A COMPARISON OF THE EFFICIENCY AND ACCURACY OF BILOG AND LOGIST

WENDY M. YEN

CTB/MCGRAW-HILL

Comparisons are made between BILOG version 2.2 and LOGIST 5.0 Version 2.5 in estimating the item parameters, traits, item characteristic functions (ICFs), and test characteristic functions (TCFs) for the three-parameter logistic model. Data analyzed are simulated item responses for 1000 simulees and one 10-item test, four 20-item tests, and four 40-item tests. LOGIST usually was faster than BILOG in producing maximum likelihood estimates. BILOG almost always produced more accurate estimates of individual item parameters. In estimating ICFs and TCFs BILOG was more accurate for the 10-item test, and the two programs were about equally accurate for the 20- and 40-item tests.

Key words: BILOG, computer program, item response theory, LOGIST, tests.

The three-parameter logistic model, which appears in (1), is being used increasingly for analyzing test data:

\[ P_i(\theta_k) = c_i + \frac{1 - c_i}{1 + \exp \left( -1.7 a_i (\theta_k - b_i) \right)} \] (1)

In this model, \( P_i(\theta_k) \) is the item characteristic function (ICF) which defines the probability that examinee \( k \) with trait value \( \theta_k \) will pass item \( i \) with parameters \( a_i \) (discrimination), \( b_i \) (difficulty), and \( c_i \) (lower asymptote). The computer program LOGIST has been available since 1973 (Wingersky & Lord, 1973) to estimate the parameters and traits for this model, and LOGIST recently underwent a major revision (Wingersky, 1983; Wingersky, Barton, & Lord, 1982). LOGIST produces maximum likelihood estimates based on the procedure developed by Lord (1974). BILOG (Mislevy & Bock, 1984) is a more recently developed program that produces estimates using marginal maximum likelihood (Bock & Aitkin, 1981); this new estimation procedure offers the possibility of increased efficiency in estimation. BILOG also offers the option of Bayesian estimation procedures not available in LOGIST.

Qualls and Ansley (1985) recently compared the performance of the two programs with simulated data for sample sizes of 200, 500, and 1000 and for test lengths of 10, 20, and 30 items. They found that BILOG uniformly took more CPU seconds than LOGIST to produce estimates, but the BILOG estimates were almost uniformly more accurate than the LOGIST estimates. However, the generality of these findings is somewhat limited by several factors. The simulated tests had uniform distributions of item difficulties and discriminations, which are not typically found with real tests. Trait and parameter estimates were compared directly to true traits and parameters; the estimated and true values
were assumed to be on the same metric because the mean and variance of the estimated traits were the same as the mean and variance of the true traits. However, given that estimated traits are not provided for all simulees (particularly very high- and low-scoring simulees) and that estimated traits contain error variance not found in true traits, it is not guaranteed that the estimated and true values were actually on the same metric. Comparisons between true and estimated values were limited to correlations and average absolute differences. Finally, it was not stated in that paper which options were used in LOGIST and BILOG; for BILOG in particular the options chosen could have a substantial effect on results.

The present paper focuses on tests of lengths that might be considered in the range frequently found in psychological and educational applications (20 and 40 items); one 10-item test is examined as an extreme case. Simulated tests, some of which are very similar to real tests in terms of their parameters, are examined. A sample size of 1000 is used, and estimated parameters are placed on the true parameter scale using the Stocking and Lord (1983) procedure; because this procedure focuses on test characteristic functions (TCF), which are expectations, it avoids the problem of error variance inflating the variances of estimates.

A variety of comparisons in addition to correlations are made between estimated and true values, and variables in addition to the item parameters and traits are examined. The true probability that each simulee passes each item, \( P_i(\theta_k) \), is compared with the probability estimated by each program, \( \hat{P}_j(\theta_k) \), where the estimated item parameters and traits replace their true values in (1). This comparison is equivalent to examining the accuracy with which ICFs are estimated; such a comparison is important because a variety of combinations of parameters can produce very similar ICFs, and in many applications the estimated ICF is more important than the individual parameters (Hulin, Lissak, & Drasgow, 1982). True scores, or expected proportion-correct scores, are also examined:

\[
T(\theta_k) = \frac{1}{n} \sum_{i=1}^{n} P_i(\theta_k).
\]  

Estimated true scores are obtained by replacing the true parameters and traits in (2) by their estimated values. Examining true scores is equivalent to examining the extent to which the estimates reproduce the TCF.

Method

Simulated Data

Normal trait distributions. Simulated right/wrong item response vectors were generated using the three-parameter logistic model for 1000 simulees “taking” one 10-item test, four 20-item tests, and four 40-item tests. Details of this data generation are presented in Yen (1984). True trait values are generated from normal (0, 1) distributions. Summary descriptions of the item parameters used in generating the responses are in Table 1. For the “reading vocabulary” tests the correlations between the true parameters were \( r_{ab} = .53, r_{cb} = .08 \), and \( r_{ac} = .00 \) for the 10-item test, \( r_{ab} = .28, r_{cb} = .31 \), and \( r_{ac} = -.22 \) for the 20-item test, and \( r_{ab} = .31, r_{cb} = .30 \), and \( r_{ac} = .12 \) for the 40-item test. The mean proportion-correct scores were .75 for the Easy tests, .65 for the Moderate tests, .55 for the Difficult tests, from .67 to .69 for the 20- and 40-item reading vocabulary tests, and .62 for the 10-item reading vocabulary test.

Nonnormal trait distributions. Additional response vectors were generated for the 20- and 40-item reading vocabulary tests for 1000 simulees from three different nonnor-