The Effect of NonNormal Ability Distributions on IRT Parameter Estimation Using Full-Information and Limited-Information Methods

J. R. Boulet

Faculty of Education

Thesis submitted to
the school of Graduate Studies and Research
in partial fulfilment of the requirements for the
PhD degree in Education

University of Ottawa

ISBN 0-612-15701-6

Canada

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ACKNOWLEDGEMENT

The author would like to thank the following people for making this dissertation possible: Cathy Lumsden-Boulet for her generous support and endless patience throughout the many years of this project; Dr. Marc Gessaroli who laboured through numerous drafts and never ran out encouragement, jokes, or red ink; my close friend Brian Bennett who periodically professed that he was looking forward to making an appointment (golf) with his doctor; Dr. Marvin Boss for his kindness, wisdom, and uncanny ability to spot improper use of commas; Dr. Bruno Zumbo for his statistical expertise, constructive criticism, and policy of immediate "open door" consultations; my family who finally realized that "are you done yet?" does little to expedite the process; and lastly, my children Nathalie and Jenna for their morning smiles and evening kisses.

ABSTRACT

The relationship between nonlinear factor analysis (FA) models and Item Response Theory (IRT) models has been well established. Furthermore, in terms of modern measurement theory, the use of nonlinear FA models to describe item-trait relationships is currently becoming more popular and may offer some statistical and/or computational advantages in the analysis of item response data. Both limited-information (LI) and full-information (FI) nonlinear FA models can be used to derive the familiar IRT parameter estimates. In general, the two approaches (LI and FI) are distinguished simply by the extent to which they use information in the data matrix of examinee (subject) responses.

The focus of this study was to compare the accuracy and efficiency of IRT parameter estimates (i.e., item difficulty, item discrimination) using both LI and FI nonlinear FA models. A Monte Carlo study was employed to investigate the precision and stability of parameter estimates in situations where a) the manifest variables (test items) are binary and there is a single underlying normally distributed latent variable and b) the manifest variables are binary and there is a single underlying latent variable that is not normally distributed. In addition, parameter recovery was explored under various simulated test lengths (number of items) and sample sizes (number of examinees).

The results of the study suggest that, for conditions involving a normally distributed latent variable, the limited-information approach incorporated in the NOHARM computer program generally provides more accurate and stable parameter estimates than the theoretically preferred FI estimator incorporated in the TESTFACT computer program. For situations involving a nonnormal distribution of the latent trait, or ability, FI estimation provided a marginally better calibration of the 2-parameter logistic response model. Both estimators were, however, prone to producing item values that were outside of feasible ranges, resulting in poor goodness-of-fit of the estimates. Furthermore, based on the conditions modelled in the study, neither the sample size, the test length, nor the sample size/test length ratio were important in terms of explaining between-program differences in the recovery of the item parameters.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I: INTRODUCTION

The relationship between factor analysis (FA) models and Item Response Theory (IRT) models is well documented (e.g., De Champlain, 1995; McDonald, 1967, 1982; Takane & De Leeuw, 1987). More importantly, in terms of modern measurement theory, the use of FA models to describe item-trait relationships is currently becoming more popular and may be advantageous in the analysis of item response data. The increased recognition of the basic unity between nonlinear FA models and IRT is a significant development and provides several interesting areas for research in measurement and testing. For example, even though there are numerous studies in which attempts have been made to evaluate competing estimation strategies based on traditional IRT methods (see, Baker, 1987), there has been little research aimed at establishing the appropriateness of parameter estimates derived from competing nonlinear FA models (i.e., limited-information versus full-information). A quantification of the relative accuracy and stability of the estimates produced under these two specifications will be of great interest to testing specialists.

Both limited-information (LI) and full-information (FI) nonlinear FA models can be used to derive the familiar IRT parameter estimates. The essential difference between LI and FI models can be defined by what McDonald (1981) calls the strong and weak principles of local independence. The strong principle states that given $\theta_1, \ldots, \theta_k$, the conditional probability of any

set of the n variables (items) is the product of the n conditional probabilities of those values. Under this specification, responses to some items can be dependent on responses to two or more other items. The latent trait not only explains item covariances but also higher order joint moments. Estimation strategies such as marginal maximum likelihood (MML) in TESTFACT (Wilson, Wood, & Gibbons, 1991), which use all of the information in the response patterns, employ the strong principle. The weak principle, which is probably more tenable in practice, states that the conditional covariances of the items are zero. From a testing perspective, one assumes the existence of one or more latent traits characterizing each examinee such that, for any fixed value of these, the responses of the examinees to the binary items are mutually statistically independent. For the unweighted least-squares (ULS) estimation in NOHARM (Fraser & McDonald, 1988), which is based only on item covariances, the weak principle is employed. In general, the two models (LI and FI) are distinguished simply by the extent to which they use information in the data matrix. However, within the LI and FI frameworks, there are a number of estimation procedures and options that can be employed.

For traditional IRT based programs (e.g., LOGIST; Wingersky, Barton & Lord, 1982), nonnormal distributions of examinee (subject) ability have been shown to result in poor parameter estimates (Yen, 1987). Under both of the nonlinear FA models mentioned above, for conditions where a single trait is presumed

to underlie test performance, it is assumed that the distribution of this latent trait (e.g., ability) is normal. When the latent trait is not normally distributed the range of unobserved abilities for the examinees can become more homogeneous. Hambleton (1993) suggests that under these conditions the item parameter estimates will be less stable. Unfortunately, little is known regarding the possible effect on NLFA parameter estimates of violating the normality assumption. Hence, there appears to be a need for an empirical study, using competing nonlinear FA approaches, that specifically addresses this issue.

The focus of this study was to compare the accuracy and stability of IRT parameter estimates using both FI and LI nonlinear FA models under both normal and nonnormal distributions of some latent trait, say ability. While there are numerous IRT, or NLFA, representations (see McDonald, 1982), the emphasis in this investigation was on parameter recovery for a unidimensional, nonlinear, dichotomous response model. In the second chapter a review of the common factor model and its potential application with dichotomous (i.e., 0/1) data is introduced. This is followed by a brief overview of the competing nonlinear FA models and associated estimation strategies. An examination of the relevant research related to IRT and FA is introduced next. The methods and Monte Carlo design are outlined in Chapter III. The results are presented in Chapter IV. Chapter V is devoted primarily to relating the methods, scope and educational implications of the results of the

research topic. The study conclusions and potential directions for future research are discussed in Chapter VI.

CHAPTER II: LITERATURE REVIEW

Issues related to the specification and use of NLFA are discussed in this Chapter. This includes an outline of the operationalization of various models and a description of some of the possible parameter estimation methods. In addition, a review of the relevant parameter recovery studies, using both LI and FI estimators, is presented. Finally, a specific explanation and rationalization of the research problem is forwarded.

Factor Analysis Models

An overview of the common factor model is presented below. The general utility of FA techniques to describe item-trait relationships is discussed, with a particular emphasis on situations where the manifest variables (test items) are dichotomously scored. This is followed by a brief overview of competing NLFA models and associated estimation strategies.

Common Factor Model

Factor analysis can be used to reveal relationships between observed and hypothetical (latent) variables (traits). If a large number of variables show substantial mutual correlation, it is reasonable to suppose that this arises from their common dependence on a smaller set of unobserved or latent variables.

The purpose of FA is to describe such a set of multivariate data in terms of one or more of these underlying variables or traits. Interrelationships among variables can expressed by a number of coefficients (e.g., Pearson Product Moment correlations (PPMC's), polychoric correlations, tetrachoric correlations), depending on the measurement scale of the items (e.g., interval, ordinal, or dichotomous). There are a number of estimation strategies that can be incorporated in FA models. These include maximum likelihood (ML), unweighted least squares (ULS), generalized (or weighted) least squares (GLS), and many associated variants.

The most common FA strategy is to use Pearson Product Moment correlations (PPMC's) and ML estimation. This requires the assumption that the observed variables (e.g., items) are measured on at least an interval scale and that the data (e.g., item responses) follow a multivariate normal distribution. In addition, it is assumed that the regressions between the observed variables and the latent (unobserved) factors/traits (e.g., ability) are linear. Furthermore, the latent traits are assumed to have a mean of zero.

The Spearman common factor model can be described as follows. Assume that the responses to j variables (items) for N subjects can be explained by k factors. The general factor model is

$$Y = \theta\Lambda' + e, \qquad (1)$$

where

Y	represents the N x j matrix of observed scores,
θ	represents the N x k matrix of latent variable(s),
Λ'	represents the k x j matrix of coefficients describing the regression of items on the latent variables (factors), and
e	represents the N x j matrix of residuals.

The variance-covariance matrix of the observed variables, denoted as Σ, can be expressed as:

$$\Sigma = \Lambda\Phi\Lambda' + \Psi \qquad (2)$$

where

Φ	is the k x k covariance matrix of θ, and
Ψ	is the j x j covariance matrix of the residuals.

Nonlinear Factor Analysis

It has been shown that the general factor analysis model described above is not appropriate when the observed scores are dichotomously scored (i.e., 0/1). The underlying problem is that when the dichotomous variables are bounded by 0/1, the assumption of linearity of regression between the latent factor/trait and the observed dichotomous variable(s) will rarely be met (McDonald, 1967a; McDonald & Ahlawat, 1974). In addition, Olsson (1979) showed that the application of FA to discrete data may lead to incorrect conclusions regarding the number of factors, biased estimates of the factor loadings, and incorrect standard errors and χ^2 tests of model fit. From a testing perspective, a

linear model would be expected to perform poorly in accounting for item-ability relationships. For example, the relationship between item performance and ability would clearly be nonlinear for high ability candidates. While raw PPMs (i.e., phi correlations) are easily obtained from the current statistical packages, their use in the FA of dichotomously scored responses does not appear to be appropriate.

It has been suggested that ULS estimation of the sample matrix of tetrachoric correlation coefficients (approximation of the correlation matrix among hypothetical multivariate normal latent response processes for the various items) should be employed. This technique may be superior to the FA of phi coefficients and generally requires far less computation than other methods specifically designed for categorical data (e.g., nonparametric techniques). Rigdon and Ferguson (1991) suggest that, if one simply needs parameter estimates such as factor loadings and uniquenesses, one can use tetrachoric correlations and binary data as a substitute for the sample covariances and multivariate normal data. Unfortunately, the calculation of the tetrachoric correlation matrix may be problematic, especially when the proportions of the 2x2 tables of item responses are extreme or when the number of observations is low. Collins, Cliff, McCormick, and Zatkin (1986), in a comparison of the performance of phi coefficients and tetrachorics in the factor analysis of binary data, found that the solutions based on tetrachorics contained many Heywood cases (communalities equal to

or greater than unity). Also, the matrix of tetrachoric correlations is not necessarily positive definite, which makes some estimations strategies (e.g., maximum likelihood) inappropriate. Finally, ordinary FA of the tetrachoric correlation matrix does not give correct standard errors or a valid χ^2 test of model fit (Bock & Lieberman, 1970).

Overall, deficiencies in the unweighted least-squares FA of phi coefficients or tetrachoric correlations restricts their use as diagnostic tools in test construction involving items with binary responses categories. However, a number of alternate strategies have been suggested that are based on the assumption that the dichotomous variables are indicators of underlying continuous variables for which FA is appropriate. These techniques, which overcome some of the computational and statistical problems outlined above, are described below. A more detailed review of nonlinear FA techniques for binary data can be found in De Champlain (1995).

Limited-Information (GLS Estimation). Christoffersson (1975) and Muthén (1978) proposed similar, and asymptotically equivalent, solutions for the FA of dichotomous data. These approaches employ tetrachoric correlations as the index of association for factoring binary data. From an IRT perspective, these solutions result in models that express the probability of correctly responding to dichotomously coded items as nonlinear functions of some ability, denoted θ. The resulting

reparameterization of the normal ogive models to common factor analysis models is straightforward. For simplicity, assume that the set of observed response variables measure a single factor or trait (e.g., math ability). The y^* variable, depicted below, can then be thought of as the tendency to respond correctly to a certain item related to math ability. When this tendency exceeds a specific threshold (t), the respondent will provide a correct response, otherwise not. This relationship can be expressed as follows:

$$y = \begin{cases} 1 \text{ if } y^* > t \\ 0 \text{ otherwise,} \end{cases} \qquad (3)$$

where

$t = \Phi^{-1}(p)$ (p = observed proportion of positive responses) and Φ^{-1} is the inverse of the cumulative normal distribution function.

This recognizes that all respondents who respond correctly to a particular math item (y=1) will not have the same level of math ability. As a result, the relationship between y and y^* leads to a nonlinear relationship between y and the latent trait (θ). This nonlinear relationship can be expressed as

$$P(y=1|\theta) = P(Y^* > t) = N(t - \lambda\theta), \qquad (4)$$

where

N is the standard normal density function.

This leads to the common factor analysis model described earlier

$$y^{*}_{j} = \Lambda_{j}\theta + \epsilon_{j}, \qquad (5)$$

which can be estimated using tetrachoric correlations and the limited-information generalized least squares (GLS) estimator (Muthén, 1984). The GLS, or weighted, estimator is similar to the unweighted least-squares (ULS) estimator in that there is an attempt to minimize residual sums of squares. However, under GLS estimation, these residuals are weighted by their sampling variability and, thus, usually provide more stable estimates.

The solutions provided by Christoffersson and Muthén are known as limited-information techniques and utilize sample information up to and including fourth-order moments. They result in an efficient use of the data because when categorical response variables are involved, other estimation strategies (e.g., those based on full-information) may lead to heavy computation (Muthén, 1983). Although the GLS techniques are computationally heavier than ULS, and still involve some loss of information, they provide statistical tests of model fit as well as standard errors of estimation. Nevertheless, the GLS estimator requires the creation of a weight matrix that grows rapidly with the number of items. Mislevy (1986) comments that the computational demands of GLS increase linearly with the number of factors but with the fourth power of the number of items. Therefore, there may be limitations on the number of items that can be employed. Muthén (1978) proposes that, due to

the largeness of the weight matrix, GLS limits the number of variables (items) to 20-25. Unfortunately, most tests have considerably more items. The limited-information GLS solution is currently available in the LISCOMP computer program (Muthén, 1987).

Limited-Information (McDonald's Harmonic Analysis). McDonald (1967a) demonstrated that conventional FA can be applied directly to binary data conforming to the normal ogive model. Using McDonald's formalization, it is possible to specify nonlinear relationships between item performance and ability. This provides a powerful methodological framework by which to estimate item parameters and test the invariance of these measurement parameters across groups.

McDonald (1986) promotes nonlinear FA as a unifying theory for the diversity of models for the analysis of multivariate data. In fact, McDonald (1982) outlined the equivalence between the general case of latent trait theory and item response theory. The essential principle is that any reasonable regular nonlinear function can be approximated closely by a linear combination of terms. Using harmonic analysis, the normal ogive can be approximated by a polynomial series of the general form:

$$z_{j1} = f_{j0} + f_{j1}\theta + f_{j2}\theta^2 + \ldots f_{jk}\theta^k. \qquad (6)$$

McDonald (1962, 1965, 1967a, 1967b) provides a theory for a common FA model in which the regressions of the observed variables on the common factors are nonlinear. By assuming that the latent trait has a normal distribution (i.e., the dichotomies are reasonably considered to have arisen from a continuous normal latent process), this model yields estimates of the item parameters based on the Spearman FA of the matrix of item tetrachoric correlations (Lord, 1952). Thus, the unidimensional (or multidimensional) normal ogive model can be looked upon as a confirmatory common factor model for binary variables. For example, the Rasch model is simply a nonlinear transformation of the special case of the Spearman single factor model in which the factor loadings are required to be the same for all items. McDonald (1980) has also shown that, provided the latent trait has a normal distribution, the 2-parameter normal ogive model can be fitted to binary data by the analysis of covariance structures.

Harmonic analysis of the normal ogive model has been used to fit the normal ogive model by substituting the best-fitting cubic model (McDonald, 1967a). Within this framework, the normal ogive function is approximated by a four term polynomial series with coefficients defined by normalized Hermite-Tchebycheff polynomials and scores defined by ratios of factor loadings (see, Fraser, 1983; Fraser & McDonald, 1988). McDonald (1982) suggests that this strategy provides a very good approximation to the

normal ogive and that terms beyond the cubic can be ignored. The unidimensional cubic model can be written as:

$$Y_i = b_{i0} + b_{i1}\theta + b_{i2}\theta^2 + b_{i3}\theta^3 + e_i, \quad (7)$$

where

b_{ijk} = factor loading of factor j on item i of polynomial of degree k, and
e_i = uniqueness component of item i.

This breakdown of the ICC into a sum of polynomials results in a restricted common FA of a matrix of item covariances, in which factor loadings are restricted to be nonlinear functions of item parameters. This is similar to the limited-information solution proposed by Muthén (1978) but uses an unweighted least squares (ULS) fitting function and a conjugate gradients algorithm. Unlike the weighted LS estimator utilized by Muthén, the ULS fitting function proposed by McDonald is computationally less burdensome and can incorporate numerous responses (subjects), variables (items) and factors (traits).

NOHARM II (Fraser & McDonald, 1988) is a computer program that can be used to fit both the unidimensional and multidimensional normal ogive models of latent trait theory. As part of the theory it is assumed that the latent trait has a normal distribution. Within NOHARM II, item difficulty (b) is estimated from a normal deviate corresponding to percent correct. Item discrimination (a) can be derived from the loading of the item in a one-factor common FA, which is based on the matrix of

joint proportions of item responses (i.e., sample product-moment matrix).

If the residuals (ϵ_i's) are multivariate normal distributed, this model is equivalent to the normal ogive model (Lord, 1980). For example, the 2-parameter normal ogive model (Lord, 1952, 1953) can be expressed as

$$E(y_j) = Pr[y_j = 1 \mid \theta] = N\{a_j(\theta - b_j)\} \quad (8)$$

where

θ is the latent trait variable (common factor),
a_j is the discrimination parameter for item j,
b_j is the difficulty parameter for item j, and
$N\{.\}$ is the normal density function.

Based on this specification, the probability that a randomly selected examinee of ability θ will correctly answer an item is dependent on the two parameters, a and b.

This model can also be written as:

$$E(y_j) = Pr[y_j = 1 \mid \theta] = N\{f_{j0} + f_{j1}\theta\} \quad (9)$$

where

$f_{j0} = -a_j b_j$ and
$f_{j1} = a_j$.

or

$$E(y_j) = Pr[y_j = 1 \mid \theta] = N[(t_j + \lambda_j\theta)/(1 - \lambda_j^2)^{1/2}] \quad (10)$$

where

t_j is the threshold value and
λ_j is the Spearman common factor loading.

This final parameterization corresponds to an interpretation of the model as arising by dichotomization of underlying continuous variables that follow the linear one-factor model (Christoffersson, 1975). It is also the most suitable transformation of straight line regression of test score on a common factor (McDonald, 1981).

McDonald's Harmonic Analysis is applicable to the two-parameter normal ogive model and a modification of the three parameter model in which the pseudo-guessing values are fixed. In addition, it can be used to fit multidimensional models and includes a residual analysis to determine model-data fit. McDonald's polynomial approximation to the normal ogive has been used to examine some problems in educational testing (e.g., Boulet & Gessaroli, 1992; De Champlain & Gessaroli, in press; Fang & Gessaroli, 1995; Miller & Hirsch, 1991). However, compared to other IRT programs, NOHARM has not received much attention in the United States (Hambleton, Swaminathan, & Rogers, 1991).

Full-Information (MML Estimation). The information in all cells of the 2^p contingency table of responses to all items is required for completely efficient estimation of parameters in the factor model for dichotomous items. Full-information (FI) approaches, based on maximum-likelihood (ML) estimation, have been developed for the analysis of factor models incorporating dichotomous items (see Bock, Gibbons, & Muraki, 1988). However,

utilizing full-information maximum likelihood (FIML) estimation of item parameters is computationally difficult in that for n items it requires the generation and inversion of a 2nx2n information matrix (Bock & Lieberman, 1970). Also, the entire 2^n table of response counts is usually unwieldy, difficult to interpret, and full of small expected values (Thissen, 1982). Unlike the LI approaches presented earlier, FI item factor analysis analyses response vectors instead of item pairs. Bock and Lieberman (1970) utilized FIML with a single latent trait and estimated that, given their ML estimation algorithm and the computer resources of the day, the upper limit on the number of items that could be analysed was 10-12. Full-information item FA based on marginal maximum likelihood (MML) estimation and the EM algorithm (see Rubin, 1991) was introduced by Bock and Aitken (1981). The incorporation of the EM algorithm for ML allows for the estimation of item parameters in the marginal distribution, integrating over the distribution of ability (see Bock & Aitken, 1981; Bock & Lieberman, 1970). As a result, the calculation of an inter-item correlation matrix is not required, and the computations are minimized. This allows for the incorporation of many more items and more than one factor and should provide tremendous computational speed advantages for long tests (Thissen, 1982). Nevertheless, the computations involved in MML still increase exponentially with the number of factors and linearly with the number of items (Bock, Gibbons, & Muraki, 1988; Mislevy, 1986).

