INTERNATIONAL JOURNAL OF TESTING, *3*(4), 365–384

Item Response Modeling With BILOG-MG and MULTILOG for Windows

André A. Rupp
Faculty of Education
University of Ottawa

Item response theory (IRT) has become one of the most popular scoring frameworks for measurement data. IRT models are used frequently in computerized adaptive testing, cognitively diagnostic assessment, and test equating. This article reviews two of the most popular software packages for IRT model estimation, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) and MULTILOG (Thissen, 1991), which are for the first time available on a single CD-ROM with new features. Most prominently, the number of items to be calibrated and examinees to be scored is now limited only by memory capacities of the hardware, MULTILOG has an interactive Windows-oriented process for creating basic command file syntax, and both BILOG-MG and MULTILOG come with a new graphics interface that displays numerous curves relevant to IRT analyses in a professional format. This article reviews the models that are and are not estimable with these programs and describes the fundamental ideas of the underlying estimation algorithms without providing detailed derivations. Moreover, the user-friendliness of both programs is assessed with a user in mind who is interested in easy-to-use IRT estimation programs within a Windows point-and-click environment. Both programs fulfill such an expectation to a large degree; yet, this review also points out some obstacles that someone relatively unfamiliar to IRT or syntax programming might have to overcome to obtain meaningful results.

Item response theory (IRT) has become one of the most popular scoring frameworks for measurement data. IRT models are used frequently in computerized adaptive testing, cognitively diagnostic assessment, and test equating. However,

Requests for reprints should be sent to André A. Rupp, Faculty of Education, University of Ottawa, 145 Jean-Jacques Lussier St., Ottawa, Ontario, Canada K1N 5N6. E-mail: arupp@rogers.com

IRT is not so much a theory as it is an umbrella term for a variety of measurement models, ranging from basic unidimensional models for dichotomously and polytomously scored items and their multidimensional analogues to models that incorporate information about cognitive subprocesses that influence the overall item response process. Even if all available models could be collected and estimated with a single software program, the differences in model structure and required estimation routines make this prohibitive. Thus, having software programs available that estimate at least the most commonly encountered IRT models is a major practical benefit. Several programs exist for this purpose and the programs BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), MULTILOG (Thissen, 1991), PARSCALE (Muraki & Bock, 1997), and TESTFACT (Wilson, Wood, & Gibbons, 1991) have proven particularly useful and reliable over the last 15 years for many applications. These four programs are now available for the first time in a Windows format on a single CD-ROM. This article reviews the first two programs in this package, BILOG-MG and MULTILOG. For this review, I follow the guidelines proposed by Gierl and Ackerman (1996), who raise questions about a software program's quality of item and examinee parameter estimation, its flexibility and ease of use, the comprehensiveness of its output, and the adequacy and timeliness of technical support. See Kim (1997) for a detailed description on installation features and sample runs for an older version of BILOG, whose results are generally applicable to the newer version as well.

This article is organized as follows: The next section provides an overview of currently available IRT models and shows which models are estimable with BILOG-MG and MULTILOG. The following section overviews the estimation theory incorporated into these programs that should be accessible for users who are comfortable with reading basic mathematical equations without intricate derivational steps. This is followed by a review of the two programs with particular attention paid to the types of knowledge the user needs to execute a successful analysis of measurement data.

This review assumes as a basic user someone who is familiar with the advantages of IRT, has a data set with item response data, and potentially has some background information on examinees at hand that he or she would like to calibrate with either BILOG-MG or MULTILOG. This review does not presume that the user is a trained measurement theorist, because most users with such training would be familiar with simple programming tasks and might find even relatively cumbersome interfaces and complex output strings acceptable. On the contrary, the availability of a software program for Windows implicitly signals the intent of its designers to make these programs accessible to users primarily or exclusively comfortable with point-and-click interfaces, such as an applied measurement analyst who would use other Windows-driven software programs for statistical analysis.

A BRIEF OVERVIEW OF IRT MODELS

The number of IRT models available to measurement analysts has increased considerably in the last 15 years due to increasing computer power and a demand for richer and more meaningful inferences grounded in complex data structures. The developments in modeling were intertwined with developments in estimation theory, most notably Bayesian estimation with associated Markov chain Monte Carlo (MCMC) algorithms (Patz & Junker, 1999a, 1999b; Rupp, Dey, & Zumbo, in press). The popularity of the IRT framework has also entailed numerous overviews in book and journals, and many connections between IRT and other statistical estimation frameworks such as factor analysis, generalizability theory, and structural equation modeling have been made repeatedly (e.g., McDonald, 1999; Mellenbergh, 1995; Muthén, 2002; Rupp, 2002; see also van der Linden & Hambleton, 1997). The following overviews the range of IRT models that have been proposed in the literature to be able to understand better the place that BILOG-MG and MULTILOG occupy in the model estimation realm.

Before beginning, however, the reader should note that all IRT models estimable with BILOG-MG and MULTILOG are based on the three assumptions of local independence, monotonicity, and unidimensionality. The first assumption, local independence, states that the conditional probability of observing any response vector can be expressed as a product, across all items and examinees, of the probabilities of observing the individual response probabilities so that the response probabilities are independent at the local item level.

