Methodology Review: Item Parameter Estimation Under the One-, Two-, and Three-Parameter Logistic Models

Frank B. Baker
University of Wisconsin

This paper surveys the techniques used in item response theory to estimate the parameters of the item characteristic curves fitted to item response data. The major focus is on the joint maximum likelihood estimation (JMLE) procedure, but alternative approaches are also examined. The literature shows that both the theoretical asymptotic properties and the empirical properties of the JMLE results are well-established. Although alternative approaches are available, such as Bayesian estimation and marginal maximum likelihood estimation, they do not appear to have an overwhelming advantage over the JMLE procedure. However, the properties of these alternative techniques have not been thoroughly studied as yet. It is also clear that the properties of the item parameter estimation techniques are inextricably intertwined with the computer programs used to implement them.

The use of item response theory (IRT) in applied settings depends upon the implementation of procedures for estimating the parameters of the items in an instrument and of the examinees taking the instrument. Because three related mathematical models for the item characteristic curve (ICC) are used in IRT, attention has focused upon the techniques for estimation of the item parameters under these models. Although models exist for graded and nominal response to items, the present paper will restrict its attention to the estimation of parameters of items that are dichotomously scored (i.e., binary items). Within this context, two areas are of interest: (1) the paradigms used to obtain the item parameter estimates, and (2) the theoretical and obtained characteristics of the estimates yielded by these paradigms.

This paper is divided into four sections. The first deals with the joint maximum likelihood paradigm for estimating item parameters under the one-, two-, and three-parameter ICC models. The second section compares and evaluates the three models with respect to the characteristics of the item parameter estimates yielded by this method. The third part examines estimation techniques that are alternatives to joint maximum likelihood. The fourth provides an evaluation of the state of the art. The overall goal of the paper is to provide a unified presentation of item parameter estimation under IRT for the three logistic models.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 2, June 1987, pp. 111–141

0146-6216/87/020111-31$2.80

Item Parameter Estimation Using the Joint Maximum Likelihood Procedure

The Likelihood Function

Under typical conditions, N examinees possessing a latent trait (e.g., ability) are tested and no assumption is necessary as to the distribution of the examinees over the trait continuum (Lord & Novick, 1968). Each of these examinees responds to the n items of the test and the responses are dichotomously scored, $u_{ij} = 0, 1$, where i ($i = 1, 2, \ldots, n$) designates the item and j ($j = 1, 2, \ldots, N$) designates the examinee. For each examinee there will be a vector of item responses of length n denoted by $(u_{1j}, u_{2j}, \ldots, u_{nj}|\theta_j)$. Under the local independence assumption, the u_{ij} are statistically independent for all examinees having the same trait level. There will be one such vector for each examinee, hence there will be N such vectors. The resulting $n \times N$ matrix of item responses is denoted by $\mathbb{U} = \|u_{ij}\|$. If Θ is the vector of the N examinee trait scores $(\theta_1, \theta_2, \ldots, \theta_N)$, the probability of $\mathbb{U}$ is given by the likelihood function

$$P(\mathbb{U}|\Theta) = \prod_{j=1}^{N} \prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} Q_i(\theta_j)^{1-u_{ij}} \quad . \tag{1}$$

Although this form of the likelihood function is the one most commonly reported in the literature, it is not convenient for computing purposes because the probability of correct or keyed response to an item must be computed for each examinee.

A more economical approach is to arrange the examinees into k groups identified by trait scores θ_j ($j = 1, 2, \ldots, k$). Of the f_j examinees having trait level θ_j, r_{ij} answered item i with the correct or keyed response and $f_j - r_{ij}$ answered with the incorrect (or non-keyed) response. Let $\mathbb{R} = \|r_{ij}\|$ be the $n \times k$ matrix of the observed number of correct/keyed responses. The probability of observing matrix $\mathbb{R}$ is given by

$$P(\mathbb{R}) = \prod_{j=1}^{k} \prod_{i=1}^{n} P_i(\theta_j)^{r_{ij}} Q_i(\theta_j)^{f_j - r_{ij}} \tag{2}$$

and the log likelihood is

$$\mathrm{L} = \log [P(\mathbb{R})] = \text{constant} + \sum_{j=1}^{k} \sum_{i=1}^{n} [r_{ij} \log P_i(\theta_j) + (f_j - r_{ij}) \log Q_i(\theta_j)] \quad . \tag{3}$$

The $P_i(\theta_j)$ are now functions of the h parameters of the ICCs and the θ scale. Because the h parameters for each of the n items and the θ parameters of each of the N examinees are unknown, they must be simultaneously estimated from the item response data. The estimation paradigm generally employed was given by Birnbaum (1968) and is the basis for many of the IRT computer programs in general use.

Birnbaum proposed using an iterative two-stage "back and forth" procedure for the joint maximum likelihood estimation (JMLE) of these parameters. Sanathanan (1974) and Mislevy and Bock (1984) referred to this JMLE approach as the "fixed effects" solution because from an analysis-of-variance perspective, the item parameters and θ parameters are considered fixed. Because the measurement scale of the parameter estimates yielded by the method is invariant only up to a linear transformation, the origin and unit of the scale must be fixed. In recent years this has been called the "identification problem." Although many different schemes could be used, the LOGIST computer program (Wingersky, Barton, & Lord, 1982) solves the identification problem by setting $\bar{\hat{\theta}} = 0$ to fix the origin and $\sigma^2_{\hat{\theta}} = 1$ to fix the unit of measurement. Consequently, in the two- and three-parameter models there are only $hn + N - 2$ free parameters to be estimated.

The JMLE procedure involves maximizing a likelihood function from three different perspectives: (1) for a single item, (2) for a single examinee, and (3) for the overall based upon the n items and N examinees. It is assumed that the JMLE procedure will maximize the overall likelihood function by means

of the parameter estimates yielded by the separate MLE procedures for the items and examinees. The intent is to obtain a global maximum for the overall likelihood function. Haberman (1977) has shown that this holds for the one-parameter logistic model; Lord (1980, p. 181) conjectured that such a proof exists for the three-parameter model.

Fischer (1981) examined the necessary and sufficient conditions for the JMLE procedure to converge to a unique solution for the Rasch model. He found that for this to occur, the data matrix $\mathbf{U}$ must be well-conditioned. A technique based upon the marginal sums was derived for the purpose of determining whether a data matrix is well-conditioned. Fischer suggested that although such a check can be useful, when the numbers of items and examinees are reasonable the data matrix is usually well-conditioned. He concluded that the conditions for the uniqueness of the JMLE solution under the Rasch model were rather mild. Samejima (1973) has shown that for small numbers of items (2 or 3), the likelihood function based upon a three-parameter model may not have a unique maximum. However, Lord (1980) indicated that non-uniqueness of the maximum should not be a problem when $n \geq 20$. There does not seem to be any empirical evidence to indicate that the JMLE procedure finds a local rather than a global maximum of the log likelihood function.

When the two- and three-parameter models are employed in the JMLE procedure, a phenomenon occurs in certain datasets: Discrimination estimates for one or more items can become very large, which in turn results in large values of θ estimates for examinees answering those items correctly. Then a feedback loop develops across the stages and the discrimination estimates approach infinity. Wingersky (1983) indicated that in the LOGIST program upper limits on the values of the discrimination estimates must be imposed to handle such datasets. The problem does not occur under a one-parameter model, as all of the discrimination indices are fixed at unity. Much of Wright's (1977) criticism of the two- and three-parameter models is based upon this phenomenon. He stated that it always happens under the models and therefore it is impossible to estimate the discrimination parameter. However, there is no evidence that it occurs in all datasets.

Recently, it has been recognized that this problem is the Heywood case of factor analysis (Swaminathan & Gifford, 1985). Mislevy (1986a) indicated that, under the common factor model, the Heywood case occurs when one or more unique variances takes a value of 0. He indicated that under maximum likelihood factor analysis, the unique variances do not appear as parameters to be estimated, hence their values are implied through the values of the discrimination indices. As a result, the Heywood case becomes apparent when one or more discrimination estimates becomes infinite. As will be indicated below, many of the recent developments are motivated, in part, by a need to cope with the Heywood case.

The Two-Parameter Logistic ICC Model

Fundamentally, item parameter estimation procedures attempt to fit a monotonically increasing mathematical function to the observed proportions of correct/keyed response to an item over the range of the θ scale. Holland (1981) discussed the conditions under which this is a proper approach. Rosenbaum (1984) provided a method for testing the local independence and monotonicity assumptions. The fitting of such functions to the observed proportion of correct response as a function of some scale has a long history in psychology and biology. Although much of the early IRT literature employed the normal ogive as an ICC model (see, e.g., Lord, 1952), it has been replaced in practice by the cumulative form of the logistic function

$$P_i(\theta_j) = \frac{1}{1+\exp[-a_i(\theta_j - b_i)]} \quad , \tag{4}$$

where a_i is the discrimination parameter,
b_i is the difficulty parameter, and
$a_i(\theta_j - b_i)$ is the logistic deviate (logit).

The first published use of this function as an ICC model appeared in Maxwell (1959). The primary reason for the switch to the logistic function is the fact that its cumulative distribution has a closed form. As a result, it has simple derivatives, and computer evaluation of the function is much less demanding than is the normal ogive.

Haley (1952) showed that if the numerical value of the scale parameter of the normal ogive was multiplied by 1.702 and used in the logistic function, the resulting curve differed from the cumulative normal by no more than .01 over the complete criterion variable scale. The tradition has been to report the parameters of the logistic function in their normal ogive equivalents, under the rationale that the logistic ogive is being used as an approximation to the normal ogive. This is an unfortunate practice, as it makes interpretation of discrimination parameters across the three logistic-based ICC models inconsistent. This author, among others, recommends that the use of the 1.702 multiplier be abandoned.

Under IRT, the logistic deviate typically is $a_i(\theta_j - b_i)$. While this parameterization can be used to obtain estimates of the item's parameters, a different parameterization is more convenient mathematically (see Bock, 1972; Bock & Aitkin, 1981). Let the logistic deviate have the linear form $(\zeta_i + \lambda_i\theta_j)$ where ζ_i is the intercept, λ_i is the slope, $a_i = \lambda_i$, and $b_i = -\zeta_i/\lambda_i$. Then the cumulative logistic distribution function is given by

$$P_{ij} = P_i(\theta_j) = \frac{1}{1+\exp[-(\zeta_i+\lambda_i\theta_j)]} \quad , \tag{5}$$

and $Q_{ij} = 1 - P_{ij}$. Taking derivatives of the log likelihood given in Equation 3 with respect to ζ_i and λ_i, for a given item, yields the following likelihood equations:

$$\frac{\partial L}{\partial \zeta_i} = \sum_{j=1}^{k} f_j(p_{ij} - P_{ij}) = 0 \tag{6}$$

$$\frac{\partial L}{\partial \lambda_i} = \sum_{j=1}^{k} f_j(p_{ij} - P_{ij})\theta_j = 0 \quad , \tag{7}$$

where $p_{ij} = r_{ij}/f_j$ is the observed proportion of correct response in group j.

Because these equations are nonlinear in the parameters, they must be solved using a Newton-Raphson procedure based upon a Taylor series approximation ignoring terms of order higher than 1. The solution equations then are

$$\begin{vmatrix} \hat{\zeta}_i \\ \hat{\lambda}_i \end{vmatrix}_{t+1} = \begin{vmatrix} \hat{\zeta}_i \\ \hat{\lambda}_i \end{vmatrix}_{t} - \begin{vmatrix} \sum_{j=1}^{k} f_j W_{ij} & \sum_{j=1}^{k} f_j W_{ij}\theta_j \\ \sum_{j=1}^{k} f_j W_{ij}\theta_j & \sum_{j=1}^{k} f_j W_{ij}\theta_j^2 \end{vmatrix}_{t}^{-1} \cdot \begin{vmatrix} \sum_{j=1}^{k} f_j W_{ij} v_{ij} \\ \sum_{j=1}^{k} f_j W_{ij}\theta_j v_{ij} \end{vmatrix}_{t} \quad , \tag{8}$$

where $W_{ij} = P_{ij}Q_{ij}$,
$v_{ij} = (p_{ij} - P_{ij})/W_{ij}$,
θ_j is the ability level of group j, and
t is the iteration index.

The Newton-Raphson equations are solved iteratively until the difference between the two successive sets of parameter estimates is sufficiently small. These equations are identical to those used in quantal response bioassay where the dosage levels are known (Baker, 1965; Mislevy & Bock, 1984). The basic data—the number of persons and the number of desired responses observed at each level of the criterion scale—are the same in both applications of these equations.