TESTFACT (Wilson, Wood, & Gibbons, 1991) is an item factor analysis computer program that uses MML to derive parameter estimates for unidimensional and multidimensional IRT models. Unlike LI techniques it uses an iterative procedure to obtain MML estimates of the item parameters via the EM algorithm. The program also provides a statistical test for the number of factors. In the unidimensional case, the normal ogive item response model without guessing is given by

$$\text{Prob}(y_j=1|\theta) = N[z_j(\theta)] \tag{11}$$

where

$$\begin{aligned} z_j(\theta) &= c_j + a_j\theta \text{ and} \\ c_j &= -a_jb_j. \end{aligned}$$

The MML estimates of the factor loadings (λ_j) and the standard difficulties (δj) are calculated from estimates of the slope parameter (a_j) and the intercept parameter (c_j):

$$\lambda_j = a_jd_j^{-1}, \ \delta_j = c_jd_j^{-1} \tag{12}$$

where

$$d_j = (1 + a_j^2)^{\frac{1}{2}}.$$

The familiar IRT discrimination and difficulty parameters are easily derivable from the factor loadings (λ_j) and intercept parameters (c_j):

$$\begin{aligned} a_j &= \lambda_j/(1 - \lambda_j^2)^{\frac{1}{2}} \text{ and} \\ b_j &= \delta_j\lambda_j^{-1}. \end{aligned}$$

Within this specification it is also assumed that the subjects (examinees) are randomly sampled from a normal θ distribution.

Comparisons of Parameter Estimation Strategies

Various FA techniques and estimation strategies can be used to summarize item-trait relationships. A review of the studies in which the utility of competing methods were empirically addressed is provided in this section. This includes studies where particular FA methods were compared as well as studies where particular estimation strategies incorporated in commonly used IRT programs were contrasted. A summary of the use of MML estimation in item calibration studies is also provided. Finally, a review of the studies in which limited- and full-information FA estimators have been specifically contrasted is presented.

Comparison of FA Methods of Dichotomous Data

Parry and McArdle (1991) used simulation methods to compare four least-squares (LS) methods of FA of dichotomous variables. Input matrices were phi correlations, tetrachoric correlations estimated from bivariate tables of the observed variables, tetrachoric correlations estimated on the basis of the latent continuous normal response variables underlying the observed variables (using LISCOMP with GLS factor extraction), and

correlations based on a sample raw product-moment matrix (NOHARM II; Fraser & McDonald, 1988). Under varying sample sizes (N=50, 100, 200), threshold values, and population loadings of a factor model, the more sophisticated third and fourth methods were not found to be markedly superior to the first two methods. However, the calculation of the matrix of tetrachoric correlations followed by a weighted LS method of factor extraction did not work well for data sets having small sample sizes. This was due to the fact that the weight matrix is not always positive definite and thus this method often fails to converge. The authors suggested the samples of less than 200 not be used with weighted LS. The estimates of population factor loadings using McDonald's NOHARM procedure were not superior to those produced by the usual ULS estimation of the tetrachoric correlation matrix. Unfortunately, the simulations in this study were limited to eight dichotomous variables (items) and one factor in all cases. As a result, the authors recommended that further simulation studies be performed in order to study the effect of changing the number of variables (items) in the model and having a broader range of sample sizes. Furthermore, they also suggested that the utility ML (e.g., MML) factor analysis procedures be investigated.

Collins et al. (1986) also performed a simulation study to investigate factor recovery in binary data sets. Artificial data were generated using the two-parameter logistic model. The authors specifically compared the performance of phi and

tetrachoric coefficients in factor structure recovery in various types of binary data sets. Four data sets (normal, low frequency, rectangular, bimodal) were produced by varying the shape of the subject ability distribution and the shape, mean, and standard deviation of the item difficulty distribution. They found that both indices performed best on the normal data sets. In the low frequency data sets the use of tetrachoric coefficients resulted in poor nontrivial factor identification, which was likely the result of large numbers of Heywood cases in this condition. Over all conditions 11% of the variables in each data set had communalities greater than or equal to one. Similar to Parry and McArdle (1991), Collins et al. concluded that, in general, phi coefficients rather than tetrachorics should be factored.

Comparison of IRT and FA Models

There have been a number of studies in which common IRT based computer programs (e.g., LOGIST, BILOG) for analyzing item response data have been compared (see Hsu & Yu, 1989). These comparisons include situations where 1) no assumption of the model is violated, 2) the ability distributions are nonnormal, and 3) the unidimensionality assumption is violated.

When trait values are generated from normal distributions BILOG (Mislevy & Bock, 1984) with MML estimation uses more processing time than joint maximum likelihood estimation (JML)

(LOGIST; Wingersky, Barton & Lord, 1982), produces estimates that are almost always more accurate and is recommended for shorter tests and/or smaller sample sizes (Yen, 1987). In a simulation study, Drasgow (1989) found that MML estimation was able to recover true parameter estimates with tests as short as five items and samples as small as 200. In general, as tests become longer and sample sizes become larger, MML estimation and JML estimation tend to converge (Lord, 1986; Mislevy & Bock, 1984). Therefore, for longer tests, either estimation strategy would appear to be appropriate.

For nonnormal distributions of ability, the shape of the ability distributions has been found to have an effect on parameter estimation in the unidimensional case. Using simulated data for 20- and 40-item tests and 1000 examinees, Yen (1987) compared the item parameter estimates from LOGIST and BILOG under the three-parameter logistic model. For data derived from nonnormal trait distributions (negatively skewed, positively skewed, symmetric but platykurtic) she found that, in almost every case, BILOG yielded more accurate estimates of item parameters than did LOGIST.

In terms of studies involving unidimensional models and multidimensional data, it has been found that major departures from unidimensionality can cause severe problems in parameter estimation (e.g., Hsu & Yu, 1989). In general, the accuracy of parameter estimation is poor when unidimensional models are applied to multidimensional data.

Item Recovery Studies Utilizing MML

Zwinderman and van den Wollenberg (1990) investigated the robustness of MML estimation in the Rasch model. Monte Carlo methods were used to look at the effect of test length and the assumed distribution of ability on the standard errors (SEs) of the parameter estimates. The number of items was 5, 10, or 15. The simulee parameters were sampled from the standard normal or exponential distribution. For the nonnormal case, simulees were sampled from extremely, moderately, and lightly skewed distributions. Sampling from exponential distributions will result in a skewed number-correct score distribution, with a larger skewness corresponding to a larger mean of the exponential sampling distribution. The only sample size incorporated was 4000. Fifty replications of each condition were undertaken. The accuracy of the parameter estimates was assessed by means of the square root of the mean square difference (RMSE) between the input parameters and the estimates, averaged over replications. The authors found that when the distribution of ability was assumed to be normal when it was actually skewed, the MML estimators lose accuracy and efficiency. In addition, the RMSEs for the MML estimates under the skewed distributions of ability varied considerably and increased as the departure from normal ability distribution increased. The average skewness of the number-correct score distribution was -2.01 for the extremely skewed distribution. Overall, the use of an invalid (nonnormal)

ability distribution led to a loss of efficiency and biased estimates when this discrepancy was large.

Stone (1992) evaluated MML estimation of item parameter and ML estimation of ability in a 2-parameter logistic model for varying test lengths, sample size, and assumed ability distributions. One hundred data sets for each combination of factors were simulated. Three possible test lengths (10, 20, 40 items) and three sample sizes (250, 500, 1000 examinees) were incorporated in the Monte Carlo experiment. In addition, three ability distributions were used (normal, positively skewed, and symmetric but platykurtic). These data sets were analysed via MML produced by the EM algorithm, which is implemented in the MULTILOG computer program (Thissen, 1986). Recovery of item parameter values was assessed by averaging information across the 100 replications. Bias in item difficulty and discrimination was assessed by examining the difference between the mean of each parameter and the population value across 100 replications. The root mean squared error (RMSE) was also used to examine the goodness or closeness of the estimates. In general, Stone found that the effects of sample size, test length, and assumed ability distribution depend on the individual item parameters. Overall, more extreme item discrimination estimates were produced with small sample sizes and short tests. However, for poorly discriminating items (a=.83/1.7) the differences in the accuracy of estimation were negligible, regardless of the number of items, the sample size, or the true distribution of ability. When

ability was normally distributed the MML estimate of item discrimination was generally precise and stable. In addition, as sample size increased the variability of the point estimates decreased. In terms of the estimation of item difficulty, somewhat more extreme parameter estimates were produced when the ability distribution was skewed or platykurtic. The MML estimates of item difficulty were, however, generally precise and stable in small samples, short tests, and varying distributions of ability. In terms of both difficulty and discrimination, greater RMSE values were associated with items that were highly discriminating and extremely easy. Stone also commented that some of the large RMSE results that were associated with the smaller sample sizes and short test lengths were probably due to a few cases where very extreme estimates were produced through MULTILOG.

Seong (1990) also investigated the sensitivity of MML estimation of item and ability parameters when the prior and underlying ability distributions were not matched. Thirty sets of 45-item test data were generated by specifying three types of underlying ability distributions (normal, positively skewed, negatively skewed). The positively skewed distributions was defined by a χ^2 distribution with eight degrees of freedom (dfs), resulting in a skewness of one. The negatively skewed distribution was simply the mirror image of the positively skewed distribution. Furthermore, the type of prior ability distribution (normal, positively skewed, negatively skewed), the

number of quadrature points (10,20), and the number of examinees (100,1000) were systematically varied. For each set of values, only five replications of the data were generated. The item response vectors were generated under the two-parameter normal ogive IRT model. Item discriminations ranged from 0.30 to 1.1. Difficulty values ranged from -1.0 to 1.0. The data were analysed via the MML/EM approach of Bock and Aitken as implemented in the BILOG computer program (Mislevy & Bock, 1986).

The author used two descriptive statistics to assess the adequacy of the MML approach. The root mean squared error (RMSE) was used to assess the adequacy of the estimates. The mean of the absolute difference between the estimates and the parameters was used as a measure of bias. He found that item discrimination and difficulty parameters were estimated more accurately when the prior and underlying ability distributions were matched and the sample size was large. However, with small datasets, the appropriate specification of the prior distribution did not increase the accuracy of parameter estimation. Nevertheless, the number of examinees had an important effect on the accuracy of item parameter estimation. In general, increasing the sample size improved the accuracy of estimation for the item discrimination and difficulty parameters. Seong suggested that the user should specify the prior ability distribution based on theoretical or empirical considerations. Unfortunately, the population distribution of ability is often impossible to

determine. In these situations, the default normal prior ability distribution is recommended.

Reise and Yu (1990) investigated parameter recovery in the graded response model using MULTILOG (Thissen, 1986). The graded response model is used to describe test taking behaviour when the responses can be classified as ordered categories. While this is not a binary data model, MULTILOG incorporates MML estimation and can be considered as an extension of the binary logistic model (BILOG; Mislevy & Bock, 1984) to the polytomous case. Therefore, issues related to parameter recovery in this study may have some bearing on MML estimation of the well established one-, two- and three parameter binary models. Reise and Yu investigated a number of conditions that may affect parameter recovery, including sample size, true θ distribution (normal, uniform, skewed), and true discrimination parameter distribution. The potential effects of test length were not investigated. A number of indices were used to examine parameter recovery, including the correlation between population and estimated parameters for each condition and the RMSE. In addition, means of the population and estimated parameters were compared for each condition. They concluded that MULTILOG is an efficient and effective estimation program for the graded response model. In general, the error of the estimates was reasonably low when compared to other studies. Also, there were no consistent positive or negative discrepancies between the population and estimated discrimination parameters. However, parameter estimation was found to be more accurate for

larger samples (N > 1000), highly discriminating items, and samples that are heterogeneous in ability (normal θ distributions). The authors also suggested that, if structural parameter recovery is important, at least 500 examinees are needed to achieve respectable correlations and RMSEs.

In previous item parameter recovery studies where the use of MML estimation was investigated it has been shown that a misspecification of the latent ability distribution can have an effect on both accuracy and efficiency of parameter estimates. Overall, when the population ability distribution is not normal, and no constraints are placed on the prior distribution, the parameter estimates tend to be less stable and precise. Unfortunately, due to methodological constraints, the use of MML in item parameter estimation would not appear to have been sufficiently investigated in these studies. For example, both Drasgow (1989) and Seong (1990) used very low numbers of replications (10 and 5, respectively). These relatively low number of replications yielded unstable results. Even Stone (1992), who incorporated 100 replications of each condition, utilized a maximum test length of 40 and a maximum sample size of 1,000. Similarly, Zwinderman and van den Wollenberg (1990) used only small numbers of items (5, 10, and 15) in their robustness study. Finally, Reise and Yu (1990) used simulation conditions that were based on a common test length of 25 items. Although the authors maintain that with MML estimation test length is not a crucial variable to manipulate, it would still be interesting

know how NLFA calibrations using MML estimation are affected by this factor.

Full-Information Versus Limited-Information

Knol and Berger (1991) compared various factor analytic and multidimensional item response theory (MIRT) models in the estimation of item parameters for dichotomous variables. These comparisons included both full-information models (e.g., MML estimation via TESTFACT) and models that use only pairwise information (e.g., McDonald's Harmonic Analysis). It should be noted that the authors chose not to utilize GLS (LISCOMP; Muthen, 1987) due to the fact that it yields similar estimates as TESTFACT (Bock, Gibbons, & Muraki, 1988). Knol and Berger generated numerous data matrices having known discrimination and difficulty parameters using sample sizes of 250, 500, and 1000. Comparisons were then made between the population item parameters and those derived through the various estimation strategies. The authors concluded that, for multidimensional data sets, NOHARM and the common FA of the tetrachoric correlation matrix performed at least as well as TESTFACT. Although full-information methods (e.g., MML as implemented in TESTFACT) should be theoretically preferred, numerical problems involved with this strategy may have impeded the derivation of more appropriate estimates. It should also be noted that for each set of item parameters and each sample size only 10 replications were generated. The

authors also found that, for the unidimensional data sets, the mean squared differences between the population and estimated parameters were comparable for NOHARM and TESTFACT. Nevertheless, Muraki and Engelhard (1985) have shown that, for items with extreme difficulties, the root mean squared difference between the estimated and generated factor loadings was considerably lower for TESTFACT than for the traditional common FA of tetrachoric coefficients.

Miller (1991) also investigated the quality of multidimensional item response theory (MIRT) model parameter estimates with those derived from NOHARM. In a Monte Carlo study, two data sets comprised of 50 items were generated under the following conditions: n=2000, $r_{\theta 1,\theta 2} = 0, 0.5$. In this study guessing parameters were set to zero, thus yielding a common factor model parameterization of the 2-parameter normal ogive model. Standardized residuals were computed for each person and each item. In order to check for the accuracy of estimation, the means of these residuals, for individual items and overall, were contrasted. The results from this limited simulation suggested that TESTFACT (using MML estimation) and NOHARM (using ULS estimation) produced little bias in the estimates and were equally effective in reproducing data under well-fitting model conditions. Even though the scope of this study was extremely limited, Miller (1991) concluded that the estimates derived from NOHARM were sufficiently accurate for practical purposes.

Research Problem

The utility of factor analytic methodologies in the analysis of dichotomous item response data has not been investigated thoroughly. More specifically, the comparison of parameter estimates from factor analytic based methodologies utilizing full-information and limited-information nonlinear FA models has not covered the wide variety of situations that would be of interest to test developers. In terms of the computational algorithms and computer programs discussed earlier, MML estimation, as implemented in TESTFACT, and ULS estimation, as implemented in NOHARM, may have distinct advantages under certain testing conditions (e.g., number of items, number of examinees, shape of underlying ability distribution). Based on empirical studies, the use of FI nonlinear FA (MML estimation) has generally provided at least equivalent, if not more accurate, parameter estimates than more traditional IRT based procedures (e.g., JML estimation as implemented in LOGIST). Furthermore, to properly implement IRT grounded methodologies and associated computer programs, especially for two- and three-parameter models, a substantial number of examinees and items is generally required. This does not appear to be necessary for MML (Drasgow, 1989). Nevertheless, the computational demands of MML estimation may be problematic, especially for tests incorporating numerous items, or traits, leaving LI strategies as the methods of choice.

In terms of the two LI solutions and estimation approaches considered earlier (i.e., GLS and ULS), McDonald's harmonic analysis would appear to be superior in that the computational demands of the GLS estimator presently precludes the analysis of tests that incorporate more than 20 or 30 items. Also, McDonald's application of nonlinear FA uses ULS estimation to find estimates for the unknown parameters. In general, techniques specifically designed for a least squares objective function can be expected to converge faster than those designed for ML estimation (Everitt, 1987). Furthermore, least squares is the simplest of the standard fitting functions and should be computationally advantageous over most, if not all, test lengths and sample sizes.

Recently, McDonald (1994) commented that "unless it can be shown that full-information methods commonly yield stable additional information, it remains a reasonable conjecture that bivariate information methods are to be preferred since they do not require pooling of answer pattern frequencies, and arguably should be computationally more efficient" (pp. 75-76). In addition, he suggests that a systematic study of full-information and bivariate (limited) information methods, using both constructed data and resampling from large empirical data sets, be performed.

It should be noted that the assumption of a normally distributed latent trait applies for McDonald's model. Although McDonald (1982) suggests that the harmonic analysis of the normal

ogive model should be reasonably robust against nonnormal distributions of examinee ability, he also advises that robustness studies be performed in order to investigate possible violations of this assumption. The presence of nonnormal distributions of ability is a realistic possibility on achievement tests (Micceri, 1989). In particular, the latent ability distributions are not always normal in educational applications of IRT, especially in populations with specific characteristics (Seong, 1990). For example, if we were to develop a test to select a small group of low or high ability persons from some population we would likely want to have items with difficulties near the cutoff. In this situation the distribution of ability, given the items, will likely be skewed. Therefore, it is important to know how the parameter estimates from LI (e.g., McDonald's Harmonic Analysis) and FI (i.e., MML in TESTFACT) methods will behave under situations where the ability distribution is not normal.

Purpose

The purpose of this study is to compare the parameter estimates from nonlinear item factor models based on MML estimation (TESTFACT) and on a simple approximation to the normal ogive function (ULS; NOHARM) under both normal and nonnormal distributions of a single latent ability. It has been demonstrated that IRT is a general form of nonlinear FA.

Therefore, the comparison of parameter estimates from these competing estimation strategies (full-information and limited-information) would appear to be warranted. This research is specifically concerned with the case where a) the manifest variables (test items) are binary and there is a single underlying normally distributed latent variable and b) the manifest variables are binary and there is a specified underlying latent variable that is not normally distributed.

There are a number of research questions that are addressed in this study. First, what is the effect on the parameter estimates (i.e., parameter recovery of item difficulty and item discrimination) of using limited- versus full-information techniques? A quantification of the differences in parameter recovery between NOHARM and TESTFACT will provide the measurement practitioner with valuable information concerning the relative utility of the competing estimation strategies. Second, are the parameter estimates derived from limited- and full-information models unduly, and equally, affected by a violation of the normality specification on the latent response variable (trait)? Finally, considering the initial two questions, how is the closeness of parameter recovery affected at the various levels of the population parameters (i.e., levels of item difficulty (denoted b) and levels of item discrimination (denoted a)) and test length/sample size characteristics?

CHAPTER III: METHODS

The simulation conditions for the Monte Carlo study are outlined in this chapter. This includes a description of modelled ability distributions as well as the initial item parameters that were used. A brief rationale for the choice of specific conditions is also provided. The study design, including specific details on both data generation and program implementation, is also presented in this section. Finally the methods used to analyse the simulated data are discussed.