Formally, for I items, J examinees, and with X denoting the manifest response variable as well as θ denoting the latent predictor variable, $P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{i=1}^{I}\prod_{j=1}^{J} P(X_{ij} = x_{ij}|\theta)$. The second assumption, monotonicity, refers to the fact that the item characteristic curves (ICCs), which trace the response probabilities as a function of the latent variable θ, are nondecreasing functions in θ. The third assumption, unidimensionality, refers to the fact that θ is assumed to be a unidimensional random variable and not a multidimensional random vector. Note that the last two assumptions are implicit when a model is chosen whereas the first one is implicitly used in the parameter estimation process. Lack of model fit and investigations of lack of invariance (LOI) such as differential item functioning (DIF) and item parameter drift (IPD) are attributable to violations of at least one of these three assumptions.

With these notions at hand, we can now look at several IRT models available in the literature and the subsets of these that are estimable with BILOG-MG and MULTILOG.

Models for Dichotomous Data (BILOG-MG and MULTILOG)

Basic unidimensional IRT models for dichotomous scores (e.g., right–wrong, forced–choice) model, for each item separately, the log-odds of the probability of a correct response for examinees as a function of a latent variable θ and one or more item parameters. The number of item parameters and function type are used to label the models and so the one-parameter logistic (1PL) model includes one estimable item parameter, the item difficulty parameter β_j, whereas the two-parameter logistic (2PL) model contains an additional estimable item discrimination parameter α_j, the three-parameter logistic (3PL) model contains an additional estimable item lower asymptote or pseudo-guessing parameter γ_j, and the four-parameter logistic (4PL) model, even though not used in practice, contains an additional estimable upper asymptote or item ceiling parameter ζ_j. The following is the equation of the 4PL model, which allows the derivation of the other models as special cases by constraining ζ_j to 1 (3PL), γ_j to 0 (2PL), and α_j to 1 (1PL) respectively:

$$P_j(\theta_i) = \gamma_j + (\zeta_j - \gamma_j)\frac{\exp(\alpha_j(\theta_i - \beta_j))}{1 + \exp(\alpha_j(\theta_i - \beta_j))}; \qquad (1)$$

$$\alpha_j > 0,\ -\infty < \beta_j, \theta_i < \infty,\ 0 \le \gamma_j < \zeta_j \le 1$$

The functional graph for each item, the ICC, is bound between 0 and 1, sigmoidal in shape, and symmetric at its inflection point located at β_j. We thus have that for sets of items in the 1PL the ICCs all have the same slope and do not intersect, for the 2PL they are allowed to have different slopes and may intersect, for the 3PL their lower asymptotes may also be greater than 0, and for the 4PL their upper asymptotes may also be less than 1. The most flexible model is thus the 4PL, but the 3PL is more commonly used due to the large number of items and examinees required to estimate the 4PL properly and a lack of efficient routines to do so. BILOG-MG estimates the 1PL, 2PL, and 3PL, whereas MULTILOG estimates the 1PL and 2PL as special cases of the Graded Response Model (Samejima, 1969, 1997b) and the 3PL as a special case of the Multiple Response Model (Thissen & Steinberg, 1984). Note that MULTILOG requires a recoding of the responses to ‘1’ and ‘2’—in contrast to the more common ‘0’ and ‘1’—because the ‘0’ category is reserved for missing data. In the new Windows version of these programs, all of these models are now estimable for an unlimited number of examinees and items; the numbers are bound only by the memory capacity of the computer on which they are installed.

The next level of complexity for unidimensional models is to decompose individual item parameters using indicator variables or scoring weights on *L* predictor variables or so-called attributes, which are often based on cognitive theory. This can be done for the difficulty parameter β_j or for both the difficulty parameter β_j and the discrimination parameter α_j; yet, a meaningful application of these models requires that the analyst have sufficient relevant background information on the response processes, which makes them of interest mainly to substantive cognitive researchers. At any rate, neither BILOG-MG nor MULTILOG estimates these models. Similarly, the natural extension of all of the models presented so far are multidimensional IRT models, which include an *M*-dimensional vector of latent predictor variables instead of a unidimensional latent predictor variable (e.g., Ackerman, 1994; Embretson, 1984) with again a possible decompositions of item parameters (e.g., Embretson, 1999). Moreover, decomposing the item response process into subprocesses according to substantive theory is possible, which leads to componential IRT models (e.g., Hoskens & deBoeck, 1995, 2001; Samejima, 1995, 1997a). However, neither BILOG-MG nor MULTILOG estimate such models because they are programs for unidimensional IRT models only and do not allow for parameter decompositions and refined response process decomposition.

Models for Polytomously Scored Items (MULTILOG)

Many assessment instruments consist solely of polytomously scored items or a mixture of dichotomously and polytomously scored items and hence require that models for items with *K* response categories be available. These categories are typically scored 1, 2, … , *K*, yet the score could also be replaced by a more general scoring function.

MULTILOG estimates three major models, which have a variety of multicategory models as special cases. Note, however, that MULTILOG does not estimate ICCs for categories that have no observed responses and that users need to manually code the scoring key to collapse categories because the program does not automatically check for zero observed frequencies. For tests with mixed item types or when different models are desired for different polytomously scored items, MULTILOG is able to calibrate all items and score all examinees on the same scale.

The first model is suitable for item scores that represent a graded scale and is aptly called the Graded Response Model (Samejima, 1969). It uses cumulative probabilities to define the response probability for a given category *k* as the difference in cumulative response probabilities of adjacent categories. The threshold location parameters β_{jk}, $k = 0, \ldots, K\!-\!1$ indicate scoring in category *k* or *higher.* Symbolically:

$$P_j\left(X_{ij} = k \middle| \theta_i\right) = P_j^*\left(X_{ij} = k\right) - P_j^*\left(X_{ij} = k+1\right) = \frac{\exp\left[\alpha_{jk}\left(\theta_i - \beta_{jk}\right)\right]}{1+\exp\left[\alpha_{jk}\left(\theta_i - \beta_{jk}\right)\right]} - \frac{\exp\left[\alpha_{jk+1}\left(\theta_i - \beta_{jk+1}\right)\right]}{1+\exp\left[\alpha_{jk+1}\left(\theta_i - \beta_{jk+1}\right)\right]} \qquad (2)$$

As special cases of this model, one can obtain, for example, the 1PL and 2PL.