The process implemented in Equation 8 is asymptotically a weighted least-squares procedure. The term $f_j W_{ij}$ is the logistic equivalent of the Urban-Müller weights for fitting the normal ogive (Thomson, 1919). The importance of this is that θ levels with smaller f_j and those not near the item's difficulty make a smaller contribution to the determination of the item parameter estimates. In addition, the relative location of the item $|b - \bar{\theta}|$ and the item discrimination are important factors, as they determine the P_{ij} in the W_{ij}. Baker (1967) showed that a uniform distribution of examinees over the θ scale yields the minimum bias and standard errors of the item parameter estimates. This result reflects the role of these weights in the estimation procedure.

The One-Parameter Logistic (Rasch) Model

The ICC employed under the Rasch model is a special case of the two-parameter logistic model in which the discrimination parameter is set a priori to a value of unity on a logistic metric. Thus, the probability of correct response is defined as

$$P_i(\theta_j) = \frac{1}{1+\exp[-(\theta_j - b_i)]} \quad . \tag{9}$$

Because the discrimination parameter is fixed at unity, all examinees having the same raw score will obtain the same estimate of θ. As a result, the item parameter estimation procedure groups the examinees by raw score, with scores of 0 and n being eliminated. The log likelihood function is

$$L = \sum_{g=1}^{n-1} g f_{g\cdot} \theta_g - \sum_{i=1}^{n} s_i b_i - \sum_{g=1}^{n-1} f_{g\cdot} \sum_{i=1}^{n} \log\,[1+\exp(\theta_g - b_i)] \quad , \tag{10}$$

where $f_{g\cdot}$ is the number of examinees in score group g ($g = 1, 2, \ldots, n-1$),
θ_g is the trait level for score group g, and
s_i is the item score, or the number of examinees answering item i correctly.
Taking the derivative with respect to a given item's difficulty yields the likelihood equation

$$\frac{\partial L}{\partial b_i} = -s_i + \sum_{g=1}^{n-1} f_{g\cdot} P_{gi} = 0 \quad . \tag{11}$$

The Newton-Raphson equation used to solve for an item's difficulty parameter is

$$\hat{b}_{i(t+1)} = \hat{b}_{i(t)} - \left| \sum_{g=1}^{n-1} f_{g\cdot} W_{gi} \right|_t^{-1} \left| s_i - \sum_{g=1}^{n-1} f_{g\cdot} P_{gi} \right|_t \tag{12}$$

where $W_{gi} = P_{gi} Q_{gi}$, and the P_{gi}, Q_{gi} are obtained by evaluating Equation 9 using the current estimated values of b_i and θ_g.

Fischer (1981) noted that the expression in Equation 11 simply equates the marginal sum (the item score) with its expectation. The iterative estimation procedure of Equation 12 attempts to minimize the difference in the elements by finding the appropriate value of $\hat{b}_i$, assuming the θ_g are known. This MLE procedure for the Rasch model was first implemented by Wright and Panchapakesan (1969) and is presented in depth by Wright and Stone (1979) and Wright and Douglas (1977a). It is the basis for the BICAL computer program (Wright & Mead, 1978; Wright, Mead, & Bell, 1980) as well as the microcomputer version of the program, MICROSCALE (Wright & Linacre, 1984). Under the Rasch model, the unit of measurement is set to 1 and the identification problem only involves the origin. In the BICAL program, the origin is set so that $\bar{\hat{b}} = 0$. This has the effect of locating the distribution of examinee θ estimates relative to the average item difficulty. This is the converse of what hapens in the LOGIST results.

The fixing of the discrimination parameter at unity in Equation 9 has an impact upon the metric of

the parameters yielded by the JMLE procedure. The JMLE procedure will expand or contract the obtained θ scale metric until the average discrimination is unity. This must be taken into account when comparing the results yielded by different test administrations, such as is done in test equating (see Baker, 1984). Thissen (1982) employed a one-parameter IRT model in which the average value of the discrimination parameter is explicitly represented and estimated.

The Three-Parameter Logistic Model

Birnbaum (1968) proposed a three-parameter model for an ICC that would incorporate a "guessing" parameter. The model is given by

$$P_i(\theta_j) = c_i + (1-c_i)\,P_{ij}^* = c_i + (1-c_i)\,\frac{1}{1+\exp[-(\zeta_i+\lambda_i\theta_j)]}\quad , \tag{13}$$

where c_i is the guessing or "pseudo-chance level" parameter, and P_{ij}^* is the probability of correct response under a two-parameter logistic model.

This model is well known in the field of bioassay as Abbott's correction for natural mortality (Abbott, 1925). The model, with ML procedures for estimating its parameters, appears in Finney's (1952) book on bioassay. In IRT, it is assumed that the propensity to answer an item correctly by guessing is independent of the θ level of the examinee. Examination of Equation 13 shows that this is the case; the c parameter does not appear in the logit but is an additive and multiplicative factor applied to the P_{ij}^* term. The introduction of the c parameter has some serious consequences, the foremost of which is that the ICC model is no longer a logistic function (Birnbaum, 1968, p. 432). Consequently, it does not share the statistical properties of the one- and two-parameter logistic models.

Taking derivatives of the log likelihood with respect to the parameters for a given item yields the following likelihood equations:

$$\frac{\partial L}{\partial \zeta_i} = \sum_{j=1}^{k} f_j(p_{ij}-P_{ij})\,\frac{P_{ij}-c_i}{P_{ij}(1-c_i)} = 0 \tag{14}$$

$$\frac{\partial L}{\partial \lambda_i} = \sum_{j=1}^{k} f_j(p_{ij}-P_{ij})\theta_i\,\frac{P_{ij}-c_i}{P_{ij}(1-c_i)} = 0 \tag{15}$$

$$\frac{\partial L}{\partial c_i} = \sum_{j=1}^{k} f_j(p_{ij}-P_{ij})\,\frac{1}{P_{ij}(1-c_i)} = 0 \tag{16}$$

Again, these equations are nonlinear in the parameters and must be solved using a Taylor series and the Newton-Raphson procedure. Under a three-parameter model, the second derivatives of the log likelihood with respect to the item parameters contain observed data. As a result, the derivatives in the Hessian are replaced by their expectations (see Kendall & Stuart, 1961, chap. 18), yielding the information matrix. This approach is known as Fisher's method of scoring. Under the one- and two-parameter models, the derivatives do not contain observed data and the Hessian is the information matrix.

In the resulting Newton-Raphson equations for the three-parameter model, the terms in the information matrix involving the ζ and λ parameters have the same basic form as under the two-parameter model, but a "compensation" term is appended. For example, the [ζ,λ] term in the information matrix is given by

$$\sum_{j=1}^{k} f_j W_{ij}\theta_j \left[\frac{P_{ij}-c_i}{(1-c_i)P_{ij}}\right]^2 = \sum_{j=1}^{k} f_j W_{ij}\theta_j \left[\frac{P_{ij}^*}{P_{ij}}\right]^2 \quad . \tag{17}$$

A compensation term appears in all of the elements in the information matrix, but its form varies according to which parameters are paired. For example, when ζ or λ are paired with c the term involved is

$$\left[\frac{P_{ij}^{*}}{P_{ij}}\right]^{2}\left[\frac{1}{(P_{ij}-c_{i})}\right] \quad . \tag{18}$$

These compensation terms reflect the fact that the lower bound of the ICC was constrained by the c parameter. Under this model, the c parameter appears in a way which is unfortunate, from a mathematical point of view.

The LOGIST computer program is based upon a modified version of the likelihood function given in Equation 1. It employs Lord's (1974) function

$$P(\mathbf{v}|\theta) = \prod_{j=1}^{N}\prod_{i=1}^{n} P_{i}(\theta_{j})^{v_{ij}}Q_{i}(\theta_{j})^{1-v_{ij}} \quad , \tag{19}$$

where $v_{ij} = 0$ or 1 if the item is attempted and $1/M$ if omitted, and M is the number of response options to an item (Lord, 1980, p. 229; Wingersky, 1983). Lord (1974) indicated that this function is not a likelihood function, but it is used as the starting point in the same mathematical derivation of the JMLE as in Equation 1. When omitted responses are filled in at random under Equation 1, it leads to the same item parameter estimates as Equation 19. For a given item, the product over examinees is taken only over those examinees reaching an item. Lord (1974) suggested that θ estimates based on Equation 19 are better than the ML estimates when omits are randomly filled in. There does not seem to be any indication of the impact of using Equation 19 upon the resultant item parameter estimates. In actual practice, the parameter estimates based upon Equation 19 have been treated as if they were ML estimates.

Evaluation of Estimation Under the Joint Maximum Likelihood Procedures

Properties of Item Parameter Estimates

One of the limitations of MLE procedures is that an occasional dataset is encountered in which the parameters of certain items cannot be estimated. For example, the difficulty estimate for an item with no discrimination is infinite, and an item that discriminates perfectly yields $\hat{a} = \infty$. In addition, the parameters of items answered correctly by all or none of the examinees cannot be estimated. To cope with these well-known limitations of the method, the computer programs for item parameter estimation incorporate a means for detecting these situations and remove such items from the dataset before estimating the parameters of the remaining items. When ML estimators are finite, they are consistent, efficient, and sufficient. Much of the interest in the properties of ML estimates has been generated by the fact that under the Rasch model the raw test score is a minimum sufficient statistic for θ, and the item score s_i is a minimum sufficient statistic for item difficulty (Rasch, 1960). In addition, Birnbaum (1968) stated that sufficient statistics do not exist for the three-parameter model.

The unresolved question is whether the properties of ML estimators hold across the stages of the JMLE paradigm. The answer with respect to consistency is complicated by a problem inherent in simultaneous estimation of structural and incidental parameters using MLE (Kiefer & Wolfowitz, 1956; Neyman & Scott, 1948). In the present context, the item parameters are the structural parameters; because the number of θ parameters increases as the sample size increases, they are the incidental parameters. When the structural and incidental parameters are estimated simultaneously, the estimates of the structural parameters do not converge to their parameter values, that is, they are not consistent (see Hambleton & Swaminathan, 1985, p. 127 for an excellent example). Andersen (1973a) demonstrated that under the

one-parameter model the ML estimators of the item difficulty parameters were not consistent for a test of fixed length and a number of examinees approaching infinity. However, Haberman (1977) has shown that if the number of items is also allowed to increase without limit, the ML estimate of difficulty is consistent. Wingersky and Lord (1984) indicated that the consistency of the item parameter estimates for the three-parameter model has not been proven.

Although all the statistical properties of the item parameter estimates yielded by the JMLE paradigm have not been demonstrated analytically for the three-parameter model, empirical evidence exists that consistency may conform to the theoretical expectations. Swaminathan and Gifford (1983) performed a simulation study examining the consistency and bias of the item parameter estimates under a three-parameter model as yielded by the LOGIST program. Tests of lengths 10, 15, and 20 items and sample sizes of 50, 200, and 1,000 examinees were used. They found that as the number of items and the sample size increased, the regression lines for the relation of the difficulty and discrimination estimates to their parameter values differed only slightly from the theoretical 45° line. Thus, the empirical data suggest that the estimates of item difficulty and discrimination in the three-parameter model are consistent. Results for the guessing parameter were not reported. Swaminathan and Gifford also examined the bias of the item parameters yielded by LOGIST under a three-parameter model. They found that for 20 replications of a 20-item test and a sample of 200 examinees, the LOGIST program yielded an overestimate of small values of a and accurate estimates of large values of a. The overall bias of the discrimination estimates was small. In the case of item difficulty, they found that LOGIST yielded a slight underestimate of negative values of b and close estimates of large positive values of b.

Lord (1983a) derived expressions for the asymptotic bias of the ML estimators of the item parameters under a three-parameter model. The bias expressions were for a single item, assuming known examinee θ parameters. To investigate the characteristics of the asymptotic biases, Lord used simulated data having parameter values roughly equal to the $\hat{a}$, $\hat{b}$, $\hat{c}$, and $\hat{\theta}$ yielded by 2,995 examinees on a 90-item verbal Scholastic Aptitude Test. In the case of item difficulty, easy and medium difficulty items had a negative bias and only difficult items were positively biased. The bias in the estimates of item discrimination was always positive. The bias in the estimate of c was negative for all items. In general, if an item's parameter estimates had large standard errors the biases were also large. However, the magnitude of an estimator's bias typically was about .1 of its standard error and very seldom was greater than .2 of its standard error. Lord concluded that because the standard errors are inversely proportional to sample size, when N is large the numerical value of the bias is probably negligible. These results, however, are for a restricted set of conditions.