Simulation Conditions

The comparison of IRT parameter estimates from NOHARM (LI estimation) and TESTFACT (FI estimation) were investigated using Monte Carlo methods. The conditions that were explored are outlined below.

Nonlinear FA Estimation Procedures

Full-Information

Marginal maximum likelihood (MML) estimation, as implemented in the TESTFACT computer program, was used.

Limited-Information

Unweighted least-squares estimation, as implemented in the NOHARM II computer program, was used.

Ability Distributions

Normal (Gaussian)

The use of normal trait (ability) distributions has generally been the standard used for previous comparative studies of parameter recovery. There have been some studies that have looked at the effects on parameter estimates for nonnormal trait distributions (e.g., Yen, 1987), but there do not appear to be any that specifically compare limited- and full-information NLFA approaches under these conditions. In this study normal trait distributions are used as a basis for comparison of the results of parameter recovery against conditions involving nonnormal trait distributions.

Nonnormal

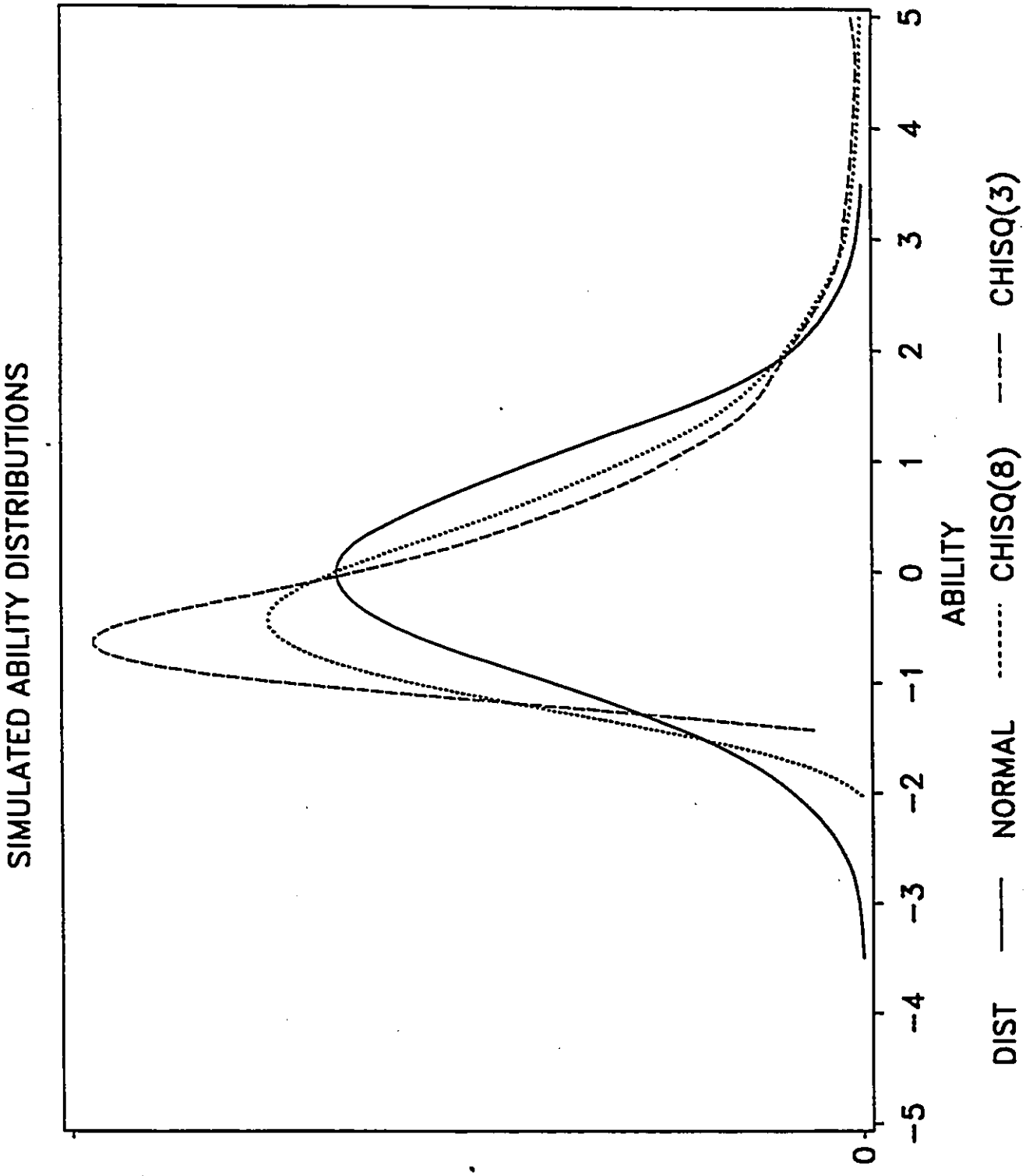
Based on empirical data, it would be expected that ability distributions for general achievement tests would likely be nonnormal (Micceri, 1989). Micceri suggests that for general achievement tests the shape of the ability distributions would likely be moderately asymmetric and contaminated (skewed). Although there are a number of nonnormal trait distributions that could be modelled (e.g., uniform, exponential, double

exponential), χ^2 distributions with relatively small degrees of freedom result in skewed, asymmetric distributions. For the present investigation, the nonnormal trait distributions were defined by χ^2 distributions with eight and three degrees of freedom (dfs). This resulted in ability distributions with skewnesses of approximately 1.00 and 1.75, respectively. The three simulated ability distributions (normal, mildly skewed, moderately skewed) are displayed in Figure 1. It is evident that, compared to the normal trait distribution, the nonnormal trait distributions are characterized by relatively fewer respondents at the tails of the distribution. Also, the homogeneity of the trait, or ability, scores increases as the skewness increases.

For all conditions the subject ability distribution was scaled to have a mean of zero and a standard deviation of one. This results in no difference in the overall level of ability in the groups being compared.

Figure 1

Simulated Ability Distributions



Sample Size and Number of Items

The sample size and number of items could both be expected to have some effect on the accuracy of parameter estimates derived from the two nonlinear FA approaches. More importantly, the ratio of sample size to the number of items may be a limiting condition for full-information models (Boulet & Gessaroli, 1992). The sample sizes and test lengths that were incorporated in order to investigate a possible ratio effect are presented in Table 1. Henceforth, the conditions modelled in Table 1 will be referred to as Study 1. The test lengths (15, 30, 45, and 60 items) were chosen in that they represent a range that is common in psychological and educational applications (Seong, 1990). The sample sizes (number of examinees) were chosen in order to derive ordered ratios of sample size to the number of items.

Table 1

Sample Sizes for the Fixed Ratio Conditions (Study 1)

	N/Items Ratio		
ITEMS	16.67	33.33	66.67
15	250	500	1000
30	500	1000	2000
45	750	1500	3000
60	1000	2000	4000

For MML estimation it could also be expected that, given sufficiently large sample sizes (e.g., 10,000 examinees), accurate and stable parameter estimates could be derived, even with extended test lengths. Therefore, a further set of conditions was modelled in order to specifically explore the individual and combined effect of sample size and test length on the accuracy and stability of the parameter estimates. Table 2 presents the sample sizes and test lengths that were utilized. The analysis of these conditions will henceforth be referred to as Study 2. It should be noted that some of the conditions explored in Study 2 (see Table 2) overlap with those depicted in Table 1 (Study 1). Furthermore, due to excessive computational demands, only the normal and moderately skewed (χ^2_3) distributions were investigated in Study 2.

Table 2

Sample Sizes for Test Length and Sample Size Conditions (Study 2)

ITEMS				
15	250*	500*	1000*	10000
30	250	500*	1000*	10000
45	250	500	1000	10000
60	250	500	1000*	10000

* cells also utilized in first set of experiments (Study 1)

Item Parameters

Item Discriminations (a)

Initial item discriminations were set at 0.5, 1.0, 1.5. Item discrimination parameter values are typically found in the interval (.4,2) (Hambleton, 1993; Hambleton & Swaminathan, 1985). Therefore, these values are realistic and also provide a suitable range for the strength of the item-factor relationships.

Item Difficulty (b)

Initial item difficulties were set at -2, -1, 0, 1, 2. These values were chosen in order to model a fairly wide range of item difficulty. Although item difficulty values can theoretically range from minus infinity to infinity, values typically vary from -2 to +2 (Hambleton, 1993; Hambleton & Swaminathan, 1985). Also, the use of extremely easy (b=-2) and difficult (b=2) items allows for the investigation of the parameter recovery of items that would normally be difficult to estimate.

Design

For interpretive purposes the investigation of parameter recovery was divided into two parts. The purpose of Study 1 was to determine the effect of the sample size/test length ratio on the accuracy and stability of the estimated parameters. For this

component there were 36 possible combinations of conditions (3 ability distributions x 4 test lengths x 3 n/item ratios (see Table 1)). By incorporating 100 replications of each condition, 3600 experiments were performed.

In Study 2 the individual and combined effects of sample size and test length on parameter estimation were determined. Thirty-two combinations of conditions were modelled (2 ability distributions x 4 sample sizes x 4 test lengths (see Table 2)). These conditions included 12 cells that were incorporated in study 1 outlined above. Therefore, only 20 additional combinations were utilized. This resulted in 2000 additional experiments.

Data Generation

Data for the various combinations of conditions were simulated using the M2PLGEN program (Ackerman, 1985) and a modification to incorporate the sampling of examinees from a nonnormal distribution of ability. For conditions involving a normal trait distribution, unidimensional 2-parameter data were generated in which the discrimination (a) and difficulty (b) parameters were fixed across items. Using the defined item parameters, IRT item responses were generated by randomly sampling from a normal ability distribution, determining the probability of a correct response according to the 2-parameter logistic response model given the item parameters and comparing

this probability with a random number sampled from a uniform (0,1) distribution. The simulated item response was scored correct if the probability of a correct response was greater than the sampled number. For the nonnormal ability distributions a variant of the algorithm noted above was employed. Instead of sampling from a normal ability distribution, the IMSL subroutine GGCHS was used to randomly select a θ from a χ^2 distribution with either three or eight degrees of freedom depending on the condition of interest. For both the normal and nonnormal conditions the ability distribution was scaled to have a mean of zero and a standard deviation of one. This adjustment was necessary in order to make meaningful comparisons of item parameters across conditions involving the normal and nonnormal trait specifications.

The population discrimination parameters were set at .5, 1.0, or 1.5. For each test length (i.e., 15, 30, 45, 60) there was an equal number of items within each fixed level of discrimination. For groups of items that had the same discrimination parameter, difficulty values of -2, -1, 0, 1, and 2 were chosen. For example, the 15 item test had five items with discriminations (a) set to .5, five items with discriminations equal to 1.0, and five items with discriminations equal to 1.5. Within this subset of items (fixed discrimination values) the set of five difficulty values mentioned above was used once. For illustrative purposes, the population item parameters for the 15 item test are presented in Table 3. The population item

parameters for the remaining tests lengths (i.e., 30, 45, 60) are simply the appropriate repetitions of this initial 15 item pattern (e.g., 3x for the 45 item test).

Table 3

Population (True) Parameter Values for the 15 Item Test

Item	Difficulty	Discrimination
1	-2	1.0
2	-1	1.0
3	0	1.0
4	1	1.0
5	2	1.0
6	-2	0.5
7	-1	0.5
8	0	0.5
9	1	0.5
10	2	0.5
11	-2	1.5
12	-1	1.5
13	0	1.5
14	1	1.5
15	2	1.5

The data were then analysed using item FA based on MML estimation (TESTFACT) and item FA based on McDonald's Harmonic Analysis (NOHARM II). One-hundred replications of each of the 56 unique combinations were undertaken, resulting in 5600 experiments. For each replication a different seed for the starting value of the random number generator was used to generate the item responses. The use of multiple replications

allows for the analysis of central tendency and variability measures across replications.

In order to investigate the validity of the data generation process, several test data sets, for various test lengths, sample sizes and ability distributions were independently analysed. Both item statistics and descriptive statistics concerning the population item values (e.g., means) and the total test score (e.g., skewness of number correct distribution) were inspected. The results of these analyses suggested that both the ability distributions and the initial item values were being generated correctly via the programming described earlier.

Program Implementation

A brief description of the NOHARM and TESTFACT program implementations is presented in the next two subsections.

NOHARM (Normal Harmonic Factor Analysis)

One-dimensional models were fitted using the NOHARM program. Exploratory factor analysis models were specified. The default convergence criterion of a change of .00001 in the fitting function was utilized.

TESTFACT

Full-information item factor analysis is implemented in the TESTFACT program. There are numerous options that can be specified. Although it is not a default option, TESTFACT allows one to set constraints on item parameters. By setting a prior beta distribution on the unique factor loadings (1 - communality) Heywood cases can be effectively controlled. Factor loadings will not exceed one and the discrimination parameters will not be excessive. A normal prior with a specified mean and variance can also be placed on the intercept parameter. This will result in no excessive or small intercepts. In the present study, the default options (no priors) were chosen in order to maintain comparability with past research. One-dimensional models were fitted in all cases. TESTFACT uses numerical quadrature in fitting models to the data. For the one-factor solution the default number of quadrature points is 10. For the EM algorithm implemented in TESTFACT the defaults of five E-step iterations and 5 M-step iterations were used. The precision criterion for convergence in the M-steps was .005. No prior distributions were placed on the intercept and slope parameters.

Generalizability

It should be remembered that the results derived from this study are idiosyncratic to the model characteristics that were investigated (e.g., normal and nonnormal common factor distributions, number of items, sample size, etc.) and the specific program implementations (e.g., no constraints on intercept and slope parameters for TESTFACT). Nevertheless, the analysis of the results from the varied conditions that are modelled should provide practical guidelines for the appropriate use of the competing estimation strategies.

Analysis

The methods and procedures used to analyse the results of the Monte Carlo experiments are presented in this section.

Descriptive Indicators of Goodness-of-Fit

There have been a number of statistics that have been used when comparing the accuracy of estimates derived from competing programs for Monte Carlo studies involving multiple replications. These include correlations between estimates and population values, mean differences, ratios of standard deviations, and many associated variants. Most commonly, the averaged signed or unsigned difference between the estimated and population

parameters is used as a measure of bias. An estimator is said to be unbiased if the mean of the sampling distribution is equal to the population (true) characteristic to be estimated. Typically, a good estimator should provide reasonable assurance that the estimates obtained will be close to the population values. For the present investigation, the difference between the mean of the estimates and the population value was used to assess the average bias in the estimates.

The square root of the mean squared difference between the population and estimated parameters (Root Mean Squared Error, RMSE), over replications, has also been widely used as a measure of goodness, or closeness, of fit. The RMSE is given by:

$$RMSE = \left[\sum_{k=1}^{r} (\alpha_{ik} - \alpha_i)^2 / r \right]^{1/2}, \tag{13}$$

where r is the number of repetitions, α_{ik} represents the estimate of the difficulty or discrimination parameter for item I in replicate k, and α_i is the population difficulty or discrimination value. This measure, while providing an indication of the spread of the estimates, can be contaminated by bias. If the estimates are biased the RMSE index will be large. Also, if the variance of the estimates is large the RMSE index will also be inflated. It should be noted, however, that relative comparisons of RMSE values, as indicators of the spread of the estimates, are only appropriate once the estimates, across the comparative conditions of interest, have been shown to be

reasonably unbiased. For conditions in which estimators are unbiased, where the RMSE would equal the standard deviation of the estimates, the estimator with the uniformly smallest RMSE would be preferred. Otherwise, outside of this class of unbiased estimators, the RMSE, or mean square error (MSE), can be looked upon as a crude measure of the goodness or closeness of an estimator (Mood, Graybill, & Boes, 1974).

An initial screening of the data indicated that, similar to Stone (1992), a number of extreme estimates were produced. A summary of these outliers is given in Chapter IV. Due to the relatively extreme nature of these improper estimates, and the possible misleading information that may be provided through any analyses based on mean values, the discrepant estimates were rescaled to more reasonable values. For the analyses that compared item recovery between the two programs it was decided to constrain extreme estimates as follows: all discrimination estimates that exceeded 4.5 were rescaled to 4.5; extreme difficulty values ($b>4.5$ or $b<-4.5$) were rescaled to 4.5 and -4.5, respectively.

Many of the researchers who have investigated item parameter recovery averaged the bias and error estimates over all items (e.g., Stone, 1992; Zwinderman et al., 1990). For the present investigation, measures of bias and error were not initially collapsed over items. Instead, values were calculated at all levels of the population item difficulty and discrimination. The examination of individual items, over replications, provides

information regarding the accuracy and variability of parameter estimation for different combinations of item discrimination and difficulty. This is important in that, for specific combinations of population item values, it would be expected that item recovery would be difficult due to a lack of sufficient examinee response strings on which to base the item parameter estimates.

Repeated Measures Analyses

The descriptive measures outlined above can be used to explore both the accuracy and goodness-of-fit of the point estimates under different modelled conditions. For the present investigation it is important, however, to determine the specific factors (e.g., sample size, test length) that lead to estimation differences between the two programs. Once these elements are determined, inspections of the mean values, over replications, can be used to indicate how the programs differ in terms of parameter recovery. Similarly, where differences in recovery do not exist, measures of spread (e.g., RMSEs) can be used to determine the comparative efficiency of the two estimation strategies.

In order to compare the item parameter recovery between the limited- and full-information estimators four repeated measures analyses of variance (RM-ANOVAS) were performed. Due to presence of extreme estimates, these analyses were based on the rescaled data, noted previously. The RM analyses provide the means by

which to test hypotheses about differences in the recovery of the estimated item parameters derived from the competing estimation strategies as a function of the conditions under which they are estimated (e.g., test length, sample size, etc.). Thus, one can identify the independent variables, and interactions, that contribute to differences in parameter recovery between the two estimation strategies. As mentioned above, any differences, or similarities, in item recovery can then be explored in order to assess the conditions under which one estimation strategy may provide more accurate, less biased, parameter values. For conditions in which LI and FI may recover the parameters equally well, descriptive measures of the stability of the estimates (e.g., RMSEs) can be used to judge relative efficiency. Since the same data were analysed with two competing computer packages the individual program (TESTFACT or NOHARM) can be looked upon as the repeated measures (RM) factor (within-subject factor which has been measured twice). Within this framework the combined effects of classification variables (e.g., sample size, test length, etc.) on item parameter recovery can be disentangled.

Fixed Ratio (Study 1)

Two separate RM ANOVAS were performed for fixed ratio experiments (described in Table 1). The first analysis investigated the differences in difficulty parameter recovery between LI and FI estimation as a function of the modelled trait distribution, the ratio of sample to test length, and the

population discrimination and difficulty values. The second analysis incorporated the same between-subject (BS) factors but used the discrimination estimates from NOHARM and TESTFACT as the RM factor (within-subject). For both analyses the dependent variable was the parameter estimates produced by NOHARM and TESTFACT. The BMDP 2V program (BMDP Statistical Software Manual, 1990) was used for these analyses.

Test Length and Sample Size (Study 2)

Two separate RM ANOVAS were also performed for this set of experiments (described in Table 2). The first analysis investigated the differences in difficulty parameter recovery between LI and FI estimation as a function of the modelled trait distribution, the sample size, the test length, and the population discrimination and difficulty values. The second analysis incorporated the same between-subject factors but used the discrimination estimates from NOHARM and TESTFACT as the RM factor. For both analyses the dependent variable was the parameter estimates produced by NOHARM and TESTFACT. The BMDP 2V program (BMDP Statistical Software Manual, 1990) was also used for these analyses.

CHAPTER IV: RESULTS

The results for the study are presented in three parts. A synthesis of the extreme estimates is provided in the first section. A detailed breakdown of the frequency of occurrence of extreme estimates provides important information regarding the nature of estimation difficulties encountered with NOHARM and TESTFACT. The results of the RM ANOVAs are presented in the second section. The statistical comparison of item recovery between LI and FI estimation, based on the rescaled data, is essential in order to delineate the nontrivial conditions in which one estimator, on average, provides meaningfully different estimates than the other. For situations in which NOHARM and TESTFACT did not provide statistically equivalent difficulty or discrimination estimates, graphical summaries of the means of the estimates, across conditions of interest, are presented. These figures allow for the interpretation of the direction and the magnitude of estimation bias for particular combinations of population conditions. Descriptive summaries are presented in section three. These summaries are provided for two reasons. First, the analyses in part two only provide for a comparison of the estimators based on means. Therefore, only difference in accuracy can be ascertained. For conditions where both estimators may provide equally unbiased or equally biased estimates, measures of error or variability are needed in order to choose the most appropriate estimation strategy. Such

measures are provided in section three. Second, descriptive summaries, across conditions identified in section two, provide an additional mechanism with which to interpret the specific interactive effects of certain population conditions that lead to estimation differences between the LI and FI approaches.

As mentioned above, the study was broadly divided into two parts (Study 1, Study 2), each corresponding to specified research questions. Therefore, the three sections noted above will also be summarized individually for each study. Furthermore, for clarity, the results for the difficulty and discrimination estimates will also be segregated.

Extreme Estimates

A preliminary examination of the data indicated that some of the difficulty and discrimination estimates produced by NOHARM and TESTFACT were outside of reasonable ranges. Although it was decided to constrain these values for the parametric analyses (i.e., RM ANOVAs), a thorough investigation of the conditions under which extreme estimates were produced is also essential in order to ascertain the comparative utility of the two estimators.

The parameter estimates derived from the two programs were initially screened for possible values that were outside of reasonable ranges. This process was performed separately for Study 1 and Study 2. Similar to Stone (1992), discrimination (a) parameter estimates that exceeded 4.5 were considered to be

improper. Likewise, for the difficulty parameter (b), estimated values greater than 4.5 or less than -4.5 were considered to be aberrant.

Fixed Ratio (Study 1)

Both TESTFACT and NOHARM produced discrimination and difficulty estimates that, based on realistic values, could be considered to be unreasonable. A summary of the patterns of the outlier generation, by estimation method, is presented here. This overview encompasses the conditions described for Study 1. The profiles of improper estimation are presented separately for the discrimination and difficulty values.

Item Discrimination

The number of outlier discrimination values, based on the criteria described above, by modelled trait distribution and estimation method is shown in Table 4. It should be noted that, based on the Monte Carlo design described previously, the total number of item parameter (a,b) pairs estimated for this part of the study was 135,000. Furthermore, there were 45,000 estimates produced under each or three modelled trait distributions (normal, χ^2_8, χ^2_3). As depicted in Table 4, the number of improper estimates of the discrimination parameter increased markedly as the modelled trait distribution became more skewed. For example, when the trait distribution was normal, NOHARM only produced 11

discrimination values that were greater than 4.5. In contrast, 788 discrepant estimates were produced under the moderately skewed (χ^2_3) distribution. Overall, both estimation methods produced relatively few outlier discrimination values when the ability distribution was normal.

Table 4

Improper Estimates of Discrimination Parameters ($a > 4.5$) by Trait Distribution (Study 1)

DISTRIBUTION OF ABILITY	PROGRAM		ESTIMATES
	TESTFACT	NOHARM	(TOTAL)
NORMAL	2	11	45000
$\chi^2(8)$	8	113	45000
$\chi^2(3)$	49	788	45000
TOTAL	59(.044)	912 (.629)	

( ) percentage of the total number of estimates

The number of outlier discrimination values by the sample size/test length ratio, trait distribution and estimation method is presented in Table 5. As previously depicted in Table 4, the number of outlier discrimination values was negligible when the trait distribution was normal. For FI estimation (TESTFACT), within each level of the trait distribution, there did not appear to be a strong relationship between the sample size/test length ratio and the number of discrepant discrimination estimates that were produced. In contrast, the number of outlier discrimination values produced by NOHARM, under the nonnormal specifications, tended to decrease as the sample size/test length ratio

increased. For example, under the χ^2_3 ability specifications, NOHARM produced 379 (0.842 %) outlier discrimination values when the ratio was 16.67. This value decreased to 153 (0.340 %) when the ratio was 66.67.

Table 5

Improper Estimates of Discrimination Parameters (a > 4.5) by Trait Distribution and n/Items Ratio (Study 1)

DISTRIBUTION OF ABILITY	RATIO	PROGRAM		ESTIMATES (TOTAL)
		TESTFACT	NOHARM	
NORMAL	16.67	2	11	15000
NORMAL	33.33	0	0	15000
NORMAL	66.67	0	0	15000
$\chi^2(8)$	16.67	5	75	15000
$\chi^2(8)$	33.33	1	27	15000
$\chi^2(8)$	66.67	2	11	15000
$\chi^2(3)$	16.67	20	379	15000
$\chi^2(3)$	33.33	12	256	15000
$\chi^2(3)$	66.67	17	153	15000

The number of discrimination outliers by distribution, sample size/test length ratio and population a and b values is shown in Table 6. This table only depicts outliers for population conditions involving a=1.5 and b=-2,+2. For both TESTFACT and NOHARM the inflated estimates of the discrimination parameter (a > 4.5) occurred almost exclusively when the

population values for a and b were 1.5 and 2, respectively.

Table 6

Improper Estimates of the Discrimination Parameter (a > 4.5) by Trait Distribution, n/Items Ratio, and Population a and b (Study 1)

Type	Ratio	a	b	NOHARM	TESTFACT	ESTIMATES (TOTAL)
Normal	16.67	1.5	-2	7	1	1000
Normal	16.67	1.5	2	4	1	1000
Normal	33.33	1.5	-2	0	0	1000
Normal	33.33	1.5	2	0	0	1000
Normal	66.67	1.5	-2	0	0	1000
Normal	66.67	1.5	2	0	0	1000
$\chi^2(8)$	16.67	1.5	-2	1	0	1000
$\chi^2(8)$	16.67	1.5	2	65	4	1000
$\chi^2(8)$	33.33	1.5	-2	0	0	1000
$\chi^2(8)$	33.33	1.5	2	23	1	1000
$\chi^2(8)$	66.67	1.5	-2	0	0	1000
$\chi^2(8)$	66.67	1.5	2	11	2	1000
$\chi^2(3)$	16.67	1.5	-2	0	0	1000
$\chi^2(3)$	16.67	1.5	2	342	19	1000
$\chi^2(3)$	33.33	1.5	-2	0	0	1000
$\chi^2(3)$	33.33	1.5	2	247	12	1000
$\chi^2(3)$	66.67	1.5	-2	0	0	1000
$\chi^2(3)$	66.67	1.5	2	151	17	1000

Only includes outliers for conditions involving a=1.5 and b=-2,+2

For NOHARM, over all conditions, there was a total of 912 discrepant discrimination values that were produced. Eight hundred and fourty-three (92.4 %) of these outlier estimates occurred when the population item parameters were a=1.5 and b=2. Similarly 56 (94.9 %) of the 59 outlier values produced by TESTFACT occurred under these conditions. For the experiments involving a normal specification on the latent trait there were also some inflated discrimination estimates corresponding to modelled items that were highly discriminating (a=1.5) and easy (b=-2). However, based on the total number of estimates produced under the normal conditions where a=1.5 and b=-2 (n=3000) these outliers are likely to be due to sampling fluctuations. Similar to the results presented in Table 5, Table 6 also shows that, for NOHARM, the number of outlier discrimination estimates produced under the nonnormal trait specifications appears to be dependent on the sample size/test length ratio. Larger ratios result in fewer discrimination outliers.

Item Difficulty

Based on the criteria for improper difficulty estimates outlined previously, both TESTFACT and NOHARM produced few improper estimates when the modelled ability distribution was normal. However, similar to results found for the estimated discrimination parameters, the number of improper estimates of the difficulty parameter increased as the ability distribution deviated from normal (see Table 7). Based on 45,000 individual

estimates, NOHARM and TESTFACT each produced only three outlier difficulty values when the ability distribution was normal. However, when the trait distribution was moderately skewed (χ^2_3) both estimation procedures resulted in significantly more outlier estimates. For example, NOHARM produced over 1,900 outlier difficulty values under this nonnormal trait condition. Nevertheless, it should be emphasized that 45,000 difficulty estimates were produced under each of the three ability specifications. Therefore, based on proportional representation, the number of discrepant estimates that were generated is still fairly small.

Table 7

Improper Estimates of the Difficulty Parameter ($b < -4.5$ or $b > 4.5$) by Trait Distribution (Study 1)

DISTRIBUTION OF ABILITY	PROGRAM TESTFACT	NOHARM	ESTIMATES (TOTAL)
NORMAL	3	3	45000
$\chi^2(8)$	48	83	45000
$\chi^2(3)$	473	1924	45000
TOTAL	524(.39)	2010(1.49)	

( ) percentage of the total number of estimates

The number of outlier difficulty values by the sample size/test length ratio, trait distribution, and estimation method is presented in Table 8. For each of the nine combinations of ratio and distribution there were 15,000 difficulty estimates

produced by each program. In contrast to the outlier results discussed for the discrimination estimates, the impact of the sample size/test length ratio, within specific ability distributions, was more pronounced for TESTFACT than for NOHARM. For example, under the χ^2_3 specification, TESTFACT produced approximately three times as many outlier difficulty values (244 versus 82) when the ratio was small (n/items=16.67) as opposed to large (n/items=66.67). Under identical conditions, NOHARM produced a relatively consistent number of difficulty outliers.

Table 8

Improper Estimates of Difficulty Parameter ($b < -4.5$ or $b > 4.5$) by Trait Distribution and n/Items Ratio (Study 1)

DISTRIBUTION OF ABILITY	RATIO	PROGRAM		ESTIMATES
		TESTFACT	NOHARM	(TOTAL)
NORMAL	16.67	3	3	15000
NORMAL	33.33	0	0	15000
NORMAL	66.67	0	0	15000
$\chi^2(8)$	16.67	38	59	15000
$\chi^2(8)$	33.33	9	20	15000
$\chi^2(8)$	66.67	1	4	15000
$\chi^2(3)$	16.67	244	644	15000
$\chi^2(3)$	33.33	147	628	15000
$\chi^2(3)$	66.67	82	652	15000

While the frequency pattern of difficulty outliers under the χ^2_8 ability distribution does suggest that the sample size/test length ratio may possibly impact on the precision, and

variability, of the difficulty estimates for both programs, the relatively small number of outlier values, compared to the total number of estimates produced under these conditions, makes this observation difficult to generalize.

The number of difficulty outliers for the ratio experiments (Study 1) by distribution, sample size/test length, and population discrimination (1.5) and difficulty (-2 or +2) values is shown in Table 9. Relatively few difficulty outliers were produced by TESTFACT (n=98/524, 18.7 %) or NOHARM (n=219/2010, 10.9 %) outside of conditions with these initial population parameter values (i.e., a=1.5, b=-2,+2). Therefore, the frequencies of improper estimation are not summarized for population values outside of combinations of these population values. Unlike the discrimination outliers, improper difficulty estimates were produced almost exclusively under population conditions involving easy (b=-2) highly discriminating (a=1.5) items. For TESTFACT, over 81 percent (425/524) of the outlier difficulty estimates were produced when the population item values were a=1.5 and b=-2. Similarly, over 89 percent (1790/2010) of the outlier values produced by NOHARM occurred when the population values were a=1.5 and b=-2. As noted previously, the generated outlier difficulty estimates did not appear to be dependent on the ratio of sample size to the number of test items, except for the FI estimates under the moderately skewed trait distribution. It should also be emphasized that, under conditions involving a normal distribution, there were

relatively few discrepant estimates produced. Of the combined

Table 9

Improper Estimates of the Difficulty Parameter (b < -4.5 or b > 4.5) by Trait Distribution, n/Items Ratio, and Population a and b Values (Study 1)

Type	Ratio	a	b	NOHARM	TESTFACT	ESTIMATES (TOTAL)
Normal	16.67	1.5	-2	0	0	1000
Normal	16.67	1.5	2	1	1	1000
Normal	33.33	1.5	-2	0	0	1000
Normal	33.33	1.5	2	0	0	1000
Normal	66.67	1.5	-2	0	0	1000
Normal	66.67	1.5	2	0	0	1000
$\chi^2(8)$	16.67	1.5	-2	46	25	1000
$\chi^2(8)$	16.67	1.5	2	0	0	1000
$\chi^2(8)$	33.33	1.5	-2	17	8	1000
$\chi^2(8)$	33.33	1.5	2	0	0	1000
$\chi^2(8)$	66.67	1.5	-2	3	1	1000
$\chi^2(8)$	66.67	1.5	2	0	0	1000
$\chi^2(3)$	16.67	1.5	-2	520	184	1000
$\chi^2(3)$	16.67	1.5	2	0	0	1000
$\chi^2(3)$	33.33	1.5	-2	567	130	1000
$\chi^2(3)$	33.33	1.5	2	0	0	1000
$\chi^2(3)$	66.67	1.5	-2	637	77	1000
$\chi^2(3)$	66.67	1.5	2	0	0	1000

Only includes outliers for conditions involving a=1.5 and b=2,-2

total of 2,534 difficulty outliers that were produced, 2,528 (or 99.8 %) were produced under conditions where the trait

distribution was not normal. Also, for both TESTFACT and NOHARM, all of the outlier difficulty estimates were less than -4.5. That is, the direction of bias in the difficulty value corresponded to the direction of the population value.

Sample Size and Test Length (Study 2)

Outlier discrimination and difficulty values for Study 2 are summarized in this section. Similar to the previous synopsis of discrepant estimates in Study 1, patterns of improper estimation are profiled separately for the discrimination and difficulty estimates. As described in the methods section, only a normal and a moderately skewed trait distribution were utilized in Study 2.

Item Discrimination

The number of outlier discrimination estimates, by distribution, for the 32 conditions involving sample size and test length (see Table 2) is displayed in Table 10. For these conditions there was a total of 120,000 item parameter pairs estimated. Similar to the values reported for the fixed ratio experiment (Study 1), the number of discrimination outliers increased substantially when the underlying trait distribution was skewed. For NOHARM, 1,207 (97.3 %) of the 1,241 improper discrimination estimates occurred when the distribution was not normal. For TESTFACT, only 20 percent of the discrimination

outliers were estimated under a normal trait specification. Overall, TESTFACT produced relatively few outlier discrimination estimates. Even for the χ^2_3 distribution, only 90 (of 60,000) discrimination estimates were in excess of 4.5. In contrast, over two percent (1027/60000) of the NOHARM estimates, under conditions involving a nonnormal trait distribution, were considered to be outliers.

Table 10

Improper Estimates of the Discrimination Parameter (a > 4.5) by Trait Distribution (Study 2)

DISTRIBUTION OF ABILITY	PROGRAM		ESTIMATES
	TESTFACT	NOHARM	(TOTAL)
NORMAL	23	34	60000
$\chi^2(3)$	90	1207	60000
TOTAL	113(.094)	1241(1.03)	

( ) percentage of the total number of estimates

In order to look at the effect of test length within a given sample size, the number of outliers was also expressed as a percentage of the total number of estimates produced within each test length. Simply comparing the raw numbers would be inappropriate in that, within a given sample size, four times as many estimates are produced for a 60 item test as opposed to a 15 item test. The actual number of outlier estimates and the percentage that this figure represents of the total number of estimates within that condition are both shown in Table 11.

Table 11

Improper Estimates of the Discrimination Parameter (a > 4.5) by Distribution, Sample Size and Test Length (Study 2)

Dist.	N	Items	NOHARM (%)	TESTFACT (%)	ESTIMATES (TOTAL)
Normal	250	15	10 (.667)	2 (.133)	1500
Normal	250	30	7 (.233)	5 (.167)	3000
Normal	250	45	4 (.089)	11 (.244)	4500
Normal	250	60	11 (.183)	5 (.083)	6000
Normal	500	15	0 (.000)	0 (.000)	1500
Normal	500	30	1 (.033)	0 (.000)	3000
Normal	500	45	1 (.022)	0 (.000)	4500
Normal	500	60	0 (.000)	0 (.000)	6000
Normal	1000	15	0 (.000)	0 (.000)	1500
Normal	1000	30	0 (.000)	0 (.000)	3000
Normal	1000	45	0 (.000)	0 (.000)	4500
Normal	1000	60	0 (.000)	0 (.000)	6000
Normal	10000	15	0 (.000)	0 (.000)	1500
Normal	10000	30	0 (.000)	0 (.000)	3000
Normal	10000	45	0 (.000)	0 (.000)	4500
Normal	10000	60	0 (.000)	0 (.000)	6000
$\chi^2(3)$	250	15	61 (4.067)	15 (1.000)	1500
$\chi^2(3)$	250	30	108 (3.600)	11 (.367)	3000
$\chi^2(3)$	250	45	144 (3.200)	7 (.156)	4500
$\chi^2(3)$	250	60	191 (3.183)	9 (.150)	6000
$\chi^2(3)$	500	15	41 (2.733)	12 (.800)	1500
$\chi^2(3)$	500	30	83 (2.767)	5 (.167)	3000
$\chi^2(3)$	500	45	119 (2.644)	1 (.022)	4500
$\chi^2(3)$	500	60	131 (2.183)	0 (.000)	6000

Table 11 (cont.)

Dist.	N	Items	NOHARM (%)	TESTFACT (%)	ESTIMATES (TOTAL)
$\chi^2(3)$	1000	15	28 (1.867)	17 (1.133)	1500
$\chi^2(3)$	1000	30	62 (2.067)	0 (.000)	3000
$\chi^2(3)$	1000	45	84 (1.867)	0 (.000)	4500
$\chi^2(3)$	1000	60	128 (2.133)	0 (.000)	6000
$\chi^2(3)$	10000	15	2 (.133)	13 (.867)	1500
$\chi^2(3)$	10000	30	6 (.200)	0 (.000)	3000
$\chi^2(3)$	10000	45	5 (.111)	0 (.000)	4500
$\chi^2(3)$	10000	60	14 (.233)	0 (.000)	6000
	TOTAL		1241	113	120000
	Normal Dist.		34	23	60000
	$\chi^2(3)$ Dist.		1207	90	60000
	a=1.5, b=2		1043	97	

( ) is the percentage that this figure represents, given the number of estimates produced within the specified condition (row of the table)

For example, with a normal distribution, a sample size of 250, and a test length of 15 items, NOHARM and TESTFACT produced 10 and 2 outlier discrimination estimates, respectively. Under these conditions there were 1,500 estimates produced (15 items x 100 replications). Therefore, based on a percentage of the estimates produced within this cell, the number of TESTFACT and NOHARM outliers represents a relatively small percentage of the total (.133 and .667 %, respectively).

Both estimation strategies resulted in the generation of significantly greater numbers of outliers under conditions

involving a skewed trait distribution. However, based on the percentage figures, test length, individually, did not appear to affect the likelihood of generating improper discrimination values based on LI estimation. For NOHARM, while the actual number of discrimination outliers tended to increase as test length increased, the values, expressed as a percentage of the number of estimates, remained relatively constant within a given sample size. For TESTFACT, however, substantially greater proportions of discrimination outliers were produced for shorter test lengths. For the shortest test length (items=15) and a skewed trait distribution approximately one percent of the discrimination estimates were greater than 4.5. In contrast, for the longest test length (items=60), less than 0.05 percent of the FI estimates were greater than 4.5. For both NOHARM and TESTFACT, under both the normal and nonnormal specifications, the number or outliers, expressed as a percentage of the estimates, decreased as the sample size increased. For test lengths of 250 and a nonnormal ability distribution, 3.36 percent of the NOHARM discrimination estimates exceeded 4.5. Under the same conditions, 0.28 percent of the TESTFACT estimates exceeded 4.5. In contrast, when the sample size was 10,000, the percentage figures for NOHARM and TESTFACT were 0.18 and 0.087, respectively.

Similar to the results reported for Study 1, outlier discrimination estimates occurred primarily for highly discriminating (a=1.5), difficult items (b=2). For NOHARM, 1,043

(84.0 %) of the total 1,241 outlier values were generated under conditions involving population values of a=1.5 and b=2 (see Table 11). Similarly, over 85 percent (97/113) of the TESTFACT outliers were produced under these initial population item values.

Item Difficulty

The number of improper difficulty estimates, by estimation strategy and trait distribution, for Study 2, is displayed in Table 12. For both NOHARM and TESTFACT there were relatively few discrepant difficulty estimates produced when the assumed trait distribution was normal.

Table 12

Improper Estimates of the Difficulty Parameter (b < -4.5 or b > 4.5) by Trait Distribution (Study 2)

DISTRIBUTION OF ABILITY	PROGRAM TESTFACT	PROGRAM NOHARM	ESTIMATES (TOTAL)
NORMAL	43	28	60000
$\chi^2(3)$	1048	2861	60000
TOTAL	1091(0.91)	2889(2.40)	120000

( ) percentage of the total number of estimates

For example, the 43 outlier estimates produced through FI estimation represents only 0.072 percent of the 60,000 total estimates. However, based on conditions involving a skewed distribution, both NOHARM and TESTFACT produced substantially

larger numbers of discrepant difficulty estimates. For NOHARM, 4.8 percent (n=2,861) of the difficulty estimates produced under the nonnormal trait distribution were categorized as outliers.

The number of improper difficulty estimates as a function of the trait distribution, sample size, and test length is shown in Table 13. The number of outliers produced by TESTFACT and NOHARM is also expressed as a percentage of the total number of estimates produced within each test length. Similar to the results presented for the discrimination estimates, both LI and FI estimation produced some discrepant difficulty estimates when the sample size was small (n=250) and the trait distribution was normal. Nevertheless, when considering the minimum number of estimates produced for particular test length and sample size combinations (n=1500), the number of discrepant estimates is relatively small. For example, under conditions involving a test length of 60 items and a sample size of 250, only 0.42 percent of the TESTFACT difficulty estimates were found to be unrealistic. For conditions involving the skewed trait distribution, both TESTFACT and NOHARM produced substantially greater numbers of outlier difficulty estimates. In fact, of the combined total of 3,980 difficulty outliers, 3,909 (98.2 %) were produced under the χ^2_3 trait distribution. The generation of LI (NOHARM) outlier difficulty estimates did not, however, appear to be dependent on test length. For each sample size the percentage of discrepant estimates was relatively constant across all test lengths.

Table 13

Improper Estimates of the Difficulty Parameter ($b < -4.5$ or $b > 4.5$) by Trait Distribution, Sample Size and Test Length (Study 2)

Distribution	N	Items	NOHARM (%)	TESTFACT (%)	ESTIMATES (TOTAL)
Normal	250	15	3 (.200)	3 (.200)	1500
Normal	250	30	4 (.133)	3 (.100)	3000
Normal	250	45	7 (.156)	13 (.289)	4500
Normal	250	60	14 (.233)	25 (.417)	6000
Normal	500	15	0 (.000)	0 (.000)	1500
Normal	500	30	0 (.000)	0 (.000)	3000
Normal	500	45	0 (.000)	0 (.000)	4500
Normal	500	60	0 (.000)	0 (.000)	6000
Normal	1000	15	0 (.000)	0 (.000)	1500
Normal	1000	30	0 (.000)	0 (.000)	3000
Normal	1000	45	0 (.000)	0 (.000)	4500
Normal	1000	60	0 (.000)	0 (.000)	6000
Normal	10000	15	0 (.000)	0 (.000)	1500
Normal	10000	30	0 (.000)	0 (.000)	3000
Normal	10000	45	0 (.000)	0 (.000)	4500
Normal	10000	60	0 (.000)	0 (.000)	6000
$\chi^2(3)$	250	15	95 (6.333)	64 (4.267)	1500
$\chi^2(3)$	250	30	182 (6.067)	123 (4.100)	3000
$\chi^2(3)$	250	45	245 (5.444)	155 (3.444)	4500
$\chi^2(3)$	250	60	317 (5.283)	175 (2.917)	6000
$\chi^2(3)$	500	15	64 (4.267)	41 (2.733)	1500
$\chi^2(3)$	500	30	133 (4.433)	64 (2.133)	3000
$\chi^2(3)$	500	45	206 (4.578)	95 (2.111)	4500
$\chi^2(3)$	500	60	269 (4.483)	124 (2.067)	6000

Table 13 (cont.)

Distribution	N	Items	NOHARM (%)	TESTFACT (%)	ESTIMATES (TOTAL)
$\chi^2(3)$	1000	15	67 (4.467)	46 (3.067)	1500
$\chi^2(3)$	1000	30	127 (4.233)	42 (1.400)	3000
$\chi^2(3)$	1000	45	192 (4.267)	49 (1.089)	4500
$\chi^2(3)$	1000	60	230 (3.833)	59 (.983)	6000
$\chi^2(3)$	10000	15	71 (4.733)	14 (.933)	1500
$\chi^2(3)$	10000	30	144 (4.800)	2 (.067)	3000
$\chi^2(3)$	10000	45	223 (4.956)	0 (.000)	4500
$\chi^2(3)$	10000	60	296 (4.933)	0 (.000)	6000
	TOTALS		2889	1091	120000
	Normal		28	43	60000
	$\chi2(3)$		2861	1048	60000
	a=1.5,b=-2		2315	716	

( ) is the percentage that this figure represents, given the number of estimates produced within the specified condition (row of the table)

In contrast, the percentage of outlier estimates produced through FI estimation, within a given sample size, decreased as test length increased. For example, under conditions involving a nonnormal trait distribution, a sample size of 1000, and a test length of fifteen items, approximately three percent of the FI difficulty estimates were outliers. This figure decreased to approximately one percent when the test length was 60. For both NOHARM and TESTFACT the estimation of discrepant difficulty values under the χ^2_3 ability distribution also appeared to be dependent on the sample size. Fewer outlier estimates were

produced as the sample size increased. This trend was more pronounced for FI estimation. Overall, FI estimation produced less than half as many improper difficulty estimates than did LI estimation.

Similar to the results presented for the ratio experiments (Study 1), improper difficulty estimates were most likely to occur for highly discriminating (b=1.5) easy items (b=-2). Of the combined total of 3,980 difficulty outliers, 3,031 (77.8 %) were produced when population item values were a=1.5 and b=-2. Under these conditions the difficulty estimates were always less than -4.5.

Summary of Extreme Estimates

The analysis of the outliers produced by NOHARM and TESTFACT revealed some interesting trends. It was clearly evident that, under almost all conditions, FI estimation produced fewer discrepant discrimination and difficulty estimates than did LI estimation. Furthermore, for both LI and FI estimation, the number of outlier values increased markedly as the modelled trait distribution deviated from normal.

In terms of the discrimination estimates, the number of outlier values produced via TESTFACT did not appear to be dependent on the sample size/test length ratio. Instead, under a skewed trait distribution, proportionately fewer outliers were produced as the test length increased. For NOHARM, the number of discrimination outliers decreased as the sample size/test length

ratio increased. However, test length, alone, did not appear to be associated with the number of discrepant discrimination estimates that were produced with LI estimation. Nevertheless, regardless of estimation method used, the majority of outlier discrimination estimates were produced when the population item parameters were a=1.5 and b=2. Under these conditions, especially with a skewed trait distribution, there will be very few examinee responses on which to base the item-trait relationship. Also, given the positive skew of the trait, or ability, distribution, it is likely that under these population values most examinees would provide incorrect responses to all items. As a result, estimation of the discrimination parameter will be difficult and likely yield extreme values. However, as expected, the number of discrimination outliers that were estimated by both NOHARM and TESTFACT decreased as the sample size increased.

In contrast to the results for the discrimination estimates, difficulty outliers occurred predominantly when the population item parameters were a=1.5 and b=-2. In addition, other than a few values under normal trait conditions involving a small sample size (n=250), most of the outlier estimates occurred under the skewed trait specifications. For NOHARM, the generation of discrepant difficulty estimates did not appear to be overly dependent on the sample size/test length ratio, except under the mildly skewed trait distribution where increased sample size/test length ratios resulted in fewer discrepant estimates. Moreover,

neither test length nor sample size, independently, appeared to have a profound effect on the generation of discrepant difficulty estimates. In contrast, under all conditions involving a skewed trait distribution, TESTFACT produced fewer outliers as the sample size/test length ratio increased. Furthermore, the number of difficulty outliers decreased with increases in sample size. Likewise, within fixed sample sizes, and a nonnormal trait distribution, the number of difficulty outliers decreased as test length increased. Overall, for both LI and FI estimation, the generation of outlier difficulty values was not problematic when the modelled trait distribution was normal.

Repeated Measures ANOVA

Repeated measures (RM) designs involve a repeated measurement on the unit of analysis (e.g., subjects) for one or more independent or between-subject variables. For the present investigation, parameter estimates were obtained from TESTFACT and NOHARM under identical sets of conditions. Therefore, two measures of the dependent variable (parameter estimate) were obtained. The estimation method can then be described as the within-subject measure and has two levels (LI- and FI estimation). There are a number of between-subject factors which include the trait distribution, test length, sample size and population parameter values (a,b).

The RM design was chosen to disentangle the sources of estimation differences between LI and FI estimation. Significant main effects, and interactions, indicate where important differences in parameter estimation between TESTFACT and NOHARM occur. For any significant effects, an inspection of the means of the estimates, across the particular between-subject factors, will show conditions where the recovery of the parameter estimates is not equivalent between LI and FI estimation. Furthermore, comparisons of the means of the estimates with the population values will indicate the degree and direction of bias for a particular estimation method.

The results from the RM analyses are outlined in the following section. Similar to the descriptions of the outlier analyses, the results are presented separately for the ratio (Study 1) and sample size (Study 2) experiments. These summaries are preceded by a brief discussion of the statistical criteria that were used to determine significant effects.

Effect Importance

Typically, standard p and α values are used to determine the significance of particular effects in RM ANOVA (between-subjects main effects, within-subjects main effects, and interactions). This strategy was not employed in the present study. In both parts of the investigation the relatively large total (n=135,000 for Study 1, n=120,000 for Study 2) and cell sample sizes

precluded the use of significance levels to determine meaningful effects. This was due to the fact that with large sample sizes small effects will be more likely to be detected as statistically significant. These potential small effects would provide little practical guidance in terms of delineating meaningful differences in parameter recovery between TESTFACT and NOHARM. Instead, it is important to determine the practical significance of an effect and use this as an index of the degree of departure from the null hypothesis.

As outlined by Cohen (1988, 1992) and Prentice and Miller (1992) effect sizes, which are usually based on standardized mean differences, can be used to assess the importance of the relationships between a set of independent and dependent variables. Cohen (1988) suggests that for multiple regression and correlational analysis, where a quantitative dependent variable is studied in relation to one or more independent factors, small, medium and large effect sizes would be of the order .02, .15 and .35, respectively. In the present study, where small differences in parameter recovery between LI and FI estimation would probably be of little practical significance, only medium and large ES values were interpreted.

Cohen (1992) operationally defines a number of ES measures that are appropriate for any given statistical test. For the RM ANOVAs the ES measure should be based on the proportion of variance of the dependent variable accounted for by the source, or sources, under study. For general data-analytic systems such

as multiple regression a measure of ES based on partial *eta* squared, which is applicable to all F and t-tests, can be calculated and used to assess the degree to which the phenomenon under study is manifested. This strategy, combined with utilization of the defined conventions for small, medium, and large ES values, was employed in the interpretation of the RM ANOVA results.

Statistical Results (RM ANOVA)

The results of the repeated measures analyses are reported here. As described above, two RM ANOVAS were run for each of the experiments (Study 1, Study 2). In the first analysis, for each experiment, the differences in the recovery of the difficulty parameter between the two estimation methods was explored. The second analysis was used to investigate differences in the recovery of the discrimination parameter between FI- and LI estimation. For all analyses, outlier values were rescored as described previously. In terms of the determination of significant effects, only effect sizes greater than .15 are reported and interpreted.

Ratio Experiments (Study 1)

An interpretation of the statistical results from the analysis of Study 1 are presented below. The results concerning the recovery of the difficulty and discrimination parameter are

reported separately. All of the analyses were based on the rescaled parameter values (i.e., discrimination estimates that exceeded 4.5 were rescaled to 4.5, difficulty estimates that were greater than 4.5 or less than -4.5 were rescaled to 4.5 and -4.5, respectively).

Difficulty values. The meaningful within subject effects for the comparison of difficulty estimates between TESTFACT and NOHARM are reported in Table 14. As stated previously, the within-subject factor is the program type (estimation method) which has two levels (TESTFACT, NOHARM). Results are given for the main effect for estimator and the significant interactions of estimator with the other modelled factors (i.e., distribution, item difficulty level).

Table 14

Tests of Hypotheses for Within Subject Effects for the Difficulty Parameter Estimates (Study 1)

Effect	df	F	ES (partial ETA^2)
Estimator	1,134865	133362	.497
Estimator x dist.	2,134865	39732	.370
Estimator x b	4,134865	54071	.620
Estimator x dist. x b	8,134865	13624	.446

The significant estimator effect indicates that, averaged over all the independent factors (i.e., distribution, sample size/test length ratio, population a and b values), there is a

difference in overall parameter recovery between NOHARM and TESTFACT. While recovery of the difficulty parameter was dependent on both the underlying trait distribution (estimator x distribution) and the population difficulty value (estimator x b), the interpretation of these effects is confounded by their interaction (estimator x distribution x b). This interaction indicates that the differences in the recovery of the difficulty parameter for NOHARM and TESTFACT can be attributed to some combinations of population b values and the three modelled trait distributions. Neither the sample size/test length ratio nor the population discrimination value were important in explaining differences in parameter recovery between the two estimation methods.

The significant two-way effects also provide information concerning the sources of estimation differences between NOHARM and TESTFACT. The significant estimator x distribution effect suggests that, averaged over other conditions, differences in parameter recovery between LI and FI estimation can be credited to the underlying trait distribution. Likewise, the significant estimator x b effect indicates that, averaged over the modelled trait distributions, estimation differences between NOHARM and TESTFACT can be attributed to the population value of the difficulty parameter.

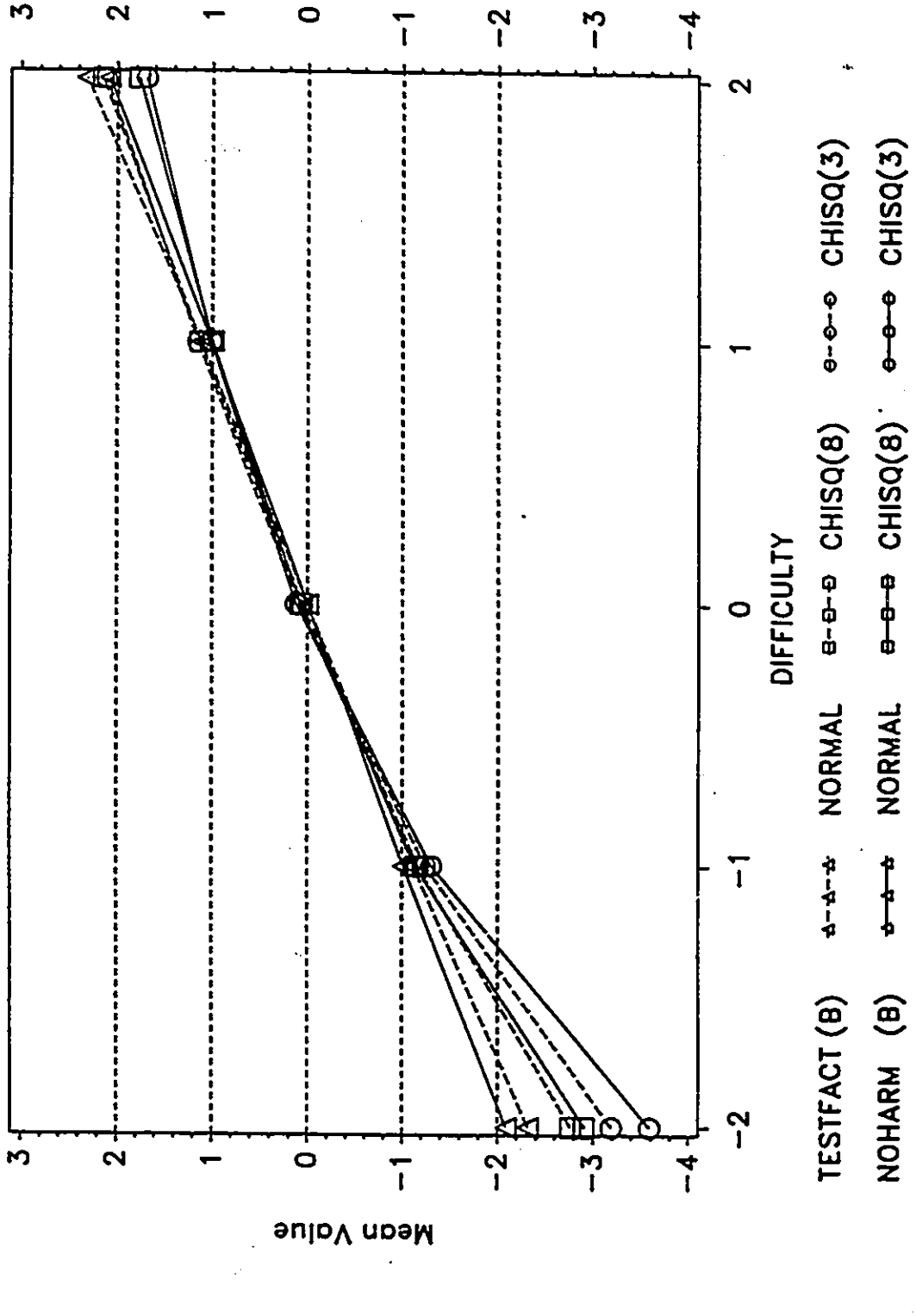
Figure 2 shows the mean values for the difficulty estimates by estimator, population difficulty values and the modelled trait distribution. These mean values are based on the rescaled data.

Visual inspection of Figure 2 shows that differences in parameter recovery resulted predominantly from items with population b values at the extremes of the difficulty range (b=-2,2). Moreover, for easy items, combined with nonnormal trait distributions, LI estimation produced more biased difficulty estimates than did FI estimation. That is, compared to the population value of b=-2, the estimates produced by NOHARM were, on average, more negatively biased than those produced by TESTFACT. For the more difficult items (b=2), and a nonnormal trait distribution, TESTFACT produced difficulty estimates that were, on average, positively biased. In contrast, under identical conditions, NOHARM produced difficulty estimates that were negatively biased.

These differences, across the population b values, account for the significant estimator x distribution x b effect noted previously. For both TESTFACT and NOHARM, parameter recovery was reasonably accurate for conditions involving a normal trait distribution. Furthermore, parameter recovery was reasonably unbiased for items of moderate difficulty (b=-1,0,1), regardless of the modelled trait distribution.

Figure 2

Mean Values for the Difficulty Estimates by Estimation Method, Population Difficulty Values and Trait Distribution (Study 1)



Discrimination values. The meaningful within subject effects for the comparison of discrimination estimates between TESTFACT and NOHARM are reported in Table 15. Similar to the results described for the difficulty parameter, differences in the estimation of the discrimination parameter were dependent on a number of factors. While some individual conditions (e.g., population a and b values), led to differences in parameter estimation between NOHARM and TESTFACT, more complex interactions of factors (i.e., estimator x a x b x distribution) are needed in order to fully describe estimation differences between the two programs. It should also be noted that the results of the RM analysis indicated that the ratio of sample size to the number of items did not, either independently or in combination with other independent factors, result in meaningful differences in the recovery of the discrimination parameter between the two programs.

Table 15

Tests of Hypotheses for Within Subject Effects for the Discrimination Parameter Estimates (Study 1)

Effect	df	F	ES (partial ETA^2)
Estimator	1,134865	139274	.508
Estimator x a	2,134865	38639	.364
Estimator x b	4,134865	44056	.567
Estimator x dist. x b	8,134865	13390	.443
Estimator x a x b	8,134865	16635	.497
Estimator x a x b x dist	16,134865	5258	.384

Figures 3a, 3b, and 3c display the mean values for the discrimination estimates by population a value, population b value, and trait distribution.

Figure 3a shows the mean values under the normal specification. The comparison of the estimates from TESTFACT and NOHARM suggests that, while the mean discrimination values for TESTFACT tend to be negatively biased, especially for items with population discrimination values of 1.5, the recovery of the discrimination parameter tends to be fairly consistent across all population b values. At extreme population b values there would, however, appear to be evidence of negative bias in the recovery of both the TESTFACT and NOHARM estimates, especially for population discrimination values of 1 and 1.5. The comparison of the LI and FI estimates with the corresponding population values suggests that NOHARM provides more accurate discrimination

estimates, especially when the population discrimination values are 1.0 or 1.5. Under these conditions TESTFACT consistently produces discrimination estimates that, on average, underestimate the population value. However, for items with population discrimination values of .5, both TESTFACT and NOHARM produce discrimination estimates with little or no bias.

When the assumed trait distribution is mildly (positively) skewed (χ^2_8) the use of LI estimation tends to produce inflated (positively biased) discrimination estimates, especially for more difficult (b=1,2) discriminating (a=1,1.5) items (see Figure 3b). In contrast, FI estimation produces relatively more accurate discrimination estimates, especially for difficult, discriminating items. Nevertheless, these FI estimates still remain negatively biased. For both estimators, under conditions involving poorly discriminating items (a=.5), parameter recovery did not appear to be overly dependent on the population difficulty value. Similar to the results presented for the normal trait distribution (Figure 3a), parameter recovery was more accurate for items in the moderate difficulty range (b=-1,0,1).

Figure 3a

Mean Values for the Discrimination Estimates by Estimation Method and Population Difficulty Values (Normal Trait Distribution)

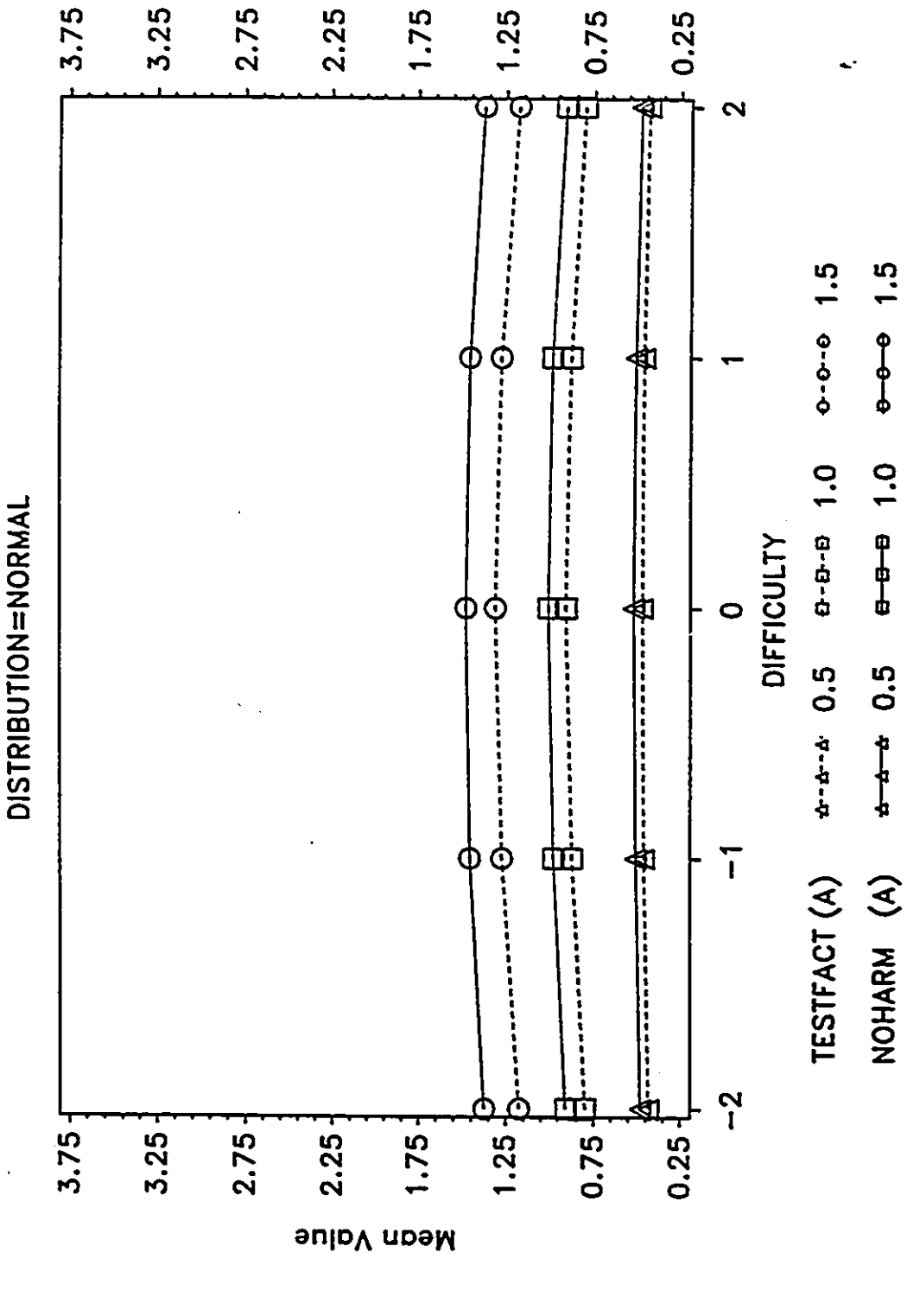


Figure 3b

Mean Values for the Discrimination Estimates by Estimation Method and Population Difficulty Values ($\chi^2_{(8)}$ Trait Distribution)

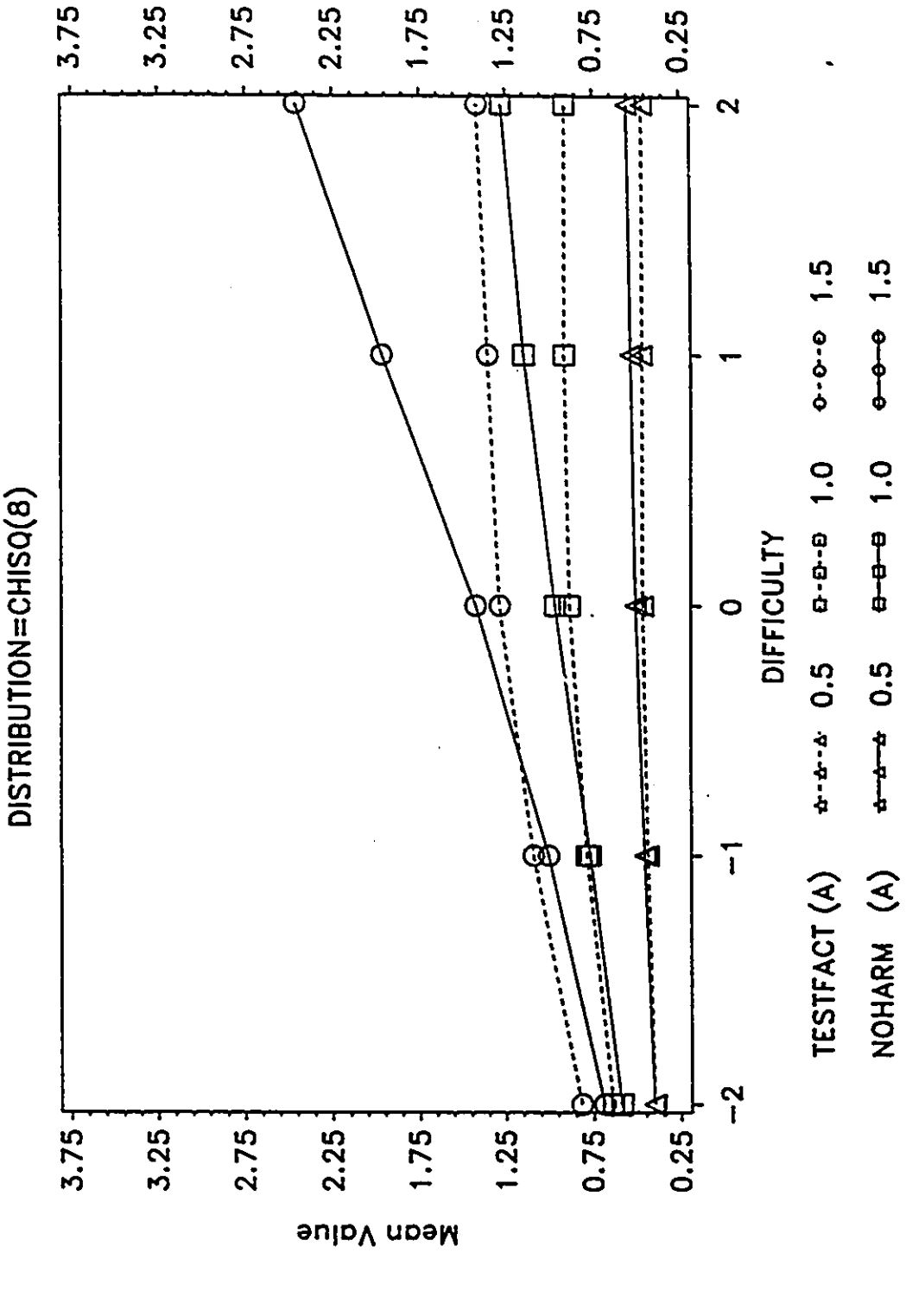
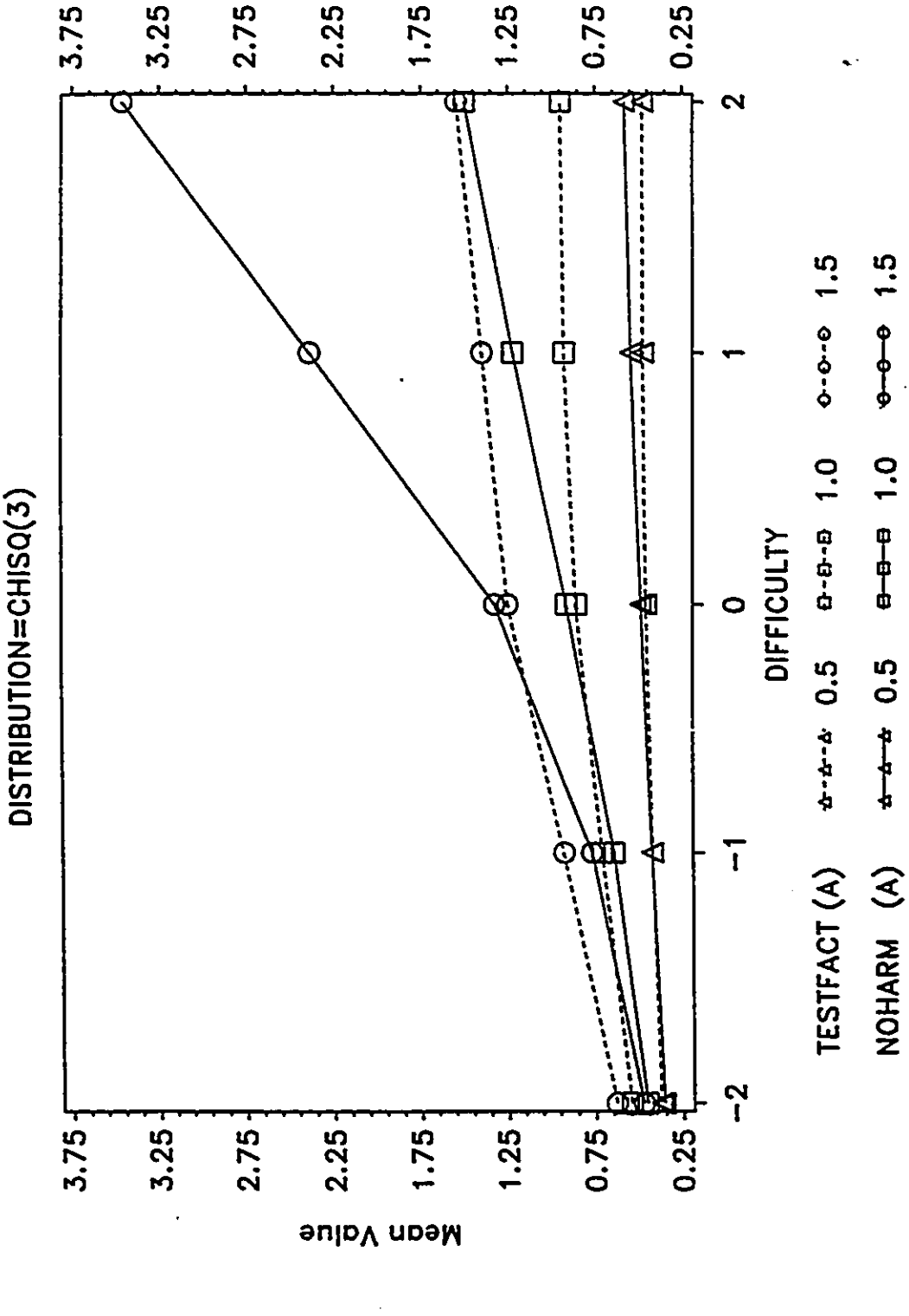


Figure 3c

Mean Values for the Discrimination Estimates by Estimation Method and Population Difficulty Values ($X^2_{(3)}$ Trait Distribution)



The mean values for the discrimination estimates over population a and b values, under an extremely skewed trait distribution (χ^2_3) are shown in Figure 3c. Similar to the results presented in Figure 3b, LI estimation produced inflated discrimination estimates, most markedly for highly discriminating (a=1.5), difficult items (b=2). While parameter recovery using FI estimation was still biased under certain conditions (e.g., a=1.5, b=-2), the discrepancies between the mean values and the population values were generally less than those resulting from LI estimation. For poorly discriminating items (a=.5), or items with average difficulty (b=0), parameter recovery was consistent, and reasonably accurate, for the two estimation strategies.

Sample Size Experiments (Study 2)

An interpretation of the statistical results from the analysis of Study 2 is presented below. Similar to the analyses described for Study 1, the results are presented separately for the difficulty and discrimination estimates.

Difficulty values. The meaningful within subject effects for the comparison of discrimination estimates between TESTFACT and NOHARM are reported in Table 16. Similar to the results presented for the sample size/test length ratio experiment (study 1), differences in the recovery of the difficulty parameter were dependent on the estimation strategy, the population b value and the modelled trait distribution. The estimator x distribution x

b effect indicates that differences in the recovery of the difficulty parameter between LI and FI estimation are dependent on some combination of the trait distribution and population b value. The significant lower-order effects (estimator, estimator x b, estimator x distribution) are identical to those found in the analysis of Study 1, and are not interpreted here. Neither sample size nor the number of items could be used to explain differences in the recovery of the difficulty parameter between the two programs.

Table 16

Tests of Hypotheses for Within Subject Effects for the Difficulty Parameter Estimates (Study 2)

Effect	df	F	ES (partial ETA^2)
Estimator	1,119520	45176	.270
Estimator x dist.	2,119520	45625	.276
Estimator x b	4,119520	16043	.349
Estimator x dist. x b	8,119520	14930	.333

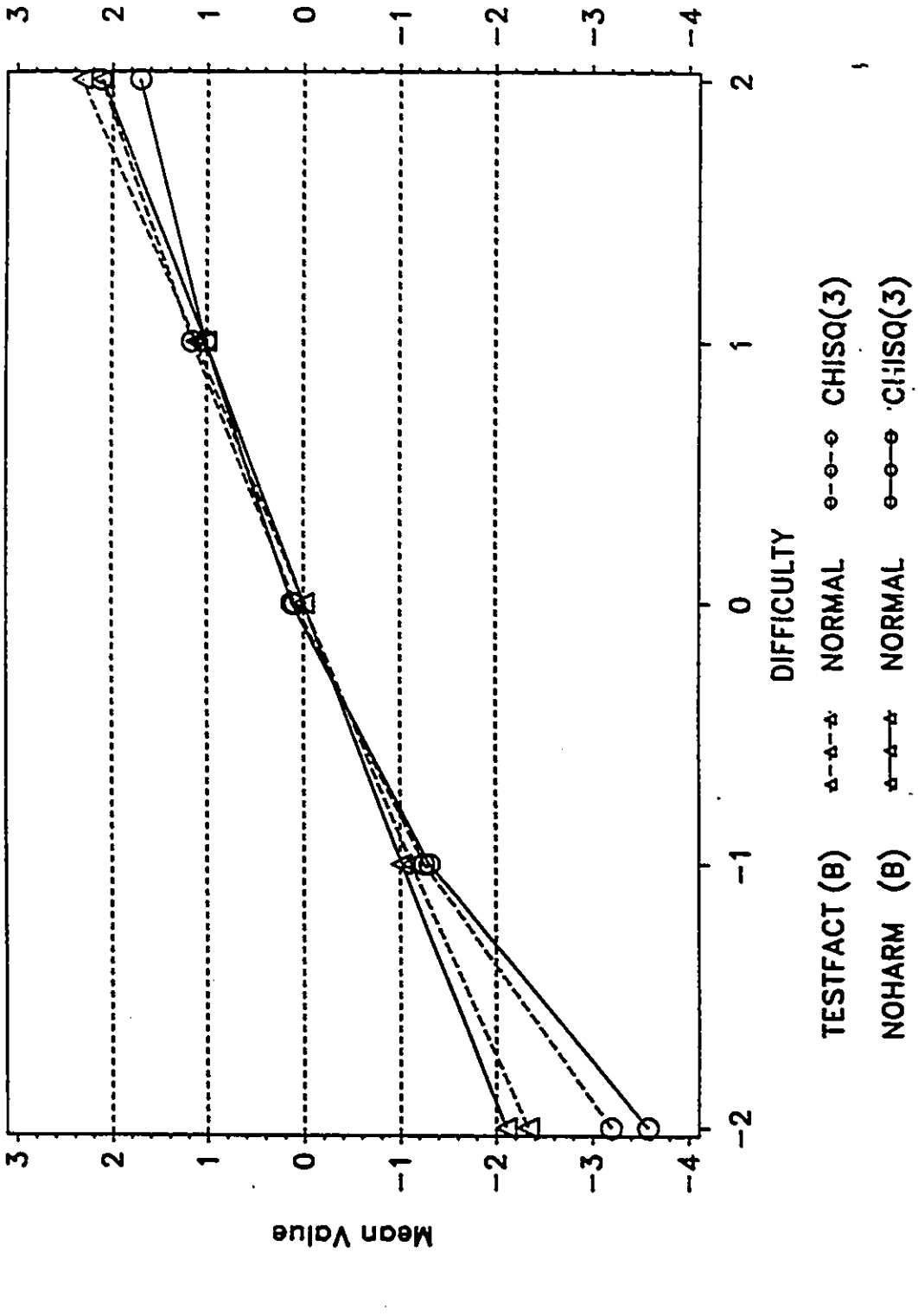
The mean values for the difficulty estimates by estimation strategy, population difficulty value and distribution are shown in Figure 4. It is readily apparent that, for items of extreme difficulty (i.e., b=-2) and a nonnormal trait distribution, both LI and FI produce negatively biased difficulty estimates. Under these conditions, NOHARM produces less accurate difficulty estimates. For more difficult items (b=2), and a nonnormal trait distribution, NOHARM produces somewhat downwardly biased

difficulty estimates whereas TESTFACT produces estimates that are slightly positively biased. The patterns noted above account for the significant estimator x distribution x b effect. While not directly displayed in Figure 4, the lower-order effects can also be interpolated. For example, averaging over trait distribution, there are differences in parameter recovery by population b values. This results in the significant estimator x b effect. Likewise, the significant estimator x distribution effect can be investigated by averaging the results, for each estimator, over the population difficulty values.

Overall, recovery of the difficulty parameter, under a normal trait distribution, is reasonably accurate for both NOHARM and TESTFACT. Moreover, for items of medium difficulty (b=-1,0,1), and either a normal or nonnormal trait distribution, recovery of the difficulty parameter was reasonably accurate, regardless of the estimation strategy that was employed. This can be seen in Figure 4 by inspecting the closeness of the mean of the estimates to the population values under the conditions that exclude extremely difficult (b=2) or extremely easy items (b=-2).

Figure 4

Mean Values for Difficulty Estimates by Estimation Method, Distribution and Population Difficulty Values (Study 2)



Discrimination values. The important within-subject effects for the comparison of discrimination estimates between TESTFACT and NOHARM are presented in Table 17. Neither sample size nor test length explain differences in the recovery of parameter estimates between NOHARM and TESTFACT. The large ES value for the estimator x distribution x a x b interaction indicates that some combination of these factors can be used to explain differences in parameter recovery between the two programs.

Table 17

Tests of Hypotheses for Within Subject Effects for the Discrimination Parameter Estimates (Study 2)

Effect	df	F	ES (partial ETA^2)
Estimator	1,119520	53971	.310
Estimator x b	4,119520	15988	.349
Estimator x t x b	4,119520	16001	.348
Estimator x a x b	8,199520	5946	.285
Estimator x dist. a x b	8,119520	5910	.283

The mean values for the discrimination estimates over the population parameter values and trait distribution are plotted in Figures 5a and 5b. Mean values for conditions involving a normal trait specification are presented in Figure 5a; mean values under the nonnormal specification are presented in Figure 5b.

Figure 5a

Mean Values for Discrimination Estimates by Estimation Method and Population a and b Values (Normal Trait Distribution)

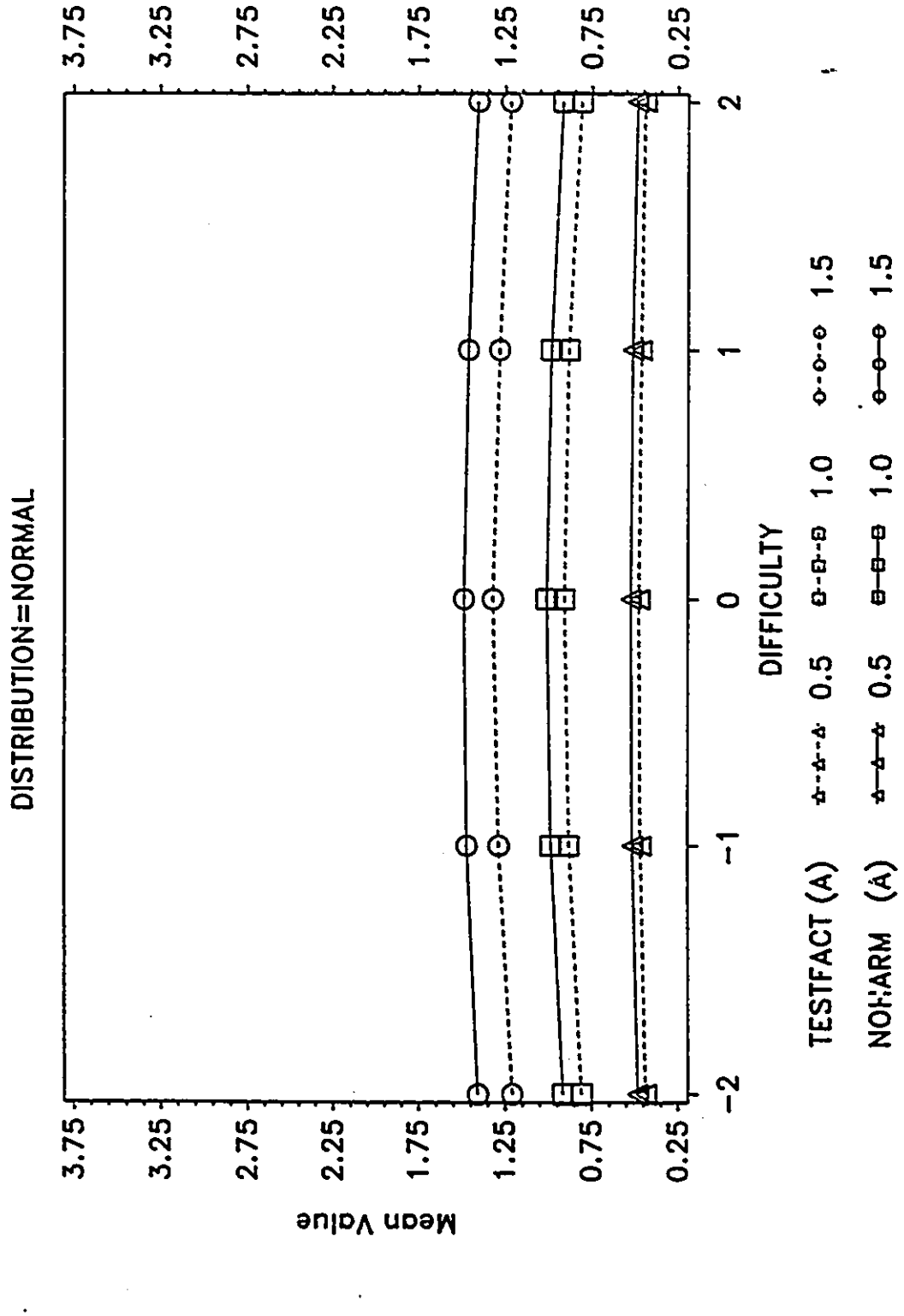
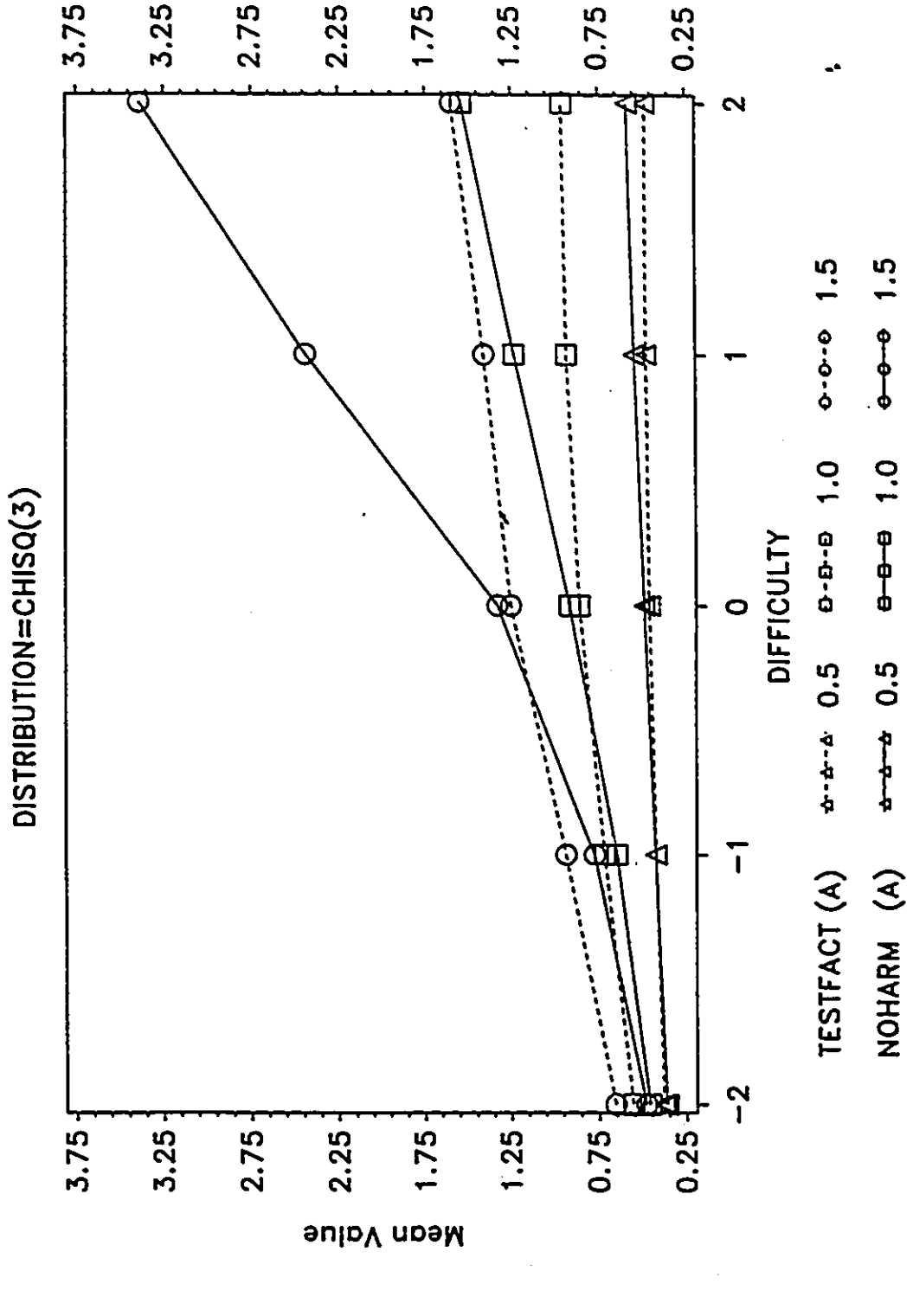


Figure 5b

Mean Values for Discrimination Estimates by Estimation Method and Population a and b Values ($X^2_{(3)}$ Trait Distribution)



From Figure 5a it is evident that, under a normal specification for the latent trait, both LI and FI recover the discrimination parameter reasonably well. Nevertheless, it is noteworthy that FI systematically underestimates the population discrimination value, especially for items with population discrimination values of 1.5. Likewise, differences in parameter recovery were slightly greater for conditions involving items with extreme population difficulty values. This pattern accounts for the significant estimator x a x b effect noted previously. For the nonnormal trait distribution both LI and FI produce biased estimates of the discrimination parameter, more so for highly discriminating, difficult, items (see Figure 5b). Under these conditions (a=1.5,b=2) LI produces the least accurate estimates. In general, FI estimation yields more precise discrimination estimates under conditions involving a nonnormal trait distribution. Nevertheless, for items with population difficulty values of zero, both estimators recover the discrimination parameter with little or no bias.

The significant estimator x distribution x a x b effect can be interpreted by comparing the results from Figure 5a and Figure 5b. It is evident that differences in the recovery of the discrimination parameter are dependent on combinations of the trait distribution, the population discrimination value and the population difficulty value. For example, when comparing the results across conditions involving extreme population difficulty values (b=-2,2), it is evident that differences in the recovery

of the discrimination parameter vary by both the trait distribution and the population discrimination value. For the skewed trait distribution, differences in LI and FI estimation vary profoundly as a function of the population discrimination and difficulty values. In contrast, where the trait distribution is normal, there are relatively minor, and consistent, differences in parameter recovery across the a x b conditions.

Summary of Statistical Results (RM ANOVA)

The RM ANOVAs indicated the important conditions that led to differences in parameter recovery between LI and FI estimates. Furthermore, an inspection of the means of the estimates, over these conditions, uncovered the nature and strength of the bias in parameter estimation. Based on the repeated measures analyses, neither the sample size, the test length nor the sample size/test length ratio were important in terms of explaining differences in parameter recovery between LI- and FI estimation. For these analyses, parameter recovery was based on the mean of the rescaled data.

The RM analyses indicated that differences in the recovery of the difficulty parameter, between TESTFACT and NOHARM, were attributable to the modelled trait distribution, the population item difficulty, and the combination of these two factors. An inspection of the means of these estimates, over this combination of factors, showed that the estimation of the difficulty

parameter was reasonably accurate for conditions involving a normal trait distribution. For both TESTFACT and NOHARM there were only minor differences between the means of the estimates and the population difficulty value. However, for conditions involving a nonnormal trait distribution, and easy (b=-2) or difficult (b=2) items, there were clear differences in the accuracy of the difficulty estimates across estimation method. Differences in the strength and direction of the bias in these estimates accounts for the significant estimator x distribution x population difficulty interaction that was found. Overall, for items of moderate difficulty (b=-1,0,1), and either a normal or skewed trait distribution, parameter recovery was reasonably accurate, regardless of the estimation strategy that was employed.

Differences in the recovery of the discrimination parameter between NOHARM and TESTFACT were dependent on the modelled trait distribution and the population item (a,b) values. Under conditions involving a normal trait distribution, NOHARM produced reasonably accurate discrimination estimates, across all population levels of a and b. In contrast, TESTFACT produced negatively biased estimates, especially for more highly discriminating items (a=1,1.5). For conditions involving nonnormal trait distribution, NOHARM produced highly discrepant mean discrimination values, especially for conditions involving population parameter values of a=1.5 and b=2. Overall, the impact of the nonnormal trait distribution was more pronounced

for LI estimation than for FI estimation. Nevertheless, bias in parameter recovery was still apparent for both estimators under conditions involving highly discriminating (a=1.5), easy (b=-2) and difficult (b=+2) items.

Efficiency

The results from the outlier and RM analyses provide a detailed profile of parameter recovery differences between NOHARM and TESTFACT. Furthermore, the inspection of the means of the estimates, across conditions of importance, indicates where parameter estimates may be unbiased. However, for conditions in which both estimators provide reasonably equivalent, or unbiased, results, a measure of error (e.g., RMSE) is needed in order to discern which estimation method is more efficient. However, as stated previously, the value of the RMSE will be contaminated by bias. Therefore, for conditions where bias exists, the RMSE can only be used as a global measure of goodness-of-fit.

A comparison of the error of the LI and FI estimates is presented in this section. For the most part, comparisons are made across conditions where both estimators recover the parameters equally well. In some instances, however, RMSE measures are contrasted under conditions where one of the estimators was shown to produce biased results. This strategy was employed for two reasons. First, if an estimator produced biased results consistently for a set of conditions then the

deviation between the population and estimated parameters could be easily corrected for in item calibration studies. Hence, the choice of a superior estimation method may be augmented by an indication of the error, or spread, in the estimates that are produced. Second, for some conditions, it appeared that the estimates were only biased for extreme population item values (i.e., a=1.5,b=-2,+2). Thus, a comparison of the goodness-of-fit of the estimates across conditions involving items of moderate difficulty (b=-1,0,1) and lower discrimination (a=0.5,1.0) will provide additional information with which to judge the appropriateness of a given estimation method under a specific subset of conditions.

The RM analyses, which were based rescaled outlier values, could possibly have yielded some misleading information regarding the true discrepancy between the estimates and the population values. For example, based on the outlier analysis, it was found that NOHARM yielded substantial numbers of discrepant discrimination estimates under conditions where the trait distribution was not normal and the population item parameters were extreme (i.e., a=1.5 and b=2). Therefore, depending on whether one scaled these discrimination values to 4.5 or used nonparametric summaries in the data analysis, slightly different results could possibly emerge. In order to investigate the effect of rescaling the outlier values, descriptive statistics, based on the median of the raw data and the mean of the rescaled data, are also presented for some conditions of interest.

Difficulty Estimates

Ratio Experiments (Study 1)

The results from the RM-ANOVA suggested that both the FI and LI estimators produced reasonably unbiased difficulty estimates when the modelled trait distribution is normally distributed. However, under nonnormal trait specifications, combined with extreme initial difficulty values (e.g., b=-2), LI estimation produced estimates that were more biased than those produced through FI estimation. Based on these results, it is important to investigate the error of the LI and FI difficulty estimates, especially under conditions involving a normal trait distribution, where both programs produced reasonably unbiased parameter estimates.

The median, mean, and RMSE values for the difficulty estimates, by distribution and population item difficulty, are presented in Table 18. As mentioned previously, the mean and RMSE summaries are based on the rescaled data (i.e., difficulty values that were greater than 4.5 or less than -4.5 were rescaled to 4.5 and -4.5, respectively). It is evident that, based on the median of the original data and the mean of rescaled estimates, both NOHARM and TESTFACT produce reasonably unbiased difficulty parameter estimates when the modelled trait distribution is normal.

Table 18

Descriptive Statistics for Difficulty Estimates by Distribution and Population b (Study 1)

Dist.	b	Md (NHb)	Md (TFb)	Mean* (NHb)	Mean* (TFb)	RMSE* (NHb)	RMSE* (TFb)
Normal	-2	-2.06	-2.32	-2.08	-2.32	0.23	0.40
Normal	-1	-1.01	-1.12	-1.01	-1.12	0.10	0.16
Normal	0	0.00	0.00	0.00	0.00	0.06	0.07
Normal	1	1.01	1.12	1.01	1.12	0.10	0.16
Normal	2	2.06	2.31	2.08	2.31	0.23	0.40
$X^2_{(8)}$	-2	-2.89	-2.74	-2.91	-2.77	1.03	0.87
$X^2_{(8)}$	-1	-1.14	-1.17	-1.13	-1.16	0.18	0.20
$X^2_{(8)}$	0	0.09	0.06	0.08	0.05	0.10	0.09
$X^2_{(8)}$	1	0.99	1.13	0.99	1.13	0.09	0.17
$X^2_{(8)}$	2	1.74	2.16	1.77	2.15	0.35	0.24
$X^2_{(3)}$	-2	-3.66	-3.13	-3.74	-3.24	1.75	1.34
$X^2_{(3)}$	-1	-1.31	-1.25	-1.30	-1.24	0.35	0.28
$X^2_{(3)}$	0	0.14	0.10	0.13	0.09	0.15	0.12
$X^2_{(3)}$	1	1.01	1.16	1.01	1.15	0.09	0.19
$X^2_{(3)}$	2	1.65	2.12	1.69	2.11	0.36	0.25

* Statistic is based on the rescaled estimates

However, even under a normal specification, FI estimation results in some scale expansion for the difficulty estimates. That is, difficult items (b=2), on average, are estimated as being more difficult (b > 2) whereas easy items (b=-2) are, on average, estimated as being less difficult (b < -2). This trend is important in that any summary measure based on data collapsed over population difficulty values will potentially average out the bias in the estimates.

Based on summary measures in Table 18, a number of other observations can be made. First, the comparison of the RMSE values for the NOHARM and TESTFACT difficulty estimates suggests that NOHARM provides more efficient estimation of the difficulty parameter under conditions involving a normal trait specification. For this condition the LI estimator produces difficulty estimates with less error than does TESTFACT. Second, conclusions regarding the potential bias of a particular estimator are consistent, regardless of the parametric (mean) or nonparametric (median) statistic that is used. Third, bias and error (RMSE) in the estimates tends to increase as modelled trait distribution deviates from normal. Fourth, for both NOHARM and TESTFACT, when extreme population parameter estimates are eliminated (i.e., b=2,-2), the estimated values lie reasonably close to the population values, more so for conditions involving a normal trait distribution. Finally, for the normal trait specification, error increases, for both LI and FI estimation, as the modelled population item difficulty deviates from zero.

Sample Size and Test Length (Study 2)

The RM analyses suggested that neither test length nor sample size had a profound effect on the comparative accuracy of the NOHARM and TESTFACT results. Test length and sample size did, however, have some effect on the error of the difficulty estimates (see Table 19). The effect of test length on the spread of the difficulty estimates can be investigated by

inspecting the RMSE values across the four sets of items within each fixed sample size. Based on the examination of the RMSEs, there did appear to be an effect due to test length, but only for the TESTFACT estimates under a normal trait specification. For fixed sample sizes, and a normal trait distribution, the RMSE values for the FI estimates increased with expanded test lengths. In contrast, there did not appear to be a strong effect due to test length on the error of the LI estimates. For conditions involving a nonnormal trait distribution, where both estimators were shown to produce estimates that deviated from the population values, the error of the estimates for both programs did not vary considerably as a function of test length. For a given test length, and a nonnormal trait specification, TESTFACT did, however, produce estimates with somewhat lower RMSEs than those obtained through NOHARM. However, given that NOHARM was shown to produce less accurate difficulty estimates under the nonnormal conditions, this result can at least be partially explained by the contaminating effects of the bias.

The effect of sample size on the error of the estimates can be examined by inspecting the RMSE values for a specific test length across the four sample sizes. For both the NOHARM and TESTFACT difficulty estimates, for a fixed test length and a normal trait distribution, the RMSE values decreased as the sample size increased (see Table 19). This effect was not readily apparent for conditions involving the nonnormal trait distribution.

Table 19

Descriptive Statistics for the Difficulty Estimates by Distribution, Sample Size and Test Length

Dist.	N	I	Md (NHb)	Md (TFb)	MEAN* (NHb)	MEAN* (TFb)	RMSE* (NHb)	RMSE* (TFb)
Normal	250	15	0.00	0.00	0.00	-0.01	0.36	0.36
Normal	250	30	0.01	0.02	0.00	0.00	0.33	0.36
Normal	250	45	0.00	0.01	0.01	0.01	0.34	0.43
Normal	250	60	0.02	0.01	0.01	0.00	0.35	0.47
Normal	500	15	0.00	0.00	-0.02	-0.02	0.25	0.25
Normal	500	30	0.01	0.01	0.01	0.01	0.24	0.29
Normal	500	45	0.00	0.00	0.00	0.00	0.22	0.32
Normal	500	60	-0.01	-0.01	0.00	0.00	0.22	0.37
Normal	1000	15	-0.00	-0.00	0.00	0.00	0.16	0.17
Normal	1000	30	0.00	0.00	0.00	0.00	0.16	0.21
Normal	1000	45	-0.00	-0.00	0.00	0.00	0.16	0.27
Normal	1000	60	0.00	0.00	0.00	0.00	0.15	0.32
Normal	10000	15	0.00	0.00	0.00	0.00	0.06	0.08
Normal	10000	30	-0.00	-0.00	0.00	0.00	0.06	0.14
Normal	10000	45	0.00	-0.00	0.00	0.00	0.06	0.22
Normal	10000	60	-0.00	-0.00	0.00	0.00	0.06	0.29
$X^2_{(3)}$	250	15	0.12	0.11	-0.40	-0.30	0.86	0.77
$X^2_{(3)}$	250	30	0.16	0.12	-0.39	-0.25	0.87	0.75
$X^2_{(3)}$	250	45	0.14	0.09	-0.39	-0.21	0.85	0.74
$X^2_{(3)}$	250	60	0.14	0.09	-0.40	-0.19	0.84	0.72
$X^2_{(3)}$	500	15	0.14	0.12	-0.39	-0.30	0.80	0.72
$X^2_{(3)}$	500	30	0.14	0.10	-0.40	-0.24	0.81	0.66
$X^2_{(3)}$	500	45	0.14	0.09	-0.41	-0.22	0.83	0.67
$X^2_{(3)}$	500	60	0.14	0.09	-0.40	-0.19	0.83	0.68

Table 19 (cont.)

Dist.	N	I	Md (NHb)	Md (TFb)	MEAN* (NHb)	MEAN* (TFb)	RMSE* (NHb)	RMSE* (TFb)
$X^2_{(3)}$	1000	15	0.13	0.12	-0.40	-0.31	0.81	0.72
$X^2_{(3)}$	1000	30	0.14	0.10	-0.41	-0.24	0.82	0.64
$X^2_{(3)}$	1000	45	0.15	0.11	-0.40	-0.20	0.82	0.62
$X^2_{(3)}$	1000	60	0.14	0.09	-0.41	-0.18	0.81	0.63
$X^2_{(3)}$	10000	15	0.14	0.13	-0.42	-0.31	0.82	0.71
$X^2_{(3)}$	10000	30	0.14	0.10	-0.42	-0.24	0.82	0.59
$X^2_{(3)}$	10000	45	0.13	0.10	-0.42	-0.20	0.82	0.57
$X^2_{(3)}$	10000	60	0.14	0.10	-0.42	-0.18	0.82	0.59

* Statistic based on the rescaled estimates

Note: median/mean of the population values equals 0.00

Overall, comparing the NOHARM and TESTFACT results, LI produced estimates with lower RMSEs under most conditions involving the normal trait specification. This is important in that both the LI and FI estimators performed reasonably well in terms of parameter recovery under conditions involving a normal trait specification. Therefore, LI can be said to be more efficient for difficulty parameter estimation under conditions involving a normal trait specification. Under the condition involving a nonnormal trait specification both LI and FI produced biased difficulty estimates. However, although both programs performed poorly in recovering the b parameter under these conditions, FI estimation was comparatively more accurate, or less biased. Furthermore, under conditions involving the

nonnormal trait distribution, the error of the difficulty estimates was almost always greater for LI estimation, regardless of sample size or test length. This finding will be due, in part, to the greater bias in the NOHARM estimates. Nevertheless, based on the conditions modelled in this study, it would appear that FI estimation of the difficulty parameter would be preferred under conditions involving a nonnormal trait distribution.

Discrimination Estimates

Ratio Experiment (Study 1)

The median, mean, and RMSE values for the discrimination estimates, by distribution and population item difficulty, are presented in Table 20. It is evident that, based on the median and the mean of the estimates, NOHARM produces reasonably unbiased discrimination parameter estimates when the modelled trait distribution is normal. However, under the normal trait specification, FI estimation leads to downwardly biased discrimination estimates, regardless of the population difficulty value. Furthermore, the bias in the discrimination estimates is greater for extreme (i.e., b=+2,-2) difficulty values. This result is borne out by both the parametric and nonparametric summaries. For conditions involving a nonnormal trait distribution, both NOHARM and TESTFACT yield biased parameter estimates. In addition, as shown previously, the discrepancy between the estimates and the population value is dependent on

the population item difficulty. For both FI and LI estimation, parameter recovery was the least accurate for easy items (b=-2). Furthermore, for NOHARM, upwardly biased discrimination estimates (estimates > population values) resulted from conditions in which the items were difficult (b=1,2).

Table 20

Descriptive Statistics for Discrimination Estimates by Distribution and Population b (Study 1)

Dist.	b	Md (NHa)	Md (TFa)	Mean* (NHa)	Mean* (TFa)	RMSE* (NHa)	RMSE* (TFa)
Normal	-2	0.90	0.78	0.92	0.80	0.20	0.27
Normal	-1	0.97	0.86	0.98	0.87	0.12	0.18
Normal	0	1.00	0.89	1.00	0.90	0.09	0.15
Normal	1	0.98	0.86	0.98	0.87	0.11	0.18
Normal	2	0.90	0.78	0.92	0.81	0.19	0.27
$X^2_{(8)}$	-2	0.57	0.62	0.56	0.62	0.54	0.46
$X^2_{(8)}$	-1	0.77	0.78	0.75	0.77	0.32	0.28
$X^2_{(8)}$	0	0.97	0.87	0.97	0.88	0.10	0.17
$X^2_{(8)}$	1	1.14	0.90	1.21	0.92	0.35	0.17
$X^2_{(8)}$	2	1.25	0.88	1.43	0.93	0.70	0.24
$X^2_{(3)}$	-2	0.43	0.52	0.43	0.52	0.68	0.58
$X^2_{(3)}$	-1	0.64	0.71	0.61	0.69	0.47	0.38
$X^2_{(3)}$	0	0.91	0.85	0.91	0.86	0.14	0.19
$X^2_{(3)}$	1	1.21	0.91	1.38	0.93	0.60	0.17
$X^2_{(3)}$	2	1.47	0.91	1.85	0.99	1.27	0.33

* Statistic based on the rescaled estimates

Note: mean/median of the population values equals 1.00

The RMSE summaries in Table 20 indicate that the error in both the NOHARM and TESTFACT discrimination estimates is dependent on the trait distribution. Inspection of the RMSE values indicates that greater error is associated with the most skewed trait distribution. However, the most skewed trait distributions also yielded the most biased discrimination estimates. For both estimators, under conditions involving a nonnormal trait distribution, the RMSE values were generally lowest for conditions involving items with moderate population item difficulties (i.e., b=-1,0,1). A comparison of the RMSE values, under conditions involving a normal trait specification, suggests that LI estimation provides less dispersed estimates across all population difficulty values. While this observation can be attributed, in part, to the fact that TESTFACT consistently produced downwardly biased discrimination estimates under these modelled conditions, it suggests that LI estimation of the discrimination parameter is preferred under conditions involving a normal trait distribution. Descriptive summaries for the discrimination estimates are also presented for combinations of the population item parameters (see Table 21). Since both NOHARM and TESTFACT yielded somewhat biased discrimination estimates under conditions involving nonnormal trait distributions, and the combined effects of distribution and other factors has previously been discussed, only the results for conditions involving a normal trait distribution are presented. The mean and median summaries, by combinations of population item

parameters, indicate that discrimination estimates derived under conditions involving extreme population values (e.g., a=1.5,b=-2,+2) are less accurate than those produced under more moderate combinations of population a and b values. Furthermore, as shown previously, the estimates from FI estimation tend to be downwardly biased, especially for easy (b=-2) and difficult items (b=2).

Table 21

Descriptive Statistics for Discrimination Estimates by a and b (Study 1)

Dist.	a	b	Md (NHa)	Md (TFa)	Mean* (NHa)	Mean* (TFa)	RMSE* (NHa)	RMSE* (TFa)
Normal	0.5	-2	0.48	0.43	0.48	0.43	0.06	0.09
Normal	0.5	-1	0.51	0.45	0.51	0.46	0.05	0.06
Normal	0.5	0	0.52	0.47	0.52	0.47	0.06	0.06
Normal	0.5	1	0.51	0.46	0.51	0.46	0.05	0.07
Normal	0.5	2	0.48	0.43	0.48	0.44	0.06	0.09
Normal	1.0	-2	0.90	0.78	0.91	0.80	0.15	0.23
Normal	1.0	-1	0.97	0.86	0.98	0.87	0.09	0.15
Normal	1.0	0	1.00	0.89	1.01	0.91	0.08	0.13
Normal	1.0	1	0.98	0.86	0.98	0.88	0.09	0.15
Normal	1.0	2	0.90	0.79	0.91	0.80	0.15	0.23
Normal	1.5	-2	1.34	1.13	1.65	1.18	0.31	0.40
Normal	1.5	-1	1.45	1.25	1.46	1.28	0.17	0.27
Normal	1.5	0	1.47	1.29	1.48	1.31	0.12	0.22
Normal	1.5	1	1.45	1.25	1.46	1.28	0.15	0.27
Normal	1.5	2	1.34	1.14	1.66	1.18	0.30	0.40

* Statistic based on the rescaled estimates

Based on the RMSE calculations shown in Table 21, NOHARM produced discrimination estimates with less error across all of the combinations of item population values. For both the TESTFACT and NOHARM estimates, error increased as the population discrimination value increased. Furthermore, error, within a fixed discrimination level, was less for items in the middle of the difficulty range (b=-1,0,1).

Sample Size and Test Length (Study 2)

Table 22 provides summary statistics for estimated discrimination values by distribution, sample size, and test length. Based on the medians of the estimates and the means of the rescaled estimates, there did not appear to be a strong effect of sample size on the recovery of the discrimination parameter. That is, for a fixed test length, the deviation of the median, or mean, of the discrimination estimates from the population value (1.0) did not vary substantially as a function of sample size. Furthermore, test length also did not appear to have an effect on the recovery of the discrimination parameter, except for FI estimation under conditions involving a normal trait distribution. For these situations, holding sample size constant, bias in the estimation of the discrimination parameter increased as the test length increased. Although the statistical analyses (RM ANOVAs) suggested that differences in recovery of the discrimination parameter were not attributable to the sample

size or the number of test items, TESTFACT, under an assumed normal trait specification, produced the least biased discrimination estimates when the test length was short (items=15). However, as shown previously, recovery of the

Table 22

Descriptive Statistics for the Discrimination Estimates by Distribution, Sample Size and Test Length

Dist.	N	Items	Md (NHa)	Md (TFa)	Mean* (NHa)	Mean* (TFa)	RMSE* (NHa)	RMSE* (TFa)
Normal	250	15	0.94	0.92	1.02	1.00	0.42	0.37
Normal	250	30	0.94	0.89	1.00	0.93	0.31	0.29
Normal	250	45	0.95	0.84	1.00	0.88	0.29	0.30
Normal	250	60	0.94	0.81	0.99	0.84	0.29	0.29
Normal	500	15	0.95	0.93	0.97	0.95	0.21	0.20
Normal	500	30	0.94	0.89	0.97	0.90	0.20	0.20
Normal	500	45	0.96	0.85	0.96	0.86	0.20	0.22
Normal	500	60	0.96	0.82	0.98	0.82	0.18	0.24
Normal	1000	15	0.96	0.94	0.96	0.94	0.16	0.15
Normal	1000	30	0.96	0.90	0.96	0.89	0.14	0.17
Normal	1000	45	0.96	0.85	0.96	0.84	0.14	0.21
Normal	1000	60	0.96	0.82	0.96	0.81	0.13	0.23
Normal	10000	15	0.96	0.95	0.95	0.93	0.09	0.11
Normal	10000	30	0.96	0.90	0.95	0.88	0.08	0.15
Normal	10000	45	0.96	0.86	0.95	0.84	0.08	0.20
Normal	10000	60	0.96	0.82	0.95	0.80	0.08	0.24
$X^2_{(3)}$	250	15	0.70	0.74	1.07	0.94	0.84	0.57
$X^2_{(3)}$	250	30	0.71	0.75	1.05	0.87	0.81	0.46
$X^2_{(3)}$	250	45	0.70	0.76	1.06	0.82	0.79	0.38
$X^2_{(3)}$	250	60	0.70	0.74	1.06	0.80	0.81	0.39

Table 22 (cont.)

Dist.	N	Items	Md (NHa)	Md (TFa)	Mean* (NHa)	Mean* (TFa)	RMSE* (NHa)	RMSE* (TFa)
$X^2_{(3)}$	500	45	0.68	0.74	1.04	0.80	0.77	0.35
$X^2_{(3)}$	500	60	0.68	0.73	1.04	0.77	0.74	0.34
$X^2_{(3)}$	1000	15	0.67	0.71	1.03	0.90	0.71	0.51
$X^2_{(3)}$	1000	30	0.66	0.74	1.04	0.82	0.74	0.33
$X^2_{(3)}$	1000	45	0.66	0.75	1.02	0.78	0.72	0.33
$X^2_{(3)}$	1000	60	0.66	0.74	1.04	0.76	0.74	0.33
$X^2_{(3)}$	10000	15	0.64	0.69	1.02	0.88	0.68	0.48
$X^2_{(3)}$	10000	30	0.64	0.72	1.02	0.81	0.68	0.32
$X^2_{(3)}$	10000	45	0.64	0.72	1.02	0.78	0.68	0.32
$X^2_{(3)}$	10000	60	0.64	0.70	1.03	0.75	0.68	0.33

* Statistic based on the rescaled estimates

discrimination parameter was more accurate for LI estimation under all conditions involving a normal trait distribution.

Based on the median values, FI estimation generally produced more accurate discrimination estimates under conditions involving a nonnormal trait specification. However, summaries based on the mean of the rescaled data, which effectively diminishes the effect of skewed distribution of discrimination estimates, suggest that NOHARM produces comparatively more accurate estimates of the discrimination parameter. Overall, it would appear that both estimators produce biased discrimination when the modelled trait distribution is not normal.

The comparison of the RMSE values shows that for conditions involving a normal trait distribution there is generally less error in the NOHARM discrimination estimates. Furthermore, for

both LI and FI estimation, sample size has an effect on the RMSE values. That is, larger sample sizes, holding test length constant, result in estimates with less error. This effect was more pronounced for the LI estimates. For NOHARM, error in the discrimination estimates did not appear to be dependent on the test length alone. In contrast, for TESTFACT, error generally increased as test length increased. This trend was not readily apparent for the smallest sample size (n=250), where outlier estimates may have partially contaminated the results.

For conditions involving a nonnormal trait distribution, there were minor decreases in error associated with increases in sample size. However, given that the discrimination estimates would appear to be biased under these conditions, and error is contaminated by bias, this trend may not be meaningful. In terms of test length, under a nonnormal trait distribution, the RMSEs based on LI estimation were relatively consistent across all test lengths, for a given sample size. For TESTFACT, the RMSE values were greatest for shortest test length (i.e., items=15).

Summary of Results

Item Difficulty

Both NOHARM and TESTFACT produced reasonably unbiased and stable estimates under conditions involving a normal specification of the latent trait. However, as evidenced by the

descriptive summaries of the mean and median values of the estimates and the patterns of improper parameter estimates, estimation accuracy was the poorest for easy (b=-2) and difficult (b=+2) items. In terms of the error of the estimates, FI estimation generally produced larger RMSEs under all conditions involving the normal trait specification. For LI estimation, the error of the difficulty estimates generally decreased as the sample size increased. In contrast, for FI estimation, error decreased as the test length decreased.

For nonnormal trait distributions, FI estimation generally produced superior, yet still somewhat biased difficulty estimates. In addition, for both TESTFACT and NOHARM, the accuracy of estimation of the difficulty parameter decreased as the modelled ability distribution deviated from normal. This was especially evident for items with high initial discrimination values (a=1.5). The recovery of the difficulty parameters under conditions involving nonnormal trait distributions did not, however, depend on either the sample size or the test length. In terms of the error of the estimates, which are contaminated by the bias noted above, the RMSE values were generally comparable between the two sets of estimates over most conditions.

Item Discrimination

Although somewhat dependent on the initial difficulty and discriminations values, LI estimation generally produced more accurate (unbiased), more stable, discrimination estimates when the modelled ability distribution was normal. Marginal Maximum Likelihood estimation, as implemented in TESTFACT, produced discrimination values which, when summarized via the median or mean, were consistently lower than the population values. Furthermore, bias in the TESTFACT discrimination estimates was not consistent and varied as a function of both the population difficulty and discrimination values. This suggests that there may be problems associated with the estimation of the unconstrained item-factor correlations. The FI estimator did, however, produce generally more accurate, yet still biased, parameter estimates when the assumed trait distribution was not normal. As expected, the accuracy and error of the discrimination estimates, for both LI and FI, was dependent on the population item parameters. The estimates for poorly discriminating items (a=.5) showed the least amount of bias and greatest stability. Similarly, the discrimination estimates were more accurate and less variable for items of average difficulty (b=-1,0,1). Overall, item recovery was generally less superior for items with extreme population values (e.g., a=1.5, b=-2 or 2). This effect was evident for both the FI and LI estimators. Extreme parameters (e.g., a=1.5, b=-2 or 2), combined with

distributions of the latent trait that were not normal and small sample size/test length ratios, resulted in the most biased unstable estimates. Under these conditions there will be few, if any, examinees in the ability continuum where the estimates must be derived. Collapsed over the initial item parameters, the sample size/items ratio had little effect on parameter recovery. The estimates produced by NOHARM and TESTFACT were, however, subject to less error as this ratio increased.

In terms of test length, sample size, and trait distribution, the recovery of the discrimination parameter varied by the estimation strategy employed. Under a normal specification for the latent variable the accuracy of the LI discrimination estimates did not improve with increases in sample size. There was, however, an associated decrease in the error of the estimates as the sample size increased. For MML estimation, under the normal specification, estimation accuracy tended to decrease as test length increased. Furthermore, the error of the estimates tended to increase with longer test lengths but decrease with larger sample sizes. For the nonnormal trait distribution both TESTFACT and NOHARM produced discrimination estimates that were biased and unstable. For both estimation strategies, parameter recovery worsened as the trait distribution deviated from normal. Under the conditions involving a nonnormal trait distribution it was difficult to discern the independent effects of sample size and test length. However, when the assumed ability distribution was not normal, neither test length

nor sample size appeared to have a substantial effect on estimation accuracy for either program. Although there were some conditions where the estimates from NOHARM were marginally more accurate than those produced via TESTFACT, the FI estimator was generally superior in recovering the true item discrimination parameters under the nonnormal trait specifications.

Conclusion

Overall, the LI approach generally produced at least equivalent, if not more accurate and stable difficulty and discrimination estimates, than the FI approach under conditions involving the normal specification of the latent variable. Full-information FA (MML estimation) did, however, provide marginally more accurate calibration of the 2-parameter model when the assumed trait distribution was not normal. In addition, the generation of improper difficulty and discrimination estimates was much less likely when FI estimation was employed. Although the MML estimates lost accuracy and stability under conditions involving a nonnormal trait distribution and extreme population item parameters, parameter recovery was still generally superior to that obtained via the LI approach.

CHAPTER V: DISCUSSION

McDonald (1994) recently stated that "given the notorious instability of the higher moments of a multivariate distribution, it seems unlikely that information in the data above the level of bivariate moments will commonly make a nonspurious contribution to the efficiency of estimation and to badness of fit" (p. 80). Thissen (1982) has also suggested that, although we can estimate the item parameters using full-information estimators such as MML, the use of the entire 2^p table of response counts may be problematic. From a practical perspective, accurate parameter estimates are extremely important in the test development process (see Hambleton & Jones, 1994; Hambleton, Jones, & Rogers, 1993). For IRT, or NLFA, items will be selected based on a consideration of their item information functions which, in turn, are determined by the item parameter estimates. If the parameter estimates are not precise then the final item pool may not provide for an optimal test. These issues, combined with the restricted research to date, establish the need for a more elaborate quantification of the conditions under which both limited- and full-information nonlinear FA can be used to produce traditional IRT based parameter estimates.

The results presented in this study suggest that, similar to Miller (1991), the LI FA technique incorporated in NOHARM provides a reasonable calibration of the 2PM logistic response model under conditions involving a normal distribution of the

latent trait. That is, information in the one-way (percent correct) and two-way (joint percent correct) is sufficient for the estimation of the location and slope parameters. Item recovery was, however, less accurate and somewhat unstable for extreme population item parameters. Full-information FA, while also providing an acceptable calibration of the 2PM logistic response model, did not provide equivalent results under well-fitting model conditions. Although the recovery of the difficulty, or threshold, parameter was reasonably equivalent across both the LI and FI approaches, the MML estimator incorporated in TESTFACT consistently yielded discrimination parameter estimates that underestimated the corresponding population values, especially for more highly discriminating items. An inspection of the results from Muraki and Englehard (1985) also suggested that FI solutions underestimated the factor loadings, especially for items with extreme threshold, or difficulty, