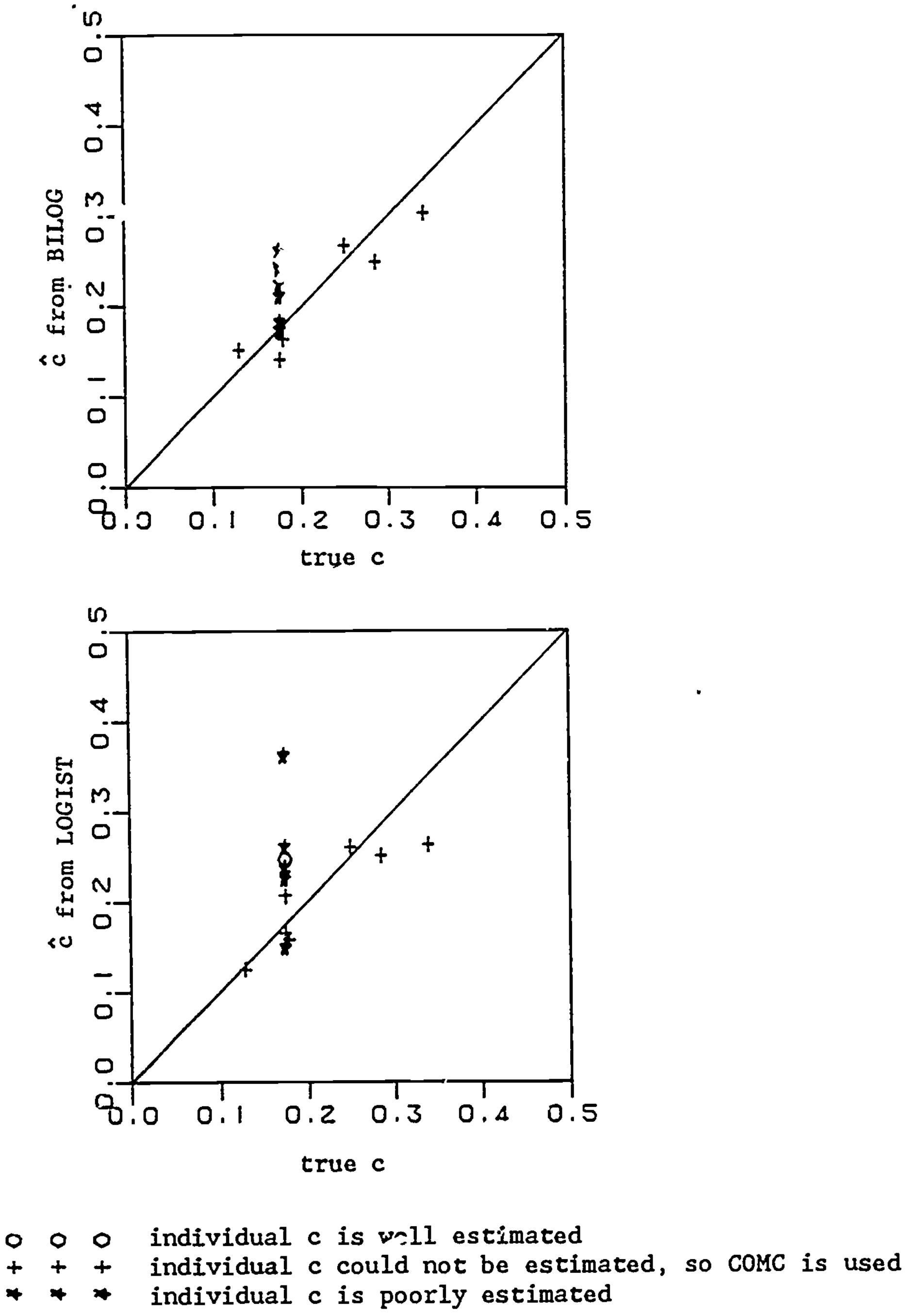
Figure 7. Estimated a's for BILOG (top) and LOGIST (bottom) compared to true a's, n = 15.