The second major model is suitable for different types of multicategory responses that do not represent a graded scale but rather alternative choices and is aptly called the Multiple Response Model (Thissen & Steinberg, 1984), which is a further modification of Samejima's (1972) modification of Bock's (1972) Nominal Model. Symbolically, in divide-by-total form (Thissen & Steinberg, 1986):

$$P_j\left(X_{ij}=k|\theta_i\right)=\frac{h^{*}\exp\left[\alpha_{jk}\theta_i-c_{jk}\right]+hd_k\exp\left[\alpha_{jk}\theta_i-c_{jk}\right]}{\sum_{s=1}^{K}\exp\left[\alpha_{js}\theta_i-c_{js}\right]} \tag{3}$$

where the parameter d_k denotes the proportion of examinees who do not know an answer but nevertheless respond in some category, the parameters h and h^* are calculated by MULTILOG and are used to obtain other models as special cases of this, and $\alpha_{jk}\theta_i - c_{jk}$ is simply an alternative parametrization of the typical 2PL kernel $\alpha(\theta - \beta)$. With the h and h^* parameters one can obtain the Multiple Choice Model ($h = 1$, $h^* = 1$), the Nominal Model ($h = 0$, $h^* = 1$), and the 3PL for dichotomous data ($h = 1$, $h^* \in \{0,1\}$) as special cases.

In general, the specification of parameter contrasts in MULTILOG allows one to develop contrasts that lead to interpretations or parameters that may be particularly meaningful in the context of a particular study or to obtain other well-known models as special cases. For example, the Nominal Model with centered polynomial contrasts for scores and constant linear contrasts across items for the α_j's is equivalent to Master's (1982) Partial Credit Model for ordered item responses. Similarly, one can obtain Muraki's (1992) Generalized Partial Credit Model and Andrich's (1978) Rating Scale Model as special cases of the Nominal Model and Muraki's (1990) rating scale version of the Graded Response Model (see Mellenbergh, 1995; Thissen & Steinberg, 1986).

The final model in MULTILOG is the Normal Linear Model for continuous responses, which can be conceived of as a limiting case for an increasing number of response categories. This model postulates a Gaussian response function with mean response level μ and standard deviation σ. With Y_{ij} denoting the continuous item variable we have, symbolically:

$$P_j\left(Y_{ij}=y_{ij}|\theta_i\right)=\left(\sqrt{2\pi\sigma_j}\right)^{-0.5}\cdot\exp[-\frac{\left(Y_{ij}-\beta_j\theta_i-\mu_j\right)^2}{2\sigma_j^2}] \tag{4}$$

The models discussed thus far cover a wide range of IRT models, but several more complex IRT models exist that these programs do not estimate, including models for preference data such as the hyperbolic cosine model (Andrich, 1995), unidimensional latent class models (e.g., Junker & Sijtsma, 2001), nonparametric models such as Mokken's models of monotone homogeneity and double monotonicity (Mokken & Lewis, 1982), and the extended forms of some of the previous models for cognitively diagnostic assessment (see Junker, 1999). For a more mathematical description of these models, see Junker (1999), Rupp (2002), and van der Linden and Hambleton (1997), as well as the citations therein. However, as stated earlier, no software program can possibly accommodate all models currently available to the practitioner in a field with myriad models, and BILOG-MG and MULTILOG are very flexible programs for estimation purposes. In particular, their ability to accommodate multilevel and multigroup data makes them so appealing for current-day applications.

Multilevel and Multigroup Data

In some situations multilevel modeling is appropriate, including scenarios where only a subset of a group of examinees is administered a follow-up assessment based on their results on a preliminary assessment, or when examinees are sampled in nested settings that lead to matrix-sampled data. These scenarios are accommodated in BILOG-MG. Moreover, when different groups of examinees are used such as in equivalent and nonequivalent groups equating and investigations of LOI such as DIF and IPD, multigroup estimation procedures are needed. Both BILOG-MG and MULTILOG accommodate data arising from these scenarios. However, as is always the case, certain limitations exist. For example, IPD and DIF analyses in BILOG-MG are restricted to the item difficulty parameter and procedures for matrix-sampled data require the administration of at least 15 randomly parallel forms each with at least one item from each content area assigned in rotation to students under identical conditions. In MULTILOG, any subset of parameters can be constrained similar to estimation procedures in LISREL (Jöreskog & Sörbom, 1996) but, depending on the model and data structure, large sample sizes are often required to obtain stable parameter estimates. Nevertheless, the convenience of having BILOG-MG and MULTILOG available in a Windows environments combined with the variety of models they estimate and the variety of data structures they accommodate make them appealing programs.