In the case of the Rasch model, Wright and Douglas (1977b) investigated the bias in the item difficulty estimates. They were concerned with the bias of the JMLE procedure, which they labeled the unconditional approach (UCON), relative to the estimates yielded by the conditional maximum likelihood approach (CML). They found that multiplying the UCON item difficulty parameters by a factor of $(n-1)/n$ resulted in very close agreement with those yielded by the conditional approach. This, however, is not the same as the difference between the parameter and its estimator.

Item Parameter Recovery Studies

A number of studies have examined the degree to which the JMLE paradigm can recover the underlying item parameters. Typically these studies use simulated data, so that the estimates and the true values can be related. Lord (1975) used simulated data for a 90-item test and 2,995 examinees. As is usual in these studies, the item parameter values were matched approximately to those of an existing test. The correlation between the discrimination estimates and their parameters was .920. When c was overestimated, a also

tended to be overestimated. Item difficulty was overestimated for large absolute values of b, $|b > 3.0|$, and slightly underestimated for medium values of b. The correlation between the difficulty estimates and the parameter values was $r = .988$. In the case of both a and b, the estimates were scattered rather tightly about the 45° theoretical relationship line. The estimates of the c parameter were widely scattered about the corresponding parameter values and generally underestimated the parameter values.

Hulin, Lissak, and Drasgow (1982) investigated the recovery of item parameters under the two- and three-parameter models. Using the root mean square (RMS) between the recovered and empirical ICCs as the criterion, they found that minimum test lengths and sample sizes depended upon the model. Under a two-parameter model, 30 items and 500 examinees gave satisfactory results in terms of RMS, while the three-parameter model required 60 items and 1,000 examinees. They also found that a tradeoff between test length and sample size gave comparable results, at least for the data used. The correlation between the estimates and the parameter values was also higher for the two-parameter model than for the three-parameter model. When $N > 500$, the correlation of $\hat{a}$ and a was roughly .9 for the two-parameter model and .5 for the three-parameter model. Under a two-parameter model, $r_{\hat{b}b}$ was greater than .94 for all sample sizes and test lengths. Under a three-parameter model, $r_{\hat{b}b}$ was greater than .94 only when $N \geq$ 1,000 and $n > 30$ items.

The characteristics of the standard errors of the item parameter estimates are also an important issue. The asymptotic standard errors of the item parameter estimates are given by the inverse of the information matrix employed in the Newton-Raphson equation. Thus, they are readily available as a byproduct of the item parameter estimation process. The values obtained in practice are those yielded by the final stage of the JMLE procedure when the overall convergence criterion has been met. Thissen and Wainer (1982) investigated the asymptotic standard errors of the item parameters for the one-, two-, and three-parameter models under the assumption that the θs of the examinees were known and normally distributed $n(0,1)$. Tables of the minimum asymptotic standard errors were reported for combinations of parameter values under each of the three models. Plots of the standard error of b for all three models showed a concave surface that increased at the extremes of the θ scale.

An interesting set of results was given by the two-parameter model and the three-parameter model when $c = 0$. Even though the numerical values of a and b would be the same, the information matrices are not. The three-parameter matrix still has a row and column corresponding to the c parameter. When an item was easy and had low discrimination, the standard errors under the two-parameter model were roughly .09 of those reported for the three-parameter model. Clearly, the two-parameter model and the three-parameter model with $c = 0$ are not the same with respect to the standard errors of the item parameter estimates. The asymptotic standard errors for item difficulty under the Rasch model were consistently smaller than those obtained for the other two models. In particular, the increase in standard error with departure of item difficulty from zero was much less pronounced.

Based upon the results, Thissen and Wainer felt that the three-parameter model was inferior to the other two models. The standard errors of the item parameters under a three-parameter model were acceptable only in the middle of the ranges of b and a with low values of c. They concluded that "the use of an unrestricted maximum likelihood estimation for the three-parameter model either yields results too inexact to be of any practical use, or requires samples of such enormous size so as to make them prohibitively expensive" (p. 403). A subsequent recommendation (p. 410) was that a tight Bayesian prior around the lower asymptote might be a vehicle for obtaining smaller standard errors. De Gruijter (1984) made a limited examination of this approach and found that the use of a prior distribution on the c parameter did reduce the asymptotic standard errors of all three parameters.

One of the limitations of the equations for the asymptotic standard errors of the item parameters used above is that they were based upon a single item and known θ. Lord and Wingersky (1985) derived

a method for computing the asymptotic sampling variance-covariance matrix of JMLEs when all parameters are unknown. They employed this method to study the asymptotic standard errors of the three-parameter model (Wingersky & Lord, 1984). Parameter sets for simulated tests of 45 and 90 items were established in which the parameter values of 15 items were replicated. Two θ distributions, bell-shaped and uniform, were used with each of two sample sizes, $N = 1{,}500$ and $N = 6{,}000$. Only the standard errors of the basal 15 items were reported. The scale used required that the mean of the difficulty parameters of certain selected items be 0 (the origin) and that the difference between two such sets of selected items be 1 (the scale unit). This metric was called the "capital" scale. The standard errors were reported for the item parameter estimates $\hat{A}$, $\hat{B}$, and $\hat{C}$ in this metric.

Lord and Wingersky (1985) concluded that, for the values of n and N employed, the standard errors of the item parameter estimates $\hat{A}$, $\hat{B}$, and $\hat{C}$ varied inversely as $N^{1/2}$ but were only slightly affected by changes in n. Using the rectangular θ distribution yielded smaller standard errors for the item parameters than did doubling the number of items under a bell-shaped θ distribution. For low As and for Cs from items with $B - 2/A$ less than 1, the standard errors computed with a rectangular distribution of θ were nearly as low as the standard errors computed with a bell-shaped distribution and quadruple the number of persons. This concurs with Baker's (1967) conclusion that the uniform distribution of θ is preferred.

De Gruijter (1985) derived equations for the asymptotic standard errors of the item difficulty parameters under the Rasch model. He noted that the size of the standard errors yielded by these formulas and by those of Wingersky and Lord (1984) depends strongly upon the restrictions introduced to solve the identification problem, that is, to anchor the θ metric. In the case of the Rasch model, anchoring the metric by setting the mean $b = 0$ and setting one item difficulty, say, $b_n = 0$ yielded different values.

One recurring theme in the IRT literature concerns the problems associated with the estimation of the c parameter of the three-parameter model. For example, Kolen (1981) found for three tests that 92%, 53%, and 39% of the cs were not successfully estimated. The index $b - 2/a$ (Lord, 1975) identifies the point on the θ scale where the ICC is within .03 of the value of its lower asymptote (c). When the criterion for the $b - 2/a$ index is not met, the LOGIST program sets the value of c for an item to the mean of the values for those items where c was successfully estimated. The current LOGIST manual (Wingersky et al., 1982) suggests an index value of -3.5 for samples of 2,000 to 3,000 and a value of -2.5 for smaller samples.

Lord (1980) and many other authors have suggested that the c parameter cannot be estimated unless a considerable lower tail to the ICC is present in the range of θ scores employed. In addition, very few studies have plotted the proportions of correct response to an item as a function of θ in order to show that a three-parameter model is needed. McKinley and Reckase (1980) plotted such data for 50 items taken from the Iowa Tests of Educational Development and a sample of 1,999 examinees. Only 8 of the 50 items showed a lower tail, suggesting that a three-parameter model was needed. The data were analyzed using the LOGIST program and the fitted ICC plotted. For these 8 items, the scatter of the empirical proportions of correct response about the lower tail of the ICC was much larger than elsewhere along the curve and would not be binomially distributed. These data suggest that c would be poorly estimated even when the criterion for Lord's index was met.

Relatively little attention has been paid to why the estimation of the parameters of the three-parameter model is fraught with so much difficulty. However, the Newton-Raphson procedure moves over the log likelihood surface in a quadratic fashion and it is possible for the increments to the item parameter estimates to improperly position the estimates on the surface. When this happens, the iterative process diverges rather than converges. Householder (1953) showed that a parameter estimate must be within a certain neighborhood of the parameter in order for the Newton-Raphson procedure to converge. Kale (1962) extended these results to MLE of multiple parameters under Newton-Raphson and Fisher's method

of scoring. He also showed that the initial estimates must be consistent estimators if the iterative estimation procedure is to converge.

The available computer programs differ with respect to the initial estimators employed in the JMLE procedure. In the LOGIST program, severe constraints are imposed upon the change in values of the item parameters at each stage of the paradigm (Wingersky, 1983). These constraints are designed to keep the movement of the item parameter estimates to a minimum until they are located within the convergence neighborhood, and also to prevent them from leaving the neighborhood. Because of this, the LOGIST program is quite expensive to use due to the slow rate of convergence within the item parameter estimation stage. The latest version of LOGIST (Wingersky et al., 1982) employs acceleration techniques (Ramsey, 1975) to speed convergence.

Goodness of Fit of ICCs

After the parameters of the ICC have been estimated, interest is focused on how well this curve fits the item response data. When the examinees have been grouped in the estimation process, a chi-square goodness-of-fit index can be obtained rather easily. Berkson (1944, 1955) has shown that the usual chi-square index can be written as

$$\chi_i^2 = \sum_{j=1}^{k} f_j \frac{(p_{ij} - P_{ij})^2}{P_{ij}Q_{ij}} \quad . \tag{20}$$

All elements in this equation are readily available from the last iteration of the item parameter estimation process; this is another justification for grouping examinees along the θ scale. Because only the P_{ij} term depends upon the ICC model, Equation 20 can be used with the one-, two-, or three-parameter models. The degrees of freedom will be the number of grouping categories less the number of item parameters estimated. The observed value of this statistic can be used to identify items where there is not a good fit to the observed item response data.

Hambleton and Swaminathan (1985) indicated that given the large sample sizes usually involved in test analysis, many significant values of the chi-square statistic may result. However, this is an artifact of the sample size rather than a lack of fit. Thus, when evaluating items, the relative values of the indices should be considered, rather than the statistical significance. The results of McKinley and Mills (1985) suggest that samples greater than 1,000 are needed before this becomes an issue. Surprisingly, the LOGIST program does not provide the user with a measure of the goodness of fit for each item. Perhaps because of this, there is relatively little discussion in the literature of how well the ICCs defined by the item parameter estimates yielded by LOGIST fit the item response data.

Yen (1981) performed a simulation study in which the item response data for 1,000 examinees and a 36-item test were generated under all three ICC models. Each dataset was then analyzed under all three models and a goodness-of-fit measure was computed for each item. The grouping intervals for the fit statistic were defined by rank ordering the examinees on the basis of their estimated θs. Then 10 intervals were established having approximately equal numbers of examinees per cell. The fit statistic Q_1 differs from the usual chi-square index given in Equation 20 in one respect: In Equation 20, the grouping intervals are usually defined a priori and can be either equally spaced over the θ scale or fractiles, and the expected number of correct responses f_jP_i is calculated from the midpoint value of θ for the interval; whereas in the case of Yen's Q_1 index, the expected frequency of correct response reflects the average rather than the midpoint of θ for the interval.

The empirical results obtained by Yen (1981) showed that when an estimating model had fewer parameters than the generating model, a lack of fit would be observed. The exception was when data

generated by the three-parameter model were analyzed using a two-parameter model. She stated that "when the data were generated by the [three-parameter] model, the [two-parameter] estimating model did a surprisingly good job of fitting the data. If a set of real data were being analyzed with all three models, it would be difficult to decide on the basis of the pattern of mean fit values whether the [three-parameter] or [two-parameter] model was more appropriate" (p. 251). She based the explanation of these results upon the fact that when three-parameter data were analyzed under a two-parameter model, the discrimination index compensates for the presence of a non-zero lower asymptote in the data. In such cases, the estimation process yields a lower value of the discrimination index and the result is a better fit to the data. She also noted that "because an item's discrimination can be affected by the item's difficulty level relative to the ability level of the group of examinees, the [two-parameter] estimation of discriminations can be sample dependent for [three-parameter] data" (p. 260).

McKinley and Mills (1985) conducted an extensive investigation of goodness-of-fit indices. They compared four such indices, those developed by Bock (1972), Yen (1981), Wright and Mead (1978), and the Likelihood Ratio (LR) statistic. The first three of these employ the standard chi-square goodness-of-fit formula and vary only with respect to the number of groups and the definition of the θ level used to compute the expected proportion of correct response. Twenty-seven 75-item tests and sample sizes of 500, 1,000, and 2,000 examinees were used to generate simulated data. The tests were created such that nine tests were generated under each of the one-, two-, and three-parameter models. In addition, the normally distributed samples had means of -1, 0, and 1 on the θ scale. All combinations of tests and groups were analyzed under the three ICC models using the LOGIST program. When the data generated by the two- and three-parameter models were analyzed under a one-parameter model, the results indicated a consistent lack of fit. As was the case with Yen's (1981) study, analyzing three-parameter data using the two-parameter model worked quite well.

McKinley and Mills (1985) concluded that the LR index appeared to yield the fewest erroneous rejections of the hypothesis of fit, while the Bock index yielded fewer erroneous conclusions of fit. However, the differences in results were slight. They also applied the four procedures to an additional nine tests having an underlying multidimensional θ structure. In all cases, the analyses yielded a high proportion of misfits. Thus, the underlying assumption of unidimensionality appears to be critical to obtaining good fit between the ICC and the observed data.

Other Estimation Techniques

Techniques Based on Relationships With Classical Test Theory

In the early days of IRT, considerable emphasis was placed upon the relation of the parameters of the normal ogive ICC model and the classical test theory item statistics. Lawley (1943), in a paper now considered the keystone of IRT, derived these relationships for the case of equivalent items. Under an assumption of normality for the distribution of examinees over the θ scale, Tucker (1946) showed that the following hold:

$$a_i = \frac{\rho_{I\theta}}{(1 - \rho_{I\theta}^2)^{1/2}} \tag{21}$$

$$b_i = \gamma_{i\theta}/\rho_{I\theta} \quad , \tag{22}$$

where $\rho_{I\theta}$ is the biserial correlation of the item variable with the ability variable, and γ_i is the normal deviate delimiting the area of the normal density corresponding to one minus the classical item difficulty.

Brogden (1971) developed a number of formulas relating the biserial and point-biserial correlations

of the item variable with either test score or θ under a three-parameter model. The biserial correlation of the item and θ can be obtained from the point-biserial correlation using

$$\rho_{I\theta} = \frac{\rho_{i\theta}[\pi_i(1 - \pi_i)]^{1/2}}{\phi(\gamma_i)(1 - c_i)}, \tag{23}$$

where π_i is the item difficulty, evaluated using the three-parameter model,
$\rho_{i\theta}$ is the point-biserial correlation of the item and θ, and
$\phi(\gamma_i)$ is the ordinate of the normal density at the deviate γ_i.

Due to the computational demands of the MLE approach to item parameter estimation, there was an interest in developing a simpler item parameter estimation procedure based upon the work of Lawley, Tucker, and Brogden. The above equations offer a computational advantage because their solutions are noniterative and involve simple terms, such as the mean score of the examinees who answered the item correctly and the item difficulty. In addition, the normal ogive must be evaluated only once for each item, rather than once at each θ level for each item and iteration (as in the JMLE procedure). In an item analysis context, Baker (1959) used item difficulty and the biserial correlation of the item with the total test score for $\rho_{I\theta}$ in Equations 21 and 22 to obtain estimates of a_j and b_j. The use of the total test score as a substitute for θ is flawed in that it results in a linear test characteristic curve (Lord, 1965).

Employing the logistic ogive as an approximation to the normal ogive, Jensema (1976) used this approach to estimate item parameters under a three-parameter model. He wrote a computer program that used the point-biserial correlation of the item and the item-excluded test score in Equation 23 to obtain the biserial correlation to use in Equations 21 and 22. This approach was termed the "simple" approach. He also developed a computer program for MLE under a three-parameter model. The two methods were compared using simulated datasets. He found correlations of $r_{\hat{a}a} = .789$ and $r_{\hat{b}b} = .963$ between the simple procedure and the underlying parameters. The c parameter was fixed at .2 and was not estimated. The corresponding values for MLE were $r_{\hat{a}a} = .863$ and $r_{\hat{b}b} = .971$. Thus, good agreement between the methods was obtained even though the item-excluded test score was used in the simple approach. Jensema apparently did not pursue the approach beyond this point. His use of the logistic ogive as an approximation to the normal ogive in a correlational context is somewhat flawed, as Gumbel (1961) has shown that the bivariate logistic and the bivariate normal distributions are not equivalent.

Urry (1974, 1976) developed a computer program based upon the Tucker and Brogden equations that estimated both item and θ parameters under a three-parameter normal model. The three-stage paradigm was as follows: In the first stage, the item parameter estimates $\hat{a}_i$ and $\hat{b}_i$ were obtained through Equations 21, 22, and 23 using item difficulty and the observed point-biserial correlations based upon the item-excluded test score. In order to estimate the c parameter, a stepwise chi-square goodness-of-fit procedure was employed. Various values of $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ ($0 \leq \hat{c} \leq .3$) that reproduced the obtained values of item difficulty and the point-biserial correlation were tried until the minimum value of chi-square was obtained for the fit of the specified ICC to the observed proportions of correct response. The examinees were grouped by test score for the chi-square calculations. Next, improved estimators using corrections based upon the item information functions were obtained. These corrections were then added to the item parameter estimates and the first stage was completed.

In the second stage, Bayesian modal estimates of θ (Samejima, 1969) were used to obtain θ estimates for each of 100 score intervals. To resolve the identification problem, the θ estimates then were rescaled to a mean of zero and unit variance. In the third stage, the item difficulty and $\rho_{I\theta}$ were recalculated using the Bayesian θ estimates instead of the item-excluded test score. This yielded new values of $\hat{a}_i$ and $\hat{b}_i$. To find $\hat{c}_i$, the stepwise goodness-of-fit procedure was again employed. The corrections were then applied to the item parameter estimates and the paradigm was completed. Although this procedure seems rather complex and convoluted, it involves only a few evaluations of the normal ogive per item and, unlike the

JMLE procedure, does not iterate until a convergence criterion is met. Thus, it offers some promise of being more economical.

One technique in the Jensema and Urry approaches is worth noting. Both used the point-biserial correlation of the item and the item-excluded test score. In the computation of the point-biserial correlation, the test score distribution was dichotomized at $\bar{X}$. If $\bar{X}$ is a good estimate of $\bar{\theta}$, then the θ scale has also been dichotomized. Thus, the point-biserial correlation of θ and the item is obtainable without actually having θ measures. Given the point-biserial, the biserial correlation of the item and θ used in Equations 21 and 22 can be obtained from Equation 23.

The item parameter recovery properties of the Urry computer program (known as ESTEM, OGIVIA, or ANCILLES in its several versions) was investigated by Urry and his co-workers. Urry (1976) found for simulated samples of 2,000 and 3,000 examinees that the root mean square of the difference between the estimates and the item parameters (RMSE) decreased over the stages of the procedure. The correlations of the estimates with the underlying item parameters for a 100-item test were $r_{\hat{b}b} = .996$, $r_{\hat{a}a} = .915$, and $r_{\hat{c}c} = .760$ with $N = 2{,}000$. Gugel, Schmidt, and Urry (1976) found a general decrease in RMSE as both sample size and the number of items increased, as well as a decrease across stages of the procedure. The root mean squares were roughly .322, .140, and .062 for $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$. However, in a number of cases the RMSE increased between the first and last stages. Schmidt (1977) reported that the Urry procedure tended to underestimate a_i and overestimate $|b_i|$. He presented formulas to correct for these biases.

Ree (1979) compared the item parameter estimates yielded by Urry's ANCILLES and OGIVIA programs with those yielded by LOGIST and with estimates derived from the biserial correlation of items and test scores in Equations 21 and 22. He termed the latter the transformation technique. Three simulated datasets of 2,000 examinees each were used. One dataset had a uniform θ distribution, with 2,000 unique values of θ over the range -2.5 to 2.5. The second dataset had a truncated uniform distribution. The third had a normal distribution. The item responses to an 80-item test were generated using a computer program. With respect to parameter recovery, Ree indicated that both ANCILLES and OGIVIA compared quite well with LOGIST. Although the estimates of b_i yielded by the transformation method for the three datasets correlated well with the underlying values ($r_{\hat{b}b} = .963, .917, .965$), the correlation for the a_i was quite low ($r_{\hat{a}a} = .59, .32, .35$).

These results are at odds with those of Urry (1976), which showed the correlation $r_{\hat{a}a}$ to be around .88. Ree (1979) did not present a detailed definition of his transformation procedure and it is difficult to determine if the data, the approach, or its implementation is responsible for the discrepancy. An interesting sidelight of the Ree study was the result that the correlation of raw test score with the underlying θ levels of the examinees ($.936 \leq r \leq .977$) was as high or higher than the correlation of θ with the θ estimates yielded by ANCILLES, OGIVIA, and LOGIST ($.935 \leq r \leq .974$). Ree (1979) concluded that the OGIVIA procedure was preferable if the examinees were normally distributed over the θ scale. He also provided some cost data in terms of computer time used to perform the analyses. On a large-scale computer, the UNIVAC 1108, LOGIST required 2,061 seconds, ANCILLES required 296 seconds, OGIVIA required 180 seconds, and the transformation technique required 38 seconds. Clearly, the procedures based upon the classical theory approximations were much less expensive than was LOGIST.

The item parameter estimation characteristics of ANCILLES and LOGIST were also compared by Swaminathan and Gifford (1983). In the simulation study described earlier, each of the 36 data combinations was analyzed by both of the computer programs. They concluded that the JMLE procedure implemented in LOGIST was, in general, superior to the Urry procedure implemented in ANCILLES. However, with long tests and large samples there was little difference in results and the cost of an ANCILLES computer run was much less.

Conditional Maximum Likelihood Estimation

Rasch (1960) showed that under his probabilistic model the raw test score was a minimum sufficient statistic for estimating θ and the item score was a minimum sufficient statistic for item difficulty. Using the Rasch results as a starting point, Andersen (1970, 1972, 1973a) developed a MLE procedure for the estimation of the item difficulty parameters that did not involve the examinees' θ parameters. These were replaced by the raw test score, that is, the item parameter estimates were conditioned upon the minimum sufficient statistics for the θ parameters. The basic equations for this conditional maximum likelihood (CML) estimation procedure (Wright & Douglas, 1977b) are presented below.

Under the assumption of local independence, the probability of person j yielding item response vector $\mathbb{u}_{ij}$ when responding to n items of a test is given by

$$P(\mathbb{u}_{ij}|\theta_j,\mathbb{b}_i) = \frac{\exp\left(r\theta_j - \sum_{i=1}^{n} u_{ij}b_i\right)}{\prod_{i=1}^{n}[1+\exp(\theta_j - b_i)]} \quad . \tag{24}$$

Because a person of trait level θ_j can obtain a given test score r in $\binom{n}{r}$ ways and the sum of the probability of these is the probability that $u_{\cdot j} = r$, then

$$P(u_{\cdot j} = r|\theta_j,\mathbb{b}_i) = \frac{\gamma_r \exp(r\theta j)}{\prod_{i=1}^{n}[1+\exp(\theta_j - b_i)]} \quad , \tag{25}$$

where

$$\gamma_r = \sum_{\mathbb{u}_{ij}}^{r} \exp\left(-\sum_{i=1}^{n} u_{ij}b_i\right) = \gamma(r;\, b_1,\, b_2,\, \ldots,\, b_n) \tag{26}$$

is the unitary symmetric function. Then the conditional probability of the item response vector $\mathbb{u}_{ij}$, given that person j has test score r, is

$$P(\mathbb{u}_{ij}|r,\mathbb{b}_i) = \frac{\exp\left(-\sum_{i=1}^{n} u_{ij}b_i\right)}{\gamma(r;\, b_1,\, b_2,\, \ldots,\, b_n)} \quad . \tag{27}$$

Thus, the probability of the item response vector is conditional upon the test score and the vector of item parameters $\mathbb{b}_i$. If N persons have responded to the set of items and the responses from different persons are assumed independent, the conditional probability distribution of the matrix of item responses, given the values of the raw scores $(r_1,\, r_2,\, \ldots,\, r_N)$ of the N examinees, is

$$P(\mathbb{u}_{ij}|r_1,\, r_2,\, \ldots,\, r_N;\, b_1,\, b_2,\, \ldots,\, b_n) = \frac{\exp\left(-\sum_{j=1}^{N}\sum_{i=1}^{n} u_{ij}b_i\right)}{\prod_{j=1}^{N}\gamma(r;\, b_1,\, b_2,\, \ldots,\, b_n)} \quad . \tag{28}$$

At this point, Andersen (1973a) took advantage of the fact that the possible values of the test score r (0,

$1, 2, \ldots, n$) are usually less than the number of persons possessing these scores. As a result, the denominator of Equation 28 can be written as

$$\prod_{j=1}^{N} \gamma(r; b_1, b_2, \ldots, b_n) = \prod_{r=0}^{n} [\gamma(r; b_1, b_2, \ldots, b_n)]^{f_r} \quad , \tag{29}$$

where f_r is the number of persons having score r. Letting $s_i = \sum_{j=1}^{N} u_{ij}$ be the item score, $b = (b_1, b_2, \ldots, b_n)$, and omitting scores of 0 and n, Equation 28 can be rewritten as

$$L(b) = \frac{\exp\left(-\sum_{i=1}^{n} s_i b_i\right)}{\prod_{r=1}^{n-1} [\gamma(r,b)]^{f_r}} \quad , \tag{30}$$

and taking logarithms yields

$$\log L(b) = -\sum_{i=1}^{n} s_i b_i - \sum_{r=1}^{n-1} f_r \log [\gamma(r,b)] \quad . \tag{31}$$

At this point, the usual iterative Newton-Raphson procedures are used to obtain the item difficulty parameter estimates. Andersen (1973a) has shown, for the Rasch model, that the CML procedure yields consistent estimators. As was the case with previous item parameter estimation procedures, there is an indeterminacy in the definition of the underlying metric. Because the Rasch model sets the unit of measurement at 1, only the location needs to be fixed. To do this, Andersen (1973a) imposed the constraint that one item have a difficulty of zero, say, $b_n = 0$. Wright and Douglas (1977b) implemented this by subtracting the derivatives for the nth item from those of the remaining items, thus eliminating one item from the estimation process.

Because this ML estimation process is conditional upon the sufficient statistics for θ, it only yields item parameter estimates. Conceptually, the approach can be turned around to obtain θ estimates conditional upon the sufficient statistics for the item parameters. However, Wainer, Morgan, and Gustafsson (1980) reported that this approach is fraught with both theoretical and computational problems. Thus, under the CML approach there is not a simultaneous estimation of both item and θ parameters. A test analysis under the CML approach would yield the item parameter estimates and the raw test scores would be used as the θ measure.

The major drawback to the use of the CML approach is that the unitary symmetric functions must be evaluated. Equations 28 and 31 involve taking products of exponentials over all of the vectors of item responses yielded by the examinees. In addition, these functions are involved in all of the derivative terms appearing in the Newton-Raphson equations. Gustafsson (1980a) indicated that a symmetric function of order r consists of a sum of $\binom{n}{r}$ products, each of which consists of r terms. When a test has 50 items, for a score of 25 there are 1.26×10^{14} terms, each of which is a product of 25 terms. Even with a large computer this is a tedious and expensive task.

Wright and Douglas (1977b) developed a computer program for the CML approach. They reported that the problem in the CML procedure was the accumulation of the round-off errors in the calculation of the symmetric functions. As a result, only tests of 10 to 15 items could be analyzed. In an attempt to minimize this problem, they assumed the items were independent (which reduces the information matrix to diagonal form) and developed an iterative procedure for evaluating the symmetric functions. This approach delayed the round-off problem to tests of 20 to 30 items. They called this the incomplete conditional maximum likelihood (ICON) approach. Several efforts at reducing the impact of round-off errors have appeared in the literature. Sanathanan (1974) reported a procedure attributed to Scheiblechner

(1971) for the recursive computation of functions of the γ terms that appears to delay the error accumulation problem. Gustafsson (1980a) reported a recursion algorithm for computing the symmetric functions that delays the errors due to the round-off problem until 60 to 80 items have been reached. This procedure has been used by Wainer et al. (1980).

The study by Wright and Douglas (1977b) appears to be the only one in which the parameter recovery properties of the CML procedure were investigated. They used simulated data based upon tests of 20 and 40 items and samples of $N = 500$ having normal and truncated normal distributions over the θ scale. Several combinations of means and standard deviations of θ were used with each distribution. Each test and sample combination was analyzed by both the JMLE and ICON procedures. In addition, 15 replications of each combination were generated and analyzed. The raw difference between the maximum value of the estimate, in the 15 replications, and the item parameter was obtained. In addition, the means of the absolute values of these differences and the RMSE were reported. The mean absolute differences were all less than .13 and most were less than .10 for both methods. The RMSE were also uniformly small, with those yielded by the ICON procedure being slightly larger.

Wright and Douglas (1977b) concluded that there was little difference in the two approaches with respect to parameter recovery. Both methods were adversely affected when the test difficulty was not appropriate to the θ level of the examinees. Skewness of the θ distributions also tended to increase the maximum difference, particularly for the ICON procedure. This was attributed to the existence of extreme item parameter values that were not estimated well by ICON due to rounding errors.

Andersen (1973b) derived a likelihood ratio test for the global goodness of fit of the Rasch model under the CML procedure. In the equations presented above, the item parameters were estimated by using the total sample of examinees. The resulting likelihood was given by Equation 30 above. If only those examinees having a particular raw score r were considered, Equation 30 would become the restricted likelihood function

$$L^{(r)}(b^{(r)}) = \frac{\exp\left(-\sum_{i=1}^{n} b_i s_i^{(r)}\right)}{[\gamma(r,b)]^{f_r}} \quad , \tag{32}$$

where $s_i^{(r)} = \sum_{(j=1)\varepsilon Gr}^{N} u_{ij}$ is the item score for those examinees having raw score r, and f_r is the number of examinees in group Gr. The item difficulties could then be estimated from this restricted likelihood using the usual Newton-Raphson procedures. There would, of course, be as many sets of item difficulty estimates as there were raw scores.

The likelihood ratio test is based upon a comparison of these two ways of estimating the item difficulties:

$$\text{LR} = \frac{L(\hat{b})}{\prod_{r=1}^{n-1} L^{(r)}(\hat{b})} \quad . \tag{33}$$

If the model is true, the value of LR should remain close to 1. If the model is not true, some values of the item difficulties yielded by one or more raw scores will differ from the overall values and LR will be small. The test statistic used is

$$z = -2 \log \text{LR} = 2 \sum_{r=1}^{n-1} \log L^{(r)}(\hat{b}) - 2 \log L(\hat{b}) \quad , \tag{34}$$

which is distributed as chi-square with $(k-1)(k-2)$ degrees of freedom as the number of examinees having each raw score $n_r \to \infty$, and k is the number of raw score groups. This test statistic is sensitive to

a lack of equality of the item discrimination parameters and hence of the appropriateness of the Rasch model.

Gustafsson (1980b) indicated that this general approach could be applied to other groupings of examinees, such as gender, school, or other demographic characteristics. In such cases the test would be a test of unidimensionality, as it would reflect the lack of group invariance of the item parameters. However, the grouping should not be related to performance differences on the test, in which case the test would also be sensitive to discrimination differences. Gustafsson (1980b) also presented a test for the hypothesis that two disjoint groups of items measure the same construct. This provides a test of unidimensionality when items are grouped a priori. Van den Wollenberg (1982) also has developed test statistics for lack of equality of item discrimination parameters and unidimensionality.

These types of tests do not provide any information on the goodness of fit of a given ICC to the item response data. Rather, they are global measures of how all the items in a test fit the Rasch model. To assess goodness of fit to a given ICC, Gustafsson (1980b) suggested using graphical procedures. Thus, the goodness-of-fit measures associated with the CML approach do not provide the type of information needed by the test constructor to evaluate individual items. However, Molenaar (1983) has provided procedures for a more detailed analysis of results under the Rasch model.

Bayesian Estimation of Item Parameters

The bulk of the older Bayesian IRT literature dealt with estimating θ within the context of adaptive testing (see Birnbaum, 1969; Kearns & Meredith, 1975; Owen, 1969, 1975) and is not of direct interest to the present review. However, Swaminathan and Gifford (1982, 1985) have developed Bayesian procedures for the joint estimation of item and θ parameters for the one- and two-parameter ICC models (see also Hambleton & Swaminathan, 1985, chap. 7, for a brief discussion of the three-parameter model). Under the Bayesian approach, prior distributions can be specified for the parameters to be estimated. These prior distributions are multiplied by the likelihood function, which is based on the ICC model and the actual item responses. The result is a posterior distribution that yields estimates of the parameters of interest. Under this approach, the prior distributions have the effect of constraining the potential values of the parameter estimates. The basic approach of these authors, under the one-parameter model, is presented in abbreviated form below.

The likelihood function is

$$L(\mathbf{U}|\boldsymbol{\theta},\mathbf{b}) = \frac{\exp\left(\sum_{j=1}^{N} r_j\theta_j - \sum_{i=1}^{n} s_i b_i\right)}{\prod_{i=1}^{n}\prod_{j=1}^{N}[1 + \exp(\theta_j b_i)]}, \tag{35}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$, $\mathbf{b} = (b_1, b_2, \ldots, b_n)$, $r_j = \sum_{i=1}^{n} u_{ij}$, and $s_i = \sum_{j=1}^{N} u_{ij}$. It is then necessary to specify the prior distributions of the θ_j and the b_i. In the first step of the hierarchical Bayesian model, it is assumed that the prior distribution of θ_j is independent and normal with mean μ_θ and variance σ_θ^2. The prior distribution of the independent item difficulties is normal with mean μ_b and variance σ_b^2. In the second step it is assumed a priori that μ_θ and μ_b are each uniformly distributed and that μ_θ, σ_θ^2 are independently distributed, as are μ_b and σ_b^2. Then, $p(\mu_\theta,\sigma_\theta^2) \cong p(\sigma_\theta^2)$. It is assumed that σ_θ^2 has the inverse chi-square distribution

$$P(\sigma_\theta^2|\nu_\theta,\xi_\theta) \propto (\sigma_\theta^2)^{-(\nu_\theta/2+1)} \exp(-\xi_\theta/2\sigma_\theta^2), \tag{36}$$

where ν_θ and ξ_θ are the parameters and are specified a priori. $p(\mu_b,\sigma_b^2) \cong p(\sigma_b^2)$ also is distributed as the

inverse chi-square with parameters ν_b and ξ_b, which are specified a priori. The joint posterior distribution function of θ and b, then, is the product of the likelihood and the prior distributions

$$p(\theta,b,\mu_\theta,\sigma_\theta^2,\mu_b,\sigma_b^2|x,\nu_\theta,\xi_\theta,\nu_b,\xi_b) \propto L(\theta,b|x)\left[\prod_{j=1}^{N} p(\theta_j|\mu_\theta,\sigma_\theta^2)\prod_{i=1}^{n} p(b_i|\mu_b\sigma_b^2)\right]p(\sigma_\theta^2|\nu_\theta,\xi_\theta)p(\sigma_b^2|\nu_b,\xi_b) \quad (37)$$

This equation is then integrated with respect to σ_θ^2 and μ_θ and then separately with respect to σ_b^2 and μ_b. Under the Bayesian approach, the estimates of the parameters θ_j and b_i are the modes of the joint posterior distribution. These are obtained by solving the following system of equations:

$$f(\theta_j) = \sum_{i=1}^{n} P_{ij} + (\theta_j - \bar{\theta})\left[\frac{\nu_\theta + N - 1}{\nu_\theta\xi_\theta + \sum_{j=1}^{N}(\theta_j - \bar{\theta})^2}\right] - r_j = 0 \quad (38)$$

$$f(b_i) = \sum_{j}^{N} P_{ij} - (b_i - \bar{b})\left[\frac{\nu_b + n - 1}{\nu_b\xi_b + \sum_{i=1}^{n}(b_j - \bar{b})^2}\right] - s_i = 0 \quad (39)$$

where $i = 1, 2, \ldots, n$,
$j = 1, 2, \ldots, N$,
ν_θ and ξ_θ are the parameters of the inverse chi-square distribution of σ_θ^2, and
ν_b and ξ_b are the parameters of the inverse chi-square distribution of σ_b^2.

Because the system of equations is nonlinear, the Newton-Raphson procedure is used to solve each of the $n + N$ equations iteratively. Swaminathan and Gifford (1982) reported that the Newton-Raphson procedure used here is not the vector version used above in the JMLE discussion, but is one where the correction term is defined by $f(x)/f'(x)$, where $f'(x)$ is the first derivative. Before the equations can be solved, the values of the parameters of the inverse chi-square distributions of the variances employed in the prior distributions of b and θ must be specified a priori. The overall paradigm parallels the two-stage JMLE procedure and is iterated until some suitable convergence criterion is met. The Bayesian approach also has the identification problem and the metric must be arbitrarily anchored. This can be accomplished by setting either $\bar{\theta}$ or $\bar{b}$ to zero.

Swaminathan and Gifford (1982) performed a simulation study to investigate the parameter recovery characteristics of their Bayesian estimation procedure for the Rasch model. Tests of 15, 25, 40, and 50 items were used with samples of $N = 20$ and $N = 50$. The values of ν_θ and ν_b were set to 5, 8, 15, and 25 with ξ_θ and ξ_b set to 10. The several datasets were analyzed using (1) a computer program implementing the Bayesian procedure and (2) the LOGIST program. The two approaches were compared with respect to the correlation of the estimates and the underlying values as well as the RMSE. The correlations of item difficulty were nearly identical for the two methods and had values ranging from .974 to .983. The correlations for θ showed the same pattern but were somewhat lower, ranging from .94 to .98. As would be expected, the RMSE values decreased as N and n increased. The RMSE values yielded by the Bayesian procedure for item difficulty were uniformly slightly smaller than those yielded by LOGIST. However, all values were less than .087. The RMSE values for θ were larger, ranging from .0384 to .2006. Except when tests of 15 items were used, the Bayesian procedure yielded somewhat smaller values of the RMSE. Overall the empirical results appeared to be insensitive to the values of ν_θ and ν_b employed, as the RMSE values generally were consistent across the values of ν.

One of the characteristics of the Bayesian IRT estimation procedure is that it tends to regress the noncentral parameter estimates toward the mean. In the case of both the ability and difficulty estimates,

the central values were very close to those yielded by JMLE. However, as the values departed from the mean, the Bayesian values were systematically smaller. This regression effect was clearly evident in a set of National Assessment of Educational Progress data analyzed by both methods in Swaminathan and Gifford (1982).

Swaminathan and Gifford (1985) also performed a simulation study using the Bayesian approach with a two-parameter logistic model. Samples of size $N = 50, 100, 200, 500$ and tests of length $n = 15, 25, 35$ were employed. Uniform priors on b_i and θ_j were employed and an informative prior, a normal approximation to the chi-square distribution, was used for the a_i. The metric identification was accomplished by setting $\bar{\theta} = 0$ and $\sigma_\theta^2 = 1$. The resulting datasets were also analyzed using the LOGIST program. The criterion variables used were the correlation of the estimates with the underlying parameter values and the mean square differences (MSD). They found that the Bayesian procedure yielded consistently higher correlations and smaller MSD than did LOGIST. In the LOGIST program, they specified an upper limit of 10 on the value of $\hat{a}_i$. When $n = 15$ all sample sizes yielded some estimates of $\hat{a}_i$ reaching this limit. A few discrepant values of $\hat{b}_i$ were observed in these datasets. Discrepant values were also observed in a few other cases when $n = 25$ and $N = 50$. The Bayesian procedure yielded no deviant values in these datasets. Thus, the priors arrested the parameter drift. As the number of items and examinees increased, the two approaches yielded similar results. Swaminathan and Gifford attributed this to the likelihood dominating the priors as the datasets become large. As this occurs, the results of the Bayesian and ML approaches become indistinguishable.

Assessment Systems Corporation (1986) has developed a microcomputer-based adaptive testing system called MICROCAT. The ASCAL test calibration part of the system uses what is termed a pseudo-Bayesian approach. The initial item parameter estimates for a and b are obtained using the Jensema (1976) approximations, with c set to the reciprocal of the number of response alternatives. Given these, Bayesian modal estimates of θ, using a normal prior, are obtained. The resulting θ distribution is divided into 20 fractiles and the mean fractile is used as the group value. The number of examinees and the number of correct responses in each group are the basic data in a MLE procedure having a Bayesian prior on the a and c parameters. In both cases, a beta distribution is employed. The parameters a, b, and c are then estimated on an item-by-item basis.

Vale and Gialluca (1985) compared the item parameter estimates yielded by ASCAL with those obtained using LOGIST for three sets of simulated data. Tests 1 and 2 differed with respect to the range of item difficulties employed; Test 3 was developed as part of an adaptive test item pool. Tests 1 and 2 had 50 items while Test 3 had 57 items. The simulated item response data ($N = 2{,}000$) were generated under a three-parameter model. Vale and Gialluca concluded that the ASCAL results were comparable with those yielded by LOGIST. Vale (personal communication, May, 1985) reported that ASCAL requires less than two hours to calibrate items on a 35-item test with 3,000 examinees using an IBM PC with an 8087 mathematics co-processor chip.

Lord (1984) examined the characteristics of the Bayesian approach from a statistical point of view. When the posterior mean is used as a parameter estimate, the overall mean squared error of estimation is minimized when an appropriate prior distribution is used. However, this is achieved by accepting an increased bias in the estimates. This bias is a result of the regression of the Bayesian estimates toward the mean. One cause of the regression of the parameter estimates toward the mean is the use of "tight" priors, such as those obtained from previous test administrations. As a result, a more diffuse prior is preferred. The Bayesian modal estimation procedures used by Swaminathan and Gifford (1982, 1985) do not minimize the mean square error unless the mean and the mode of the posterior distribution are the same. Lord (1984) indicated that they do not coincide in the case of IRT estimation problems. He also stated that the use of Bayesian priors has several practical advantages:

1. Infinite estimates of θ do not occur.

2. Item discrimination estimates will not become infinite.
3. The estimates of the guessing parameter, c, will not come out at implausible values, even when easy items are involved.

He concluded that Bayesian priors should probably be used for the a and c parameters, as regression toward the mean has less serious consequences here than for b and θ.

A major advantage of the Bayesian approach over the ML approach is that the former yields estimates for examinees with perfect and null raw scores as well as for items answered by all or by none of the examinees. It also appears to cope with the Heywood case by limiting the drift of the values of $\hat{a}$ toward infinity. However, the Bayesian procedures require assumptions about the prior distributions of the parameters. Unfortunately, these in turn require additional assumptions about the distributions of the parameters of the prior distributions. At present, it is not clear whether making appropriate assumptions and setting proper a priori values of these parameters will be difficult in practice.

Marginal Maximum Likelihood Estimation Procedures

One solution to the problem of incidental θ parameters is to replace them with their minimal sufficient statistics. The item parameters are then estimated conditional upon the sufficient statistics. While this conditional approach can be used with the Rasch model, it cannot be employed for the two- and three-parameter models. Because of this, Bock and Lieberman (1970) took a very different approach to the problem of incidental parameters, in which the θ parameters were assumed to be independently and identically distributed. The basic element in their approach was to estimate the structural (item) parameters by ML in the marginal distribution obtained by integrating over the distribution of the incidental (θ) parameters. From an analysis-of-variance perspective, Bock and Lieberman employed a mixed-effects model where items are fixed and examinees are random. The observed data employed consist of the pattern (vector) of correct and incorrect responses of an examinee to the n items of the test. Because there are 2^n possible patterns of response, the Bock and Lieberman approach is practical for only a few items, say, $n < 12$. Despite this practical limitation, this approach is very important from a theoretical point of view, as it was the initial conceptualization of what is now known as marginal maximum likelihood (MML) estimation.

Bock and Aitkin (1981) reformulated the Bock and Lieberman approach in a manner that is essentially equivalent to the EM algorithm (Dempster, Laird, & Rubin, 1977). Under this approach the examinees are assumed to be randomly sampled from a normal θ distribution denoted by $g(\theta)$. There are $\ell = 1, 2, \ldots, s$ distinct patterns of response to the n items, where $s = \min(N, 2^n)$. Then the probability of an examinee with trait level θ responding with pattern $u_\ell = (u_{\ell 1}, u_{\ell 2}, \ldots, u_{\ell n})$ is given by

$$L_\ell = P(u = u_\ell) = \int_{-\infty}^{\infty} P(u = u_\ell|\theta)g(\theta) \quad . \tag{40}$$

In order to evaluate the integral it is approximated by numerical quadrature and then

$$L_\ell = \sum_{k=1}^{q} P(u = u_\ell|X_k)\, A(X_k) \quad , \tag{41}$$

where X_k ($k = 1, \ldots, q$) is a quadrature point on the θ scale, and $A(X_k)$ is the quadrature weight. Both of these are obtained from Stroud and Secrest (1966, Table 5).

Recall that in the item stage of the JMLE procedure the examinees were grouped by θ. The basic data employed in the MLE of an item's parameters were the number of examinees in each group, f_j, and the number in each group responding correctly to the item, r_j. The key to the EM approach is that these two quantities are replaced by artificial data consisting of their expected values, which are finite-dimensional sufficient statistics.

The EM algorithm is a two-stage process based upon this likelihood. The first stage is the expectation (E)-step, in which the provisional values of the item parameters are used to compute for each item the "expected" number of examinees ($\bar{N}_k$) at each quadrature point and the number of these ($\bar{r}_{ik}$) correctly responding to an item. Essentially, what is done is to determine the part of each observed item response pattern's frequency that should be allocated to these artificial N_k and r_{ik} within each homogeneous quadrature (θ) group labeled by X_k. Then these frequencies are added within a given group to yield

$$\bar{r}_{ik} = \frac{\sum_{\ell=1}^{s} r_\ell u_{\ell i} \mathrm{L}_\ell(X_k)\mathrm{A}(X_k)}{\tilde{P}_\ell} \quad , \tag{42}$$

the expected frequency of correct response to item i at level k, and

$$\bar{N}_k = \frac{\sum_{\ell=1}^{s} r_\ell \mathrm{L}_\ell(X_k)\mathrm{A}(X_k)}{\tilde{P}_\ell} \quad , \tag{43}$$

the expected sample size at level k,

where $u_{\ell i}$ is the binary response to item i within pattern ℓ,
$\mathrm{L}_\ell(X_k)$ is the relative density at $\theta = X_k$,
$\mathrm{A}(X_k)$ is the quadrature coefficient, and

$$\tilde{P}_\ell = \sum_{k=1}^{q} \mathrm{L}_\ell(X_k)\mathrm{A}(X_k) \quad . \tag{44}$$

The second stage is the maximization (M)-step, in which the improved estimates of the item parameters are obtained by performing a conventional MLE logit analysis, such as is used in the item stage of the JMLE procedure. This is done using X_k as the independent variable with $\bar{r}_{ik}$ and $\bar{N}_k$ as the observed data. Using the slope/intercept parameterization of a logistic ICC model, the logit equations are

$$\sum_{k}^{q} [\bar{r}_{ik} - \bar{N}_k P_i(X_k)] = 0 \tag{45}$$

and

$$\sum_{k}^{q} [\bar{r}_{ik} - \bar{N}_k P_j(X_k)] X_k = 0 \quad . \tag{46}$$

Thus, the M-step consists of independent solutions for each of the n items, just as in the JMLE procedure.

The EM cycles are repeated until the estimates become stable to the required number of decimal places. Bock and Aitkin (1981) reported that the convergence is only geometric and slows up as the solution point is approached. They suggested using the acceleration technique of Ramsey (1975) to speed convergence. The convergence properties of the EM algorithm have been studied analytically by Wu (1983). He showed that if the likelihood function is unimodal and certain differentiability conditions are satisfied, any EM sequence converges to the unique ML estimates of the parameters.

The remaining feature of the MML approach is that the normal prior distribution of θ can be replaced by some empirically defined distribution of the examinees over the θ scale. Thus, rather than integrating over the normal density, the quadrature procedures are applied to the empirical distribution. The remainder of the process is unchanged.

Once the item parameters have been estimated, they are considered as known and the examinee's θ levels can be estimated as a separate process. Bock and Aitkin (1981) employed either ML, Bayesian maximum a posteriori estimation (MAP), or Bayesian expected a posteriori (EAP) estimation. Each has its own advantages and disadvantages.

According to Swaminathan and Gifford (1985), "While marginal maximum likelihood estimators are superior to joint maximum likelihood estimators of item parameters, at least in small samples, they do not offer protection from Heywood type cases where inadmissible estimates of the discrimination parameters are obtained" (p. 350).

The marginal maximum likelihood estimation (MMLE) procedures of Bock and Aitkin (1981) have been implemented in the BILOG computer program (Mislevy & Bock, 1982, 1984, 1985). The program employs logistic ICC models. It also enables imposition of Bayesian prior distributions on the item parameters. Mislevy and Bock (1984) discussed the use of such priors in terms of dealing with the problem of multi-colinearity among the item parameter estimates. The estimates then are Bayesian modal rather than MLE. (See Mislevy, 1986b, for an in-depth discussion of such Bayesian modal estimators.) The program manual contains a very concise, lucid comparison of the JMLE, CML, and MML approaches to item parameter estimation. Correct use of BILOG requires knowledge of the very large number of options available.

Yen (1985) compared the results yielded by BILOG and LOGIST. Simulated data were generated under a three-parameter model for 1,000 examinees and four 20-item and four 40-item tests. In three of the tests, in each set, the values of a and c were fixed and the values of b varied. The fourth test was modeled after an existing reading vocabulary test. When BILOG employed MLE of θ, LOGIST was about 25% faster, in terms of CPU time, than BILOG for the 20-item tests and about the same for the 40-item tests. When BILOG employed Bayesian EAP estimates of ability, the two programs took about the same amount of CPU time for the 20-item tests and BILOG was about 40% faster for the 40-item tests. When options in BILOG for allowing omits to be given partial credit were used, it took 62% to 94% more time than when omits were ignored. Hence, BILOG was considerably slower than LOGIST in this situation.

The item parameter estimates yielded by BILOG were generally more accurate estimates of the underlying parameter values than those yielded by LOGIST. Yen (1985) attributed this to the procedures used in LOGIST to estimate the c parameter, which result in a positive correlation between the true b and the amount of error in the estimates of b and c. When BILOG employed MLE of θ, the ICCs defined by the item parameter estimates were nearly identical with those yielded by LOGIST. Other than execution time when omits are analyzed, BILOG and LOGIST generally appeared to behave comparably.

Tsutakawa (1984) derived a MML procedure employing the two-parameter logistic ICC model, which he labeled the MLF procedure (ML for maximum likelihood and F because some of the linear effects are treated as unknown fixed constants). His method differs in the manner in which the prior distribution of θ is handled, but for the special case of a discrete empirical prior it is equivalent to that of Bock and Aitkin (1981). Tsutakawa analyzed a 50-item arthritis knowledge test administered to 162 examinees, using both his MLF procedure and LOGIST. After appropriate rescaling to take metric differences into account, the values of the discrimination and difficulty parameters yielded by the two methods were very similar. However, the MLF procedure tended to yield fewer extreme values of the difficulty parameter estimates.

Tsutakawa (1984) also used simulated data to evaluate the parameter recovery capabilities of the two procedures. Two hundred simulated examinees having a unit normal distribution and a 50-item test with representative values of a and b were used. The simulated item response data were analyzed by both LOGIST and the MLF procedure. The estimated item parameters were plotted against the underlying parameter values. The plots showed a close agreement between the two methods as well as a general 45° line relating the estimates and the parameters. The scatter of the as about the line was much greater than that of the bs. For these data at least, the results yielded by LOGIST and MLF agreed quite well. Again, no data were provided as to absolute costs of performing the analyses under the two approaches.

Thissen (1982) derived a MMLE procedure for a modified one-parameter model. Rather than fixing

all the a_i at unity, he assumed that the items shared a common but unknown value of a. This parameter was then estimated along with the n item difficulties. He reanalyzed the Law School Aptitude Test Scale 6 data and found that his MML estimates agreed with the CML estimates to within .002. However, the MML solution required 27 EM cycles even when Ramsey's (1975) acceleration techniques were employed. Thissen reported that the amount of processing time was 65% of that required to analyze the items using the Bock and Lieberman (1970) approach. However, the cost was not reported and it appears that his MML procedure converges quite slowly.

Rigdon and Tsutakawa (1983) also derived a MLF procedure for the estimation of item difficulty under the Rasch model, employing an extended form of the EM algorithm (Dempster, Rubin, & Tsutakawa, 1981) in which the item difficulty parameters were considered a fixed effect and the θ parameters were considered a random effect. They also developed a modified procedure where the posterior mean of each examinee's θ is used in the item estimation process, called the CMLF procedure (where C denotes conditional). In these two procedures, the item parameters, b_i, and the variance of the examinees' θ distribution are estimated using the EM algorithm.

A set of item response data from an administration of the Scholastic Aptitude Test (1,000 examinees and 20 items) was analyzed using both the MLF and CMLF approaches. This dataset had been previously analyzed by Andersen (1970) using CML. The estimated item difficulties yielded by the three methods agreed to within .05 for 18 items and to within .12 for the remaining 2 items. Simulated responses of 50 and 200 examinees with normal and uniform distributions to 50 items were also generated and analyzed. Again, very close agreement was observed for the mean values of the item difficulties under the three methods. The mean square difference of the estimates yielded by the MLF and CMLF procedures was about 19% smaller than for CML.

It should be noted that in the MML procedures based upon Bock and Aitkin's approach, the estimation of the item parameters is completed before the θ parameters are estimated. Under the other approaches, a joint estimation procedure is used in which the item and θ parameters are estimated in stages, as in JMLE.

Discussion

Because both the LOGIST and BICAL computer programs have been available for many years, the characteristics of the JMLE procedure have been thoroughly investigated. The results of the many simulation studies and test analyses exhibit a considerable level of consistency with respect to item parameter recovery, goodness of fit to the empirical data, bias, and standard errors. As a result, the general properties of the item parameter estimates yielded by the JMLE paradigm have been reasonably well established.

However, JMLE has some properties that are less than optimal. The literature suggests that the properties of the item parameter estimates for one- and two-parameter ICC models are better than those for a three-parameter model. Both the theoretical and empirical standard errors of the item parameter estimates are smaller under these two models than under a three-parameter model. This is particularly true for noncentral values of the parameters where the asymptotic standard errors under a three-parameter model are dramatically larger.

The empirical ICC data of McKinley and Reckase (1980) show that the variability of the observed proportions of correct response around the fitted ICC is much larger in the lower tail than in other sections of the ICC. This would suggest that for items meeting Lord's $b - 2/a$ rule, the c parameter would be poorly estimated. The empirical results confirm this, as this parameter is poorly estimated in terms of both bias and standard error and the estimates do not correlate well with the underlying values in simulation studies. Troubles in the estimation of c tend to carry over into the estimation of b and a. In particular, the estimation of difficulty is affected as error in c results in a shift in $\hat{b}$. The carryover is also evident in the larger standard errors of $\hat{a}$ and $\hat{b}$.

Because of the problems inherent in the simultaneous estimation of θ and item parameters under a three-parameter model, the LOGIST program imposed a complex set of constraints and procedures upon the JMLE paradigm. Thissen and Wainer (1982) suggested that what Wingersky and Lord had done was to impose informal priors upon the parameter distributions.

While much of the problem stems from the manner in which the c parameter appears in the model, the use of the $b - 2/a$ rule in the LOGIST program also appears to be a factor. The effect of applying this rule is that a varying proportion of the items is arbitrarily assigned the mean c value of the difficult items where c was successfully estimated. This in turn forces the estimates of b and a for the easy items to conform to this value of c rather than the actual value of c present in the data. In addition, there does not seem to be any evidence that the average level of guessing is the same for easy and difficult items. A common thread in the IRT literature is that the characteristics of the estimates under the three-parameter model leave much to be desired; for this reason, it would appear that the fundamental problem is an interaction of the model with the estimation process. Therefore, some effort should be devoted to developing a new model to cope with the issue of guessing. Only limited efforts have been made in this regard (e.g., Lord, 1983b) and none have developed into serious contenders.

Because of the problem of estimating structural (item) parameters in the presence of incidental (θ) parameters, considerable discussion in the literature has focused on the role of sufficient estimators. However, the empirical results for the Rasch model suggest that the item parameter estimates yielded by the JMLE procedure do not differ materially from those in which the estimation procedure is conditioned upon the sufficient statistics for θ. In addition, other simulation studies suggest that the item parameter estimates yielded under the two- and three-parameter models are empirically consistent. These same studies show that the JMLE procedure recovers the underlying parameters very well; the plots of the estimates against the parameter values generally show a 45° linear relationship with an acceptable degree of scatter around the line. In addition, $r_{\hat{b}b}$ is uniformly in the high .90s under all three models. The value of $r_{\hat{a}a}$ is generally in the upper .80s or low .90s under two- and three-parameter models. However, $r_{\hat{c}c}$ is usually quite low, ranging from the low .20s to the low .60s. It should be noted that with the exception of Yen (1985), the parameter recovery studies have neglected to equate the metric of the obtained item parameter estimates to the underlying metric used to generate the data. Although this will have no impact upon the correlations, it could have a significant impact upon other measures of agreement.

The two-stage JMLE paradigm involves a feedback loop between the estimates of item discrimination and θ. In certain datasets, large values of a lead to large values of $\hat{\theta}$. In the extreme case, $\hat{a}$ becomes infinite and the JMLE procedure fails. In the manual for the first publicly available version of the LOGIST computer program (Wingersky & Lord, 1973), this phenomenon was referred to as drift of the θ metric. In part due to Wright's persistent criticism, it has recently been recognized that this problem is the Heywood case of factor analysis. Although the case does not arise in all datasets, its mere existence requires that the computer programs for the two- and three-parameter models cope with it. In LOGIST, an upper limit on the values of $\hat{a}$ was imposed to control this loop. In the Rasch model the loop is controlled by the fact that all as have a fixed value of unity. In Bayesian methods, the loop is controlled by using prior distributions on the parameters. The MMLE approaches are also susceptible to this problem. At the present time, it is unclear which characteristics of the item response dataset lead to the Heywood case.

In the item stage of Birnbaum's (1968) paradigm for JMLE of the item and examinee parameters, parameter estimation for a given item is identical to the procedure used in quantal response bioassay. The basic process is one of fitting a monotonic response curve to the proportions of correct response, given a known θ (dosage) scale. In addition, the ML procedure employed is essentially a weighted least-squares technique where the weights, under a logistic model, are $f_j P_i(\theta_j) Q_i(\theta_j)$. This is rarely recognized in the IRT literature, even though it accounts for many of the characteristics of the item parameter estimates.

For example, the bias and standard errors of the item parameter estimates are minimized when the distribution of the examinees over the θ scale is uniform and when the item difficulty matches the mean θ of the examinees. When a non-uniform θ distribution is used, an interaction occurs between the number of examinees at each θ level and the estimated parameters of the ICC. This interaction is sensitive to both the form of the frequency distribution of θ and the location of the item relative to the θ distribution. The empirical results under all three ICC models exhibit this interaction.

The primary motivation for developing the Bayesian approach to item parameter estimation was to deal with the issue of estimating structural parameters in the presence of incidental parameters. Although only a few empirical results are available, they suggest that the numerical values of the obtained estimates are comparable to those yielded by the JMLE procedure. However, the Bayesian estimates exhibit a regression toward the mean not present in other estimation methods. The major advantage of the Bayesian approach is that the use of prior distributions on the parameters tends to inhibit obtaining unusual values for the estimates, thus providing a way to cope with the Heywood case problem. Also, estimates can be obtained even when an item has been correctly answered by all or by none of the examinees. However, the need to specify distributions of parameters for prior distributions of θ and item indices is a serious drawback to the method. It appears doubtful that the usual practitioner would be able to specify such distributions without an extensive set of guidelines. Perhaps a technique could be developed where these specifications would be determined from the context and integrated into a computer program. The lack of a widely available computer program for the Bayesian approach is a limiting factor on the examination of the method's properties. It should be noted that an option in the BILOG program implements a Bayesian modal estimation of the item parameters.

The development of the marginal maximum likelihood estimation procedure also was motivated by the structural/incidental parameters problem. The standard MMLE approach is based upon the EM algorithm. In sharp contrast to the JMLE procedure, item parameter estimation and θ estimation are treated as separate processes under MMLE. Item parameter estimation is achieved using the same standard bioassay procedure as in the JMLE approach. Under JMLE, the observed number of examinees and the number of correct responses at each θ level are the basic data. Under MMLE, a technique is used in the E stage to essentially distribute a given examinee's θ and response to an item over the θ scale. Then the aggregate expected number of examinees and expected number of correct responses at each θ level (quadrature point) are obtained from this "artificial data." This allows estimation of the item parameters in the M stage by means of the usual bioassay MLE process without actually having a θ score for each examinee. The net result is that the parameters of the items have been estimated and a metric for the θ scale established. In the θ estimation stage, the item parameters are considered known, and either maximum likelihood or Bayesian procedures are employed to obtain the examinees' θ estimates.

The initial empirical results suggest that the item parameter estimates obtained from MMLE seem to be similar to those yielded by the JMLE approach. However, the existing results are limited and it is not possible to draw a definitive conclusion as to the characteristics of the estimates yielded by the MMLE approach. A microcomputer version of BILOG (Mislevy & Bock, 1986) is available and further studies of the MMLE approach can now be accomplished more readily.

In the early stages of the development of IRT, there was considerable interest in using the classical theory approximations to the item parameters developed by Lawley, Tucker, and Brogden. The computational advantage over the JMLE approach was considerable and in those days computer time was very expensive. This approach is best exemplified by Urry's (1974, 1976) computer programs. In these programs, a multistage procedure based upon these approximations was used to obtain the item parameter estimates. A Bayesian procedure was then used to estimate θ for an examinee. The overall paradigm is structurally similar to the MMLE approach. The empirical data suggest that when long tests and large sample sizes are employed, the obtained item parameter estimates are very similar to those yielded by

JMLE. The approach probably capitalizes upon the fact that as sample size increases, the raw score becomes a reasonable estimator of the true score which has a monotone relation to θ. As a result, the large sample estimates of the item-θ correlations improve. The major attraction of this approach is its rather minimal computational demands as compared to all of the other paradigms. Because of this, it might be useful to reexamine this approach to determine whether microcomputer implementation could yield acceptable parameter estimates.

Conditional maximum likelihood estimation can only be employed with the Rasch model where sufficient statistics for the examinees' θs exist. It estimates the item difficulties only after having replaced the examinees' θs by their sufficient statistics, the raw test scores. Although it is possible to estimate θ conditional upon the sufficient statistics for the item difficulties, this is not done in practice. The major theoretical advantage of CML is that the item parameter estimates are consistent. However, the empirical results indicate that the values of the obtained difficulty estimates differ only slightly from those yielded by the JMLE approach. Even with recent advances in techniques for computing the symmetric functions, the computational demands of this approach are formidable. Thus, the issue reduces to one of whether the theoretical advantages outweigh the computational demands. With the shift toward the use of microcomputers, the routine use of the conditional approach appears to be unlikely.

The item parameter estimation issue is inextricably interlinked with the computer software that implements the estimation procedures. The characteristics of the obtained item parameter estimates are critically dependent upon the manner in which the underlying mathematics are implemented in the software. In any program, the computer programmer makes numerous decisions resulting from implementation choices rather than from the underlying mathematics. Unfortunately, these decisions usually are not apparent to the user of the program. For example, the BICAL program divides the item difficulty by 2 when its value becomes excessive. In the LOGIST program, it was necessary to impose severe constraints upon the incremental values of the parameter estimates for all items to cope with the problems associated with the estimation of c and the Heywood cases. The more complex the computer program and the smaller the computer, the larger the number of such decisions.

Also, as programs get more complex, the number of analysis options increases. Both LOGIST and BILOG have a very large number of options that influence the manner in which the analyses are performed, and hence the results. However, it has not been standard practice to report the pattern of options employed in a given study. As a result, it is not always possible to determine the exact program configuration used. This in turn makes comparisons across studies difficult. It would appear that each published article should specify the options used in each computer program employed and the metric of the parameter estimates.

Many articles dealing with IRT parameter estimation employ computer programs that have not been released for general use (see, e.g., Swaminathan & Gifford, 1982, 1985; Thissen, 1982; Tsutakawa, 1984). These estimation programs often have been written for a specific study and do not have the level of development and documentation appropriate for general distribution. As a result, the reader has a limited basis for judging the validity of the software implementation. In addition, most studies based upon such software fail to report any cost data. The lack of such data makes it difficult to decide if a theoretically promising approach is economically viable.

Despite its inherent problems and the availability of alternative methods, the JMLE approach, suitably constrained, appears to be the de facto standard item parameter estimation technique. Consequently, the examination of alternative estimation procedures typically involves comparison of the results with those under the JMLE approach. The current, but limited, results suggest that each of the alternative methods has some specific advantage over JMLE. However, the advantage in terms of the estimates produced is not overwhelming and not sufficient to cause total abandonment of the JMLE approach in favor of one of the alternatives.

Because most of the alternative approaches are relatively new, there is a need to explore their

characteristics to the depth and breadth that the JMLE paradigm has been investigated. In addition, the alternative procedures employ the Newton-Raphson technique at some point in their paradigms to linearize the solution equations. Thus there is a common procedural thread that runs through all the available estimation methods. As a result, they share the limitations, such as the need for good initial estimators and well-conditioned matrices, of the Newton-Raphson technique. This aspect of the item parameter estimation procedures has received very little attention. In particular, the interaction of the three-parameter model and the Newton-Raphson procedure needs to be examined.

A salient feature of the item parameter estimation techniques discussed above is that they are based upon asymptotic statistical theory. As a result, they require rather large datasets to produce acceptable item parameter estimates. This basis is suitable for those with access to large datasets, such as commercial testing organizations. However, it poses a problem for routine test analysis at the classroom and similar levels. The empirical results cited above show that the asymptotic procedures can yield parameter estimates with large biases and large standard errors of estimate for datasets of the size typically encountered at this level. As a result, there is little motivation for teachers and others to make the transition from classical test theory to item response theory. Thus, there is a need for development of small-sample approaches to item parameter estimation. Such techniques should be able to cope with the vagaries of small datasets and yield estimates with acceptable sampling characteristics. Implementation of such procedures in a "user-friendly" manner on a microcomputer is an absolute requirement.

References

Abbott, W. S. (1925). A method of computing the effectiveness for insecticide. *Journal of Economic Entomology, 18,* 265–267.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B, 32,* 283–301.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 34,* 42–54.

Andersen, E. B. (1973a). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31–44.

Andersen, E. B. (1973b). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140.

Assessment Systems Corporation. (1986). *MICROCAT: A computer program for computerized adaptive testing.* St. Paul MN: Author.

Baker, F. B. (1959). Univac scientific computer program for test scoring and item analysis [Computer program abstract]. *Behavioral Science, 4,* 254–255.

Baker, F. B. (1965). Origins of the item parameters X_{50} and β as a modern item analysis technique. *Journal of Educational Measurement, 2,* 167–180.

Baker, F. B. (1967). The effect of criterion score grouping upon item parameter estimation. *British Journal of Mathematical and Statistical Psychology, 20,* 227–238.

Baker, F. B. (1984). Ability metric transformations involved in vertical equating under item response theory. *Applied Psychological Measurement, 8,* 261–271.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association, 39,* 357–365.

Berkson, J. (1955). Maximum likelihood and minimum chi-square estimates of the logistic function. *Journal of the American Statistical Association, 50,* 120–162.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–472). Reading MA: Addison-Wesley.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6,* 258–276.

Bock, R. D. (1972). Estimating item parameters and latent ability when 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: An application of an EM algorithm. *Psychometrika, 46,* 443–459.

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

Brogden, H. E. (1971). *Latent ability and the structure of tests.* (Mimeographed). West Lafayette IN: Purdue University.

de Gruijter, D. N. M. (1984). A comment on some standard errors in item response theory. *Psychometrika, 49,* 269–272.

de Gruijter, D. N. M. (1985). A note on the asymptotic variance-covariance matrix of item parameter estimates in the Rasch model. *Psychometrika, 50,* 247–249.

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.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance component models. *Journal of the American Statistical Association, 76,* 341–353.

Finney, D. J. (1952). *Probit analysis: A statistical treatment of the sigmoid response curve*. Cambridge, England: Cambridge University Press.

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika, 46,* 59–77.

Gugel, J. F., Schmidt, F. L., & Urry, V. W. (1976). Effectiveness of the ancillary estimation procedure. In C. L. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing* (pp. 103–106). Washington DC: Personnel Research and Development Center, U.S. Civil Service Commission.

Gumbel, E. J. (1961). Bivariate logistic distributions. *Journal of the American Statistical Association, 56,* 335–349.

Gustafsson, J. E. (1980a). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement, 40,* 377–385.

Gustafsson, J. E. (1980b). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 33,* 205–233.

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

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error* (Technical Report No. 15). Stanford CA: Stanford University, Applied Mathematics and Statistics Laboratory.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79–92.

Householder, A. S. (1953). *Principles of numerical analysis*. New York: McGraw-Hill.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6,* 249–260.

Jensema, C. J. (1976). A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement, 36,* 705–715.

Kale, B. K. (1962). On the solution of likelihood equations by iteration processes: The multiparameter case. *Biometrika, 49,* 479–486.

Kearns, J., & Meredith, W. (1975). Methods for evaluating empirical Bayes point estimates of latent trait scores. *Psychometrika, 40,* 373–394.

Kendall, M. G., & Stuart, A. (1961). *The advanced theory of statistics* (Vol. 2). New York: Hafner.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887–906.

Kolen, J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18,* 1–11.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61A,* 273–287.

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

Lord, F. M. (1965). A note on the normal ogive or logistic curve in item analysis. *Psychometrika, 30,* 371–372.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39,* 247–264.

Lord, F. M. (1975). *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (Research Bulletin RB-75-33). Princeton NJ: Educational Testing Service.

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

Lord, F. M. (1983a). Statistical bias in maximum likelihood estimators of item parameters. *Psychometrika, 48,* 425–435.

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

Lord, F. M. (1984). *Maximum likelihood and Bayesian parameter estimation in item response theory* (Research Report RR-84-30-ONR). Princeton NJ: Educational Testing Service.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1985). Sampling variances and covariances of parameter estimates in item response theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 69–88). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Maxwell, A. E. (1959). Maximum likelihood estimates of item parameters using the logistic function. *Psychometrika, 24,* 221–227.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9,* 49–57.

McKinley, R. L., & Reckase, M. D. (1980). *A comparison of the ANCILLES and LOGIST parameter estimation procedure for the three-parameter logistic model using goodness of fit as a criterion* (Research Report 80-2). Columbia MO: University of Missouri Tailored Testing Laboratory.

Mislevy, R. J. (1986a). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3–31.

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

Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items*. Chicago: International Educational Services.

Mislevy, R. J., & Bock, R. D. (1984). *BILOG I maximum likelihood item analysis and test scoring: Logistic model*. Mooresville IN: Scientific Software.

Mislevy, R. J., & Bock, R. D. (1985). Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 189–202). Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software Inc.

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48,* 49–72.

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

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Bulletin 69–92). Princeton NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351–356.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Ree, M. J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement, 3,* 371–385.

Rigdon, S. E., & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567–574.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–435.

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

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38,* 221–233.

Sanathanan, L. (1974). Some properties of the logistic model for dichotomous response. *Journal of the American Statistical Association, 69,* 744–749.

Scheiblechner, H. H. (1971). *A simple algorithm for CML parameter estimation in Rasch's probabilistic measurement model with two or more categories of answers* (Research Bulletin No. 5). Vienna: Psychologisches Institut der Universität Wien.

Schmidt, F. L. (1977). The Urry method of approximating the item parameters of latent trait theory. *Educational and Psychological Measurement, 37,* 613–620.

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

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7,* 175–191.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 13–30). New York: Academic Press.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two parameter logistic model. *Psychometrika, 50,* 349–364.

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

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397–412.

Thomson, G. H. (1919). A direct deduction of the constant process used in the method of right and wrong cases. *Psychological Review, 26,* 454–464.

Tsutakawa, R. K. (1984). Estimation of two-parameter logistic item response curves. *Journal of Educational Statistics, 9,* 263–276.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11,* 1–13.

Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement, 34,* 253–269.

Urry, V. W. (1976). Ancillary estimators for the item parameters of mental tests. In W. A. Gorham (Chair), *Computers and testing: Steps towards the inevitable*

conquest (PS-7C-1) (pp. 14–18). Washington DC: Personnel Research and Development Center, U. S. Civil Service Commission.

Vale, C. D., & Gialluca, K. A. (1985). *ASCAL: A microcomputer program for estimating logistic IRT item parameters*. St. Paul MN: Assessment Systems Corporation.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123–140.

Wainer, H., Morgan, A., & Gustafsson, J. E. (1980). A review of estimation procedures for the Rasch model with an eye toward longish tests. *Journal of Educational Statistics, 5,* 35–64.

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* (pp. 45–56). 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. (1973). *A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses* (Research Memorandum RM-73-2). 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.

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

Wright, B. D., & Douglas, G. A. (1977a). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1,* 281–295.

Wright, B. D., & Douglas, G. A. (1977b). Conditional versus unconditional procedures for sample-free analysis. *Educational and Psychological Measurement, 37,* 573–586.

Wright, B. D., & Linacre, J. M. (1984). *MICROSCALE* [Computer program]. Boston MA: Mediax Interactive Technologies.

Wright, B. D., & Mead, R. J. (1978). *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23A). Chicago: University of Chicago, Statistical Laboratory.

Wright, B. D., Mead, R. J., & Bell, S. R. (1980). *BICAL: Calibrating items with the Rasch model* (Research Memorandum No. 23C). Chicago: University of Chicago, Statistical Laboratory, Department of Education.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23–48.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics, 3,* 95–103.

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 Psychometric Society Meetings, Nashville TN.

Author's Address

Send requests for further information to Frank B. Baker, University of Wisconsin, Department of Educational Psychology, 1025 West Johnson Street, Madison WI 53706, U.S.A.

Reprints

Reprints of this article may be obtained *prepaid* for $2.50 (U.S. delivery) or $3.00 (outside U.S.; payment in U.S. funds drawn on a U.S. bank) from Applied Psychological Measurement, N657 Elliott Hall, University of Minnesota, Minneapolis MN 55455, U.S.A.