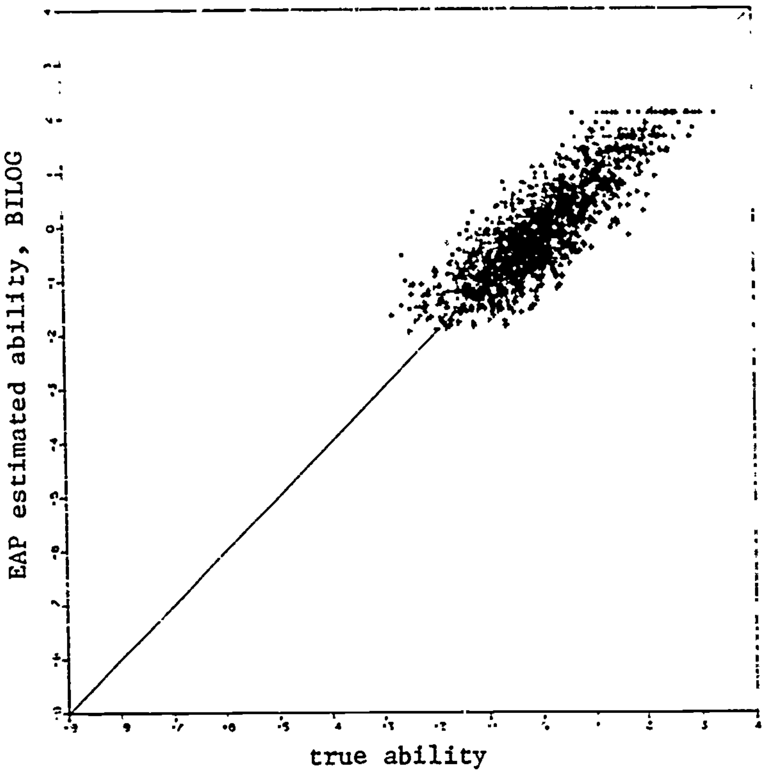
Point Not Plotted
(-1.83, -4.43)

+ + + individual c is well estimated
o o o individual c could not be estimated, so COMC is used
✶ ✶ ✶ individual c is poorly estimated

Figure 8. Estimated b's for BILOG (top) and LOGIST (bottom) compared to true b's, n = 15.



Fi; ure 9. Estimated c's for BILOG (top) and LOGIST (bottom) compared to true c's, n = 15.



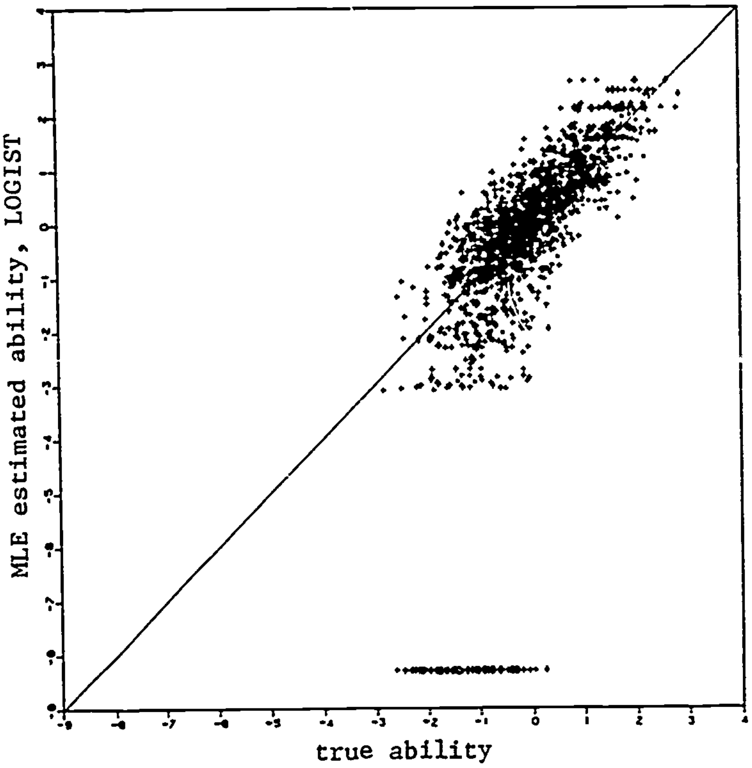


Figure 10. Estimated abilities from BILOG (top, EAP estimates) and LOGIST (bottom, MLE estimates) compared to true values, n = 15.