ESTIMATION OF MODEL PARAMETERS IN BILOG-MG AND MULTILOG

Writing a description of an estimation process is always a bit odd because to understand estimation theory properly a fairly advanced training in mathematical statistics is required; yet, even analysts who lack such a training should become familiar with the basic ideas. Fortunately, several excellent articles are available that describe the general estimation process in IRT models. At a more advanced level, I have found the unifying description of the expectation-maximization (EM) algorithm by Harwell, Baker, and Zwarts (1993) and its extension to Bayesian estimation (Harwell & Baker, 1991) to be most useful because it facilitates an understanding of the even more technical and denser articles on said algorithm and the

Bayesian estimation paradigm (e.g., Bock & Aitkin, 1981; Bock & Lieberman, 1970; Dempster, Laird, & Rubin, 1977; Mislevy, 1984, 1986). At a simpler level, chapter 3 of the accompanying manual for the CD-ROM reviewed here (du Toit, 2003) provides a relatively accessible overview of the estimation process in BILOG-MG without going into too many subtle details and derivations. However, the chapter does not address the estimation procedures in MULTILOG in more detail but rather references the original articles in the literature due to the more complex estimation routines.

This article focuses on the general concepts in the estimation procedures of both BILOG-MG and MULTILOG, which are applicable, conceptually, to most modern estimation enterprises. At this point it is worth remarking on what constitutes a sufficiently large sample size for item calibration. This depends of course first and foremost on the model because different models require different numbers of parameters to be estimated per item (e.g., J parameters for a 1PL, $2 \times J$ parameters for a 2PL, and $3 \times J$ parameters for a 3PL) and also some of these parameters are harder to estimate than others due to a lack of information in the data about them (e.g., the lower-asymptote parameter in a 3PL). Specifically, the degree of bias and estimation error for parameter estimates depends on factors such as the number of parameters whose true values are extreme, the degree of skewness and kurtosis of the true underlying examinee parameter distribution, the match of the prior distribution to this underlying distribution, and the variance of the prior distribution. As might be expected, the influence of these factors decreases as the number of examinees increases for a fixed number of items. If any general guidelines can be given, it appears that for tests with between 15 and 50 items, approximately 250 examinees are required for the 1PL and 2PL, and approximately 500, maybe even 1,000, examinees are required for the 3PL and the Graded Response Model to achieve stable parameter estimates (e.g., Harwell & Janosky, 1991; Reise & Yu, 1990; Seong, 1990; Stone, 1992; Yen, 1987; see also Drasgow, 1989; Kirisci, Hsu, & Yu, 2001).

The following description outlines some fundamental estimation steps and comprises two sections. First, some basic equations that illustrate common concepts in estimation theory, such as marginalization and likelihood, are introduced. Second, these concepts are used to describe the major estimation process in BILOG-MG and MULTILOG, which is marginal maximum likelihood (MML) estimation within a fully Bayesian framework.

Basic Equations for Estimation

In the following, I will again denote examinees by $i = 1, \ldots, I$, items by $j = 1, \ldots, J$, the latent predictor variable by θ, and the response probability for a correct response to a given item or item category by P. The following description of parameter estimation is not tied to a single IRT model, but instead is applicable to all basic unidimensional models. Consequently, I use the general expression $P_j(X_{ij} = x_{ij} \mid \theta_i)$ to denote such an IRT model. For example, for the 3PL:

$$P_j\left(X_{ij} = x_{ij} \middle| \theta_i\right) = \gamma_j + \left(1 - \gamma_j\right) \frac{\exp\left[\alpha_j\left(\theta_i - \beta_j\right)\right]}{1 + \exp\left[\alpha_j\left(\theta_i - \beta_j\right)\right]}. \qquad (5)$$

Under the assumption of local independence, the probability of observing a response vector x_i for a given examinee is:

$$P\left(\mathbf{X}_i = \mathbf{x}_i \middle| \theta_i\right) = \prod_{j=1}^{J} P_j\left(X_{ij} = x_{ij} \middle| \theta_i\right). \qquad (6)$$

This is a conditional probability because the response probability depends on the latent variable θ. Given that the responses of the examinees are independent of one another, the conditional probability of observing all response patterns (i.e., the data), can thus be computed as the double product:

$$P\left(\mathbf{X} = \mathbf{x} \middle| \theta\right) = \prod_{i=1}^{I} \prod_{j=1}^{J} P_j\left(X_{ij} = x_{ij} \middle| \theta_i\right). \qquad (7)$$

This probability, if thought of as a function for the unknown parameter vector, is also known as the likelihood for the data, $L(\theta \mid \mathbf{X} = \mathbf{x})$. If one conceives of θ as a random variable (i.e., of the examinees being randomly drawn from a population with a distribution for θ), one can further integrate out θ. The distinction between conceiving of θ as either a random variable or not is analogous to analysis of variance models where one distinguishes between random effects and fixed effects. Indeed, MULTILOG also calibrates item parameters for the fixed effects θ case using maximum likelihood (ML) estimation, but this case is rarer in practice.

Under the assumption of θ as a random effect, one thus obtains the unconditional or marginal probability of observing the data:

$$P(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^{I} \left\{ \int_{\Theta} \prod_{j=1}^{J} P_j\left(X_{ij} = x_{ij} \middle| \theta_i\right) g(\theta) d\theta \right\}. \qquad (8)$$

This function is also known as the marginal likelihood, $L(\mathbf{X} = \mathbf{x})$. Here, $g(\theta)$ denotes the probability distribution of θ in the population, which is often assumed to be standard normal for estimation purposes (e.g., if $\theta \sim N(0,1)$ the domain for integration, Θ, becomes the real numbers).

For practical estimation purposes, the distribution g(θ) needs to be estimated, just as is the case with any estimation of an integral in numerical analysis, and several techniques are available for that purpose. They all have in common that a suitable subset of the real numbers is chosen (e.g., the interval from –5 to 5 if θ ~ N(0,1)), along with a number of K evaluation points, which are typically equally spaced. For each evaluation point, the approximate value of the density function needs to be computed, which can be done for a theoretically selected distribution (e.g., a standard normal distribution) or for an empirically estimated distribution (i.e., one that is estimated from the data). For a general normal distribution density weights for evaluation points are available in the literature (Stroud & Secrest, 1966) and can be easily modified to obtain the requisite values for the standard normal distribution.

If we now denote the *k*th density weight by $A(T_{\mathrm{k}})$, we can write equation (8) as:

$$\tilde{P}(\mathbf{X}=\mathbf{x}) = \prod_{i=1}^{I}\left[\sum_{k=1}^{K}\left\{\prod_{j=1}^{J} P_j\left(X_{ij}=x_{ij}\,|T_k\right)\right\}A\left(T_k\right)\right]. \qquad (9)$$

Note that the letter *T* is used to denote that the unobservable θ value has been replaced with the observable (i.e., determined) evaluation point value *T*. We now have the basic machinery to describe how the previous equations are used in the MML parameter estimation procedure implemented in BILOG-MG.

MML Estimation

The parameters of interest in a testing situation are typically both the item and examinee parameters. The former are always conceived of as fixed whereas the latter can either be conceived of as random or fixed as stated earlier. If the examinee parameters are conceived of as random, the interest is in estimating the parameters of their distribution g(θ) (e.g., the mean and variance of a normal distribution); if they are conceived of as fixed, the interest is in estimating the *I* values for the *I* examinees who took the test.

Historically, item parameters and examinee parameters were either estimated jointly in an iterative fashion using a procedure called joint maximum likelihood (JML), or, for models where a sufficient statistic for θ was available—such as the 1PL—using a procedure called conditional maximum likelihood (CML). However, the parameter estimates from JML do not have desirable properties in most cases (i.e., they are not consistent), and CML is useful only for the limited class of models that have sufficient statistics for θ that are directly observable from the data.

Thus, the basic idea in MML is to overcome the iterative dependency of previous parameter estimates in JML by first integrating out θ (see Equations 8 and 9) and then to maximize the marginal log-likelihood (Log-$\mathrm{L_M}$) to obtain the MML estimates of the item parameters using the first- and second-order derivatives. The item parameter estimates can then be used to obtain estimates of the examinee parameters if needed. Examinees are commonly grouped by observed response patterns and the observed frequencies of each response pattern are used to reduce the number of computations, but that step is not reproduced in this exposition to preserve notational clarity.

Recalling the form of the marginal likelihood function from Equation 8 and its estimate from Equation 9, the estimated Log-L_M using the density weights A(T_k) looks like this:

$$\log-\mathrm{L}_\mathrm{M} \cong \sum_{i=1}^{I} \log\left[\sum_{k=1}^{K}\left\{\prod_{j=1}^{J} P_j\left(X_{ij}=x_{ij}\middle|T_k\right)\right\}A\left(T_k\right)\right]. \tag{10}$$

In BILOG-MG and MULTILOG, initial theoretical (i.e., prior) density weights are chosen to start the estimation routine, which then become adjusted or replaced at each iteration cycle by empirically estimated (i.e., posterior) weights (Mislevy, 1984). The practical maximization process of Equation 10 is done via a modification of a missing-data algorithm, the EM algorithm (Bock & Aitkin, 1981; Bock & Lieberman, 1970; Dempster et al., 1977; Harwell & Baker, 1991; Harwell, Baker, & Zwarts, 1988). This algorithm uses the expected number of examinees at each evaluation point, $\bar{n}_{jk}$, and the expected number of correct responses at each evaluation point, $\bar{r}_{jk}$, as artificial data and then maximizes the Log-L_M at each iteration.

Moreover, the EM algorithm employs Bayes theorem. In general, the theorem expresses the posterior probability of an event, after observing the data, as a function of the likelihood for the data and the prior probability of the event, before observing the data. To implement the EM algorithm, the kernel of Log-L_M is expressed with respect to the posterior probability for θ, which is:

$$P_i\left(\theta_i\middle|\mathbf{X_i}=\mathbf{x_i}\right)=\frac{L\left(\mathbf{X_i}=\mathbf{x_i}\middle|\theta_i\right)g(\theta)}{\int_{\Theta} L\left(\mathbf{X_i}=\mathbf{x_i}\middle|\theta_i\right)g(\theta)d\theta}. \tag{11}$$

Here, g(θ) represents our prior beliefs about the population distribution of θ. Using this theorem, the MML estimation process within the EM algorithm comprises three steps, which are repeated until convergence of the item parameter estimates is achieved.

First, the posterior probability of θ for each examinee *i* at each evaluation point *k* is computed via

$$P_{ik}\left(T_k\middle|\mathbf{X_i}\right)=\frac{\left\{\prod_{j=1}^{J} P_j\left(X_{ij}=x_{ij}\middle|T_k\right)\right\}A\left(T_k\right)}{\sum_{s=1}^{K}\left\{\prod_{j=1}^{J} P_j\left(X_{ij}=x_{ij}\middle|T_s\right)\right\}A\left(T_s\right)} \tag{12}$$

as an approximation to Equation 11 at evaluation point *k*. This is accomplished by using provisional item parameter estimates from the previous iteration to compute $\mathrm{P}_\mathrm{j}(\mathrm{X}_\mathrm{ij} = \mathrm{x}_\mathrm{ij} \mid \mathrm{T}_\mathrm{k})$ for a chosen model. Second, using these posterior probabilities, the artificial data for each item *j* at each evaluation point *k* are generated using

$$\bar{n}_{jk} = \sum_{i=1}^{I} P_{ik}\left(T_k \middle| \mathbf{X_i}\right)$$
$$\bar{r}_{jk} = \sum_{i=1}^{I} X_{ij} P_{ik}\left(T_k \middle| \mathbf{X_i}\right). \tag{13}$$

Third, the first-order derivatives of the estimated Log-L$_M$ function Equation 10 with respect to the item parameters are set to 0 to determine their maxima (i.e., to determine the item parameter values) and the information matrix at their estimates using the Newton-Gauss (i.e., Fisher scoring) algorithm. For that purpose, Equation 10 and its derivatives are rewritten using the artificial data in Equation 13. The entire process is then repeated until convergence of parameter estimates has been achieved.

Instead of just performing these steps, however, BILOG-MG and MULTILOG allow for a fully Bayesian estimation process. Hence, prior distributions can be specified for item and examinee distribution parameters as well, which are then incorporated into the Log-L$_M$ and its derivatives. For the 3PL for example, it is reasonable to specify normal priors for α_j, β_j, and the logit of γ_j (e.g., α_j ~ N(1,1), β_j ~ N(0,3), logit[γ_j] ~ N(–1.4,1)) or, alternatively, a joint multivariate normal distribution with hyperpriors for all three (see e.g., Mislevy, 1986). Both BILOG-MG and MULTILOG allow the specification of normal priors for item parameters, and BILOG-MG allows for the estimation of empirically updated priors at each cycle. The choice of priors does not have a strong influence on the item and examinee parameter estimates if the sample size is large enough to provide sufficient information about the parameters from the data, but the parameter estimates tend to drift toward the modal or mean values of the priors if the sample size is small. In such a case, if a prior is chosen that does not reflect the true behavior of the item parameter, then the estimation results become unduly biased and may also lead to seemingly erroneous results for fit statistics. For this reason, many practitioners go with the default values provided in both BILOG-MG and MULTILOG because they provide reasonable estimates that work sufficiently well over a wide range of applications (see Gifford & Swaminathan, 1990; Harwell & Janosky, 1991; Seong, 1990).

Table 1 shows several default options for the estimation process in both programs. Note that a row is included for the acceleration parameter, which Ramsay (1975) introduced for faster convergence.

Both programs allow for overriding the starting values, for the equating of item parameters and examinee distribution parameters across groups, and allow all of these to be done for all items, tests, and groups or for selected subsets. Whereas BILOG-MG allows for these things to be specified via interactive menus, MULTILOG requires them to be specified through syntax. Moreover, BILOG-MG also allows for differential weighting of cases due to allocation sampling of respondents and even allows for the specification of different random number seeds.

TABLE 1
Default Values for Marginal Maximum Likelihood

Estimation Aspect	*Default BILOG-MG*	*Default MULTILOG*
Maximum number of iterations for entire EM cycle	10	25
Maximum number of iterations for M-step in EM cycle	2	4 (times number of parameters)
Convergence threshold for M-step in EM cycle	.01	.001
Convergence threshold for entire EM cycle	.01	.0001
Acceleration parameter	1 (single group) 0.5 (multiple group)	0
Number of evaluation points for g(θ)	10 (1 group or < 50 items) 20 (> 1 group or > 50 items)	19 in [–5, 5]

Note. EM = expectation-maximization.

Scoring Examinees

Obtaining examinee parameters (i.e., scoring examinees) presumes that the chosen model not only provides stable parameter estimates, but also that it fits the data well as measured by some fit index (e.g., the likelihood-ratio test using the G^2 statistic). The scoring itself can be done via ML estimation in a frequentist framework or with either expected a posteriori (EAP) or maximum a posteriori (MAP) estimation in a fully Bayesian framework employing prior distributions. All three approaches consider the item parameter estimates that were obtained by the steps described previously as the true values; alternatively, one can of course import known item values for scoring.

Under ML estimation, the log-likelihood for each examinee or examinee group is maximized with respect to each θ yielding an ML estimate for θ for each examinee or examinee set. Under EAP, the means of the posterior distributions for θ (see Equations 11 and 12) are used as the θ estimates for each examinee or examinee group whereas in MAP, the modes of the posterior distributions for θ are used. Of course, the posterior distributions have to be empirically estimated using numerical integration with evaluation points and density weights as before. Finally, one should note that after all point estimates for θ have been obtained, their precision needs to be determined as well. This amounts to either inverting the estimated information matrices at the point estimates (for the ML approach), inverting the estimated posterior information matrices (for the MAP approach), or computing the standard deviation of the posterior distributions directly (for the EAP approach). BILOG-MG and MULTILOG allow users to choose among these three methods via interactive menus and syntax commands respectively with EAP being the default.

Similar ideas about estimation such as MML estimation in a Bayesian framework prevail for estimation of models for polytomously scored items; hence, the description presented here should prove useful for those interested in delving into the literature. The remainder of this article discusses the input procedures, output files, and interface of BILOG-MG and MULTILOG.

ANALYZING DATA WITH BILOG-MG AND MULTILOG

No software program frees its users from needing basic knowledge about the routines and models it uses; therefore, users of BILOG-MG and MULTILOG should familiarize themselves with the models they are estimating. The following program descriptions assume that users have some basic knowledge of IRT models (e.g., that users know a 3PL model has three parameters and know how they are interpreted), but that they are not familiar with FORTRAN, the code on which BILOG-MG and MULTILOG are based.

Input

Visually, the first thing to note about both programs is that they are built around point-and-click and click-and-drag interfaces, which significantly facilitates their use in a Windows environment for users who are not entirely comfortable with command syntax. Users will immediately recognize the typical Windows pull-down menus and icons with their basic choices for file management, editing, viewing, help, and numerous other choices once syntax command files are opened or created. To prepare a calibration or scoring process, either a new command file needs to be created or an existing one needs to be loaded. The latter can be easily achieved, but both programs do not automatically direct the user to the appropriate directory on the hard drive. Once the files are opened, however, new pull-down menus are activated and users can start to build a command file for the data. For users preferring command line syntax, both programs allow the creation of such files without going through any interactive Windows menus. In MULTILOG, once a complete syntax file is opened, the user has two options: run the file or look at generated output. Multiple other options for building a basic command file through a sequence of interactive Windows are activated only when a new one is created. In BILOG-MG, all program options for file management, data setup, estimation, and other features are directly accessible through menus, which is much more user-friendly than MULTILOG, which contains interactive menus only for certain subtasks.

In both BILOG-MG and MULTILOG, users can interactively select the assessment characteristics such as the number of items, tests, examinee groups, or distinct response patterns. In MULTILOG, a dialogue box opens that allows users to select radio buttons for the model they wants to use, and they can specify either the same model for all items or different models for different items via point-and-click selection. This process may be a bit tedious for very long tests when multiple model combinations for different items need to be tried out, but the program is generally an extremely user-friendly setup. Being able to go back and forth among the windows in this program during the process of creating the command file to change previous choices is helpful. Once all is done, MULTILOG automatically generates the syntax code for running. In BILOG-MG, similar options facilitate the creation of command syntax through the setup option; the major difference is that options are presented in Windows with separate tabs rather than as a sequence of separate windows.

Despite the fact that both programs facilitate the creation of basic command file syntax through interactive menus considerably, users must be aware that the goal is still to assemble an internally logical syntax file. Particularly in MULTILOG, this requires that FORTRAN code needs to be input at some points. For example, if users would like to input item response vectors for MML item calibration, they need to specify the appropriate data format, which is a rather cryptic code such as (2X, 2A1, T1, 1X, 2A1). Because the code has to match the data structure exactly, this process requires a bit of practice if the data are stored in a separate file and are not an integral part of the command syntax. Even though I can understand the structure of these codes with several examples, I have not found the MULTILOG manual to be very clear and helpful in this regard. In particular, the examples in MULTILOG highlight how the data file layout changes for different applications, so for new users who prefer a didactic method, the process of learning through multiple examples requires some patience.

Most importantly, the interactive windows-driven menu in MULTILOG only assembles a basic command file. If additional specifications such as prior distributions for item parameters or specifying parameter contrasts to obtain alternative models are desired, they have to be manually inserted into the basic command file after its completion. For guidance with this task, the layout of chapter 5 of the manual is rather confusing because even though it describes each optional command in an ordered sequence and contains cross-references, the lack of indentations and larger syntax segments that show placements of sequences of command lines made assembling a complete syntax file difficult. Certainly, examples are provided from which one can pick and choose syntax commands, but if one is unfamiliar with the programming code, familiarizing oneself with the program syntax takes time. In BILOG-MG, however, the interactive creation of syntax command is much more complete and even more advanced choices can be made via point-and-click environments.

Both BILOG-MG and MULTILOG allow the user to influence the calibration and scoring process. In MULTILOG, for example, the user may treat examinee parameters as fixed or random, a choice selected through an interactive menu, or specify prior distributions, number of iteration cycles for the EM algorithm, and starting values via command syntax. In BILOG-MG, similar choices can be made directly through the interactive menus, which also include differential weighting of cases as well as test- and group-specific calibration and scoring specifications. The calibration and scoring took only a few seconds for all of the data sets that were included in the programs on a Pentium II 400 MHz processor with 128 MB. Even if a calibration were to take a few minutes for a large data file, this would hardly pose problems for a typical user of these programs who is not doing extensive simulation work.

Output

Output in BILOG-MG and MULTILOG can be inspected in two ways. Both programs provide extensive text output in Windows, which is both helpful for experienced users and a hinderance for new users for whom digging through such output might be cumbersome and confusing. In this regard, I found the bookmark function in BILOG-MG to be most helpful because it allowed me to quickly mark important output segments and jump among them, although I also got used to the output structure quickly. Moreover, BILOG-MG separates the Windows-accessible output into three separate files with respective windows corresponding to general calibration details, item parameter estimation details, and examinee parameter estimation details, which is helpful. Standard output in both programs includes observed and estimated response frequencies, classical item statistics, estimated item parameters, posterior item and test information functions, goodness-of-fit information, and scale scores for participants with associated standard errors. In addition, BILOG-MG reports things such as detailed information on estimation cycles, IPD and DIF statistics, assumed prior distributions of item and examinee parameters, posterior distributions for different groups, correlations among subtest scores, and rescaling information, whereas MULTILOG reports things such as the specified contrast coefficients and their associated standard errors. Both programs had warning statements in their output that certain statistics are not valid if samples sizes (e.g., overall examinee sample size or number of responses for a given category) are small.

Apart from the textual output, both programs come with a new graphics module, which is an interface for displaying IRT graphics in a professional manner. The module displays ICCs (both for individual items and in a multiple-item matrix), item information curves, and test information curves with standard error curves, as well as histograms of the estimated latent distributions. All graphic displays can be interactively manipulated and easily exported into other programs such as Microsoft Word™ or Microsoft PowerPoint™ for presentation purposes. Because this interface is compatible with both BILOG-MG and MULTILOG, it is even more versatile for someone who frequently relies on both programs; I would consider the new graphics interface to be one of the most impressive and useful features of these programs for practitioners.

Help

The information in the CD-ROM manual is a rearranged reproduction of the online help information from both programs and as such is extremely comprehensive, which again can be both a help and a hindrance. The manual is well structured, but would benefit from more clearly formatted tables that collect default values and options, both of which are hidden among myriad other pieces of information in both programs. I have also found that the indexing function in BILOG-MG more frequently provided me with the exact answers to my questions and had categories divided into more useful subcategories. For example, finding detailed information about the type of output provided in BILOG was easy, but it was not easily accessible in MULTILOG. Nevertheless, both programs provide examples, explanations, and screen shots that help users to familiarize themselves with their functionalities. In addition, I have received prompt, appropriate, and courteous service from Scientific Software International when I had questions regarding workings of the software or technical difficulties.

SUMMARY

At the beginning of the 21st century, IRT models have definitely become the models of choice for many psychometric data analysts and BILOG-MG and MULTILOG continue to introduce IRT models to a wider audience. The improved capabilities of both programs, the new Windows interface for MULTILOG, and the graphics interface for both programs encourage users to explore the advantages IRT modeling has to offer. For users who prefer point-and-click Windows interfaces, BILOG-MG should be extremely easy to maneuver for both basic and advanced analyses whereas MULTILOG still requires the user to learn some syntax commands for more advanced analyses. The output both programs provide is comprehensive and can now be professionally presented due to the new graphics interface.

Both programs flexibly estimate a wide variety of IRT models, but of course not all available models. Most notably, multidimensional IRT models, componential IRT models, and models for cognitively diagnostic assessment are not estimable with these programs. On the other hand, both programs can handle data from multiple groups and perform LOI analyses; BILOG-MG can even handle matrix-sampled multilevel data. Despite these exciting possibilities, these programs still require responsible users who have a solid understanding of IRT models and that is, as always, good knowledge to have.

ACKNOWLEDGMENT

I thank Leo Stam at Scientific Software International not only for providing a review copy, but also for providing timely, comprehensive, and courteous support.

REFERENCES

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 7,* 255–278.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement, 19,* 269–290.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 40,* 443–459.

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

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, 39, Series B,* 1–38.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13,* 77–90.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Lincolnwood, IL: Scientific Software International.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175–186.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64,* 407–433.

Gierl, M. J., & Ackerman, T. (1996). XCALIBRE™ Marginal Maximum-likelihood estimation program, Windows™ Version 1.10 [Software review]. *Applied Psychological Measurement, 20,* 303–307.

Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters in item response theory models. *Applied Psychological Measurement, 14,* 33–43.

Harwell, M. R., & Baker, F. B. (1991). The use of prior distributions in marginalized Bayesian item parameter estimation: A didactic. *Applied Psychological Measurement, 15,* 375–389.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics, 13,* 243–271.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15,* 279–291.

Hoskens, M., & de Boeck, P. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement, 32,* 364–384.

Hoskens, M., & deBoeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement, 25,* 19–37.

Jöreskog, K. G., & Sörbom, D. (1996). LISREL 8: User's reference guide. Chicago: Scientific Software International.

Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Retrieved August 28, 2003, from http://www.stat.cmu.edu/~brian/nrc/cfa

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258–272.

Kim, S.-H. (1997). BILOG 3 for Windows: Item analysis and test scoring with binary logistic models (software review). *Applied Psychological Measurement, 21,* 371–376.

Kirisci, L., Hsu, T.-C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25,* 146–162.

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

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19,* 91–100.

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.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417–430.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59–71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM alogorithm. *Applied Psychological Measurement, 16,* 159–176.

Muraki, E., & Bock, R. D. (1997). PARSCALE 3: IRT based test scoring and item analysis for graded items and rating scales (computer software). Chicago: Scientific Software International.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 20,* 81–117.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146–178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342–366.

Ramsay, J. O. (1975). Solving implicit equations in psychometric data analysis. *Psychometrika, 40,* 337–360.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the Graded Response Model using MULTILOG. *Journal of Educational Measurement, 27,* 133–144.

Rupp, A. A. (2002). Feature selection for choosing and assembling measurement models: A building-block based organization. *International Journal of Testing, 2,* 311–360.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (in press). To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to item response modeling. *Structural Equation Modeling.*

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded responses. *Psychometrika Monograph Supplement, No. 17.*

Samejima, F. (1972). A general model for free response data. *Psychometrika Monograph Supplement, No. 18.*

Samejima, F. (1995). Acceleration model in the heterogeneous case of the general Graded Response Model. *Psychometrika, 60, 549–572.*

Samejima, F. (1997a). Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. *Psychometrika, 62,* 471–493.

Samejima. F. (1997b). Graded Response Model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer-Verlag.

Seong, T.-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14,* 299–311.

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliffs, NJ: Prentice Hall.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16,* 1–16.

Thissen, D. (1991). MULTILOG: Multiple category item analysis and test scoring using item response theory [Computer software]. Chicago: Scientific Software International.

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

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response theory models. *Psychometrika, 51,* 567–577.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). TESTFACT: Test scoring, item statistics, and item factor analysis [Computer software]. Chicago: Scientific Software International.

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

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago: Scientific Software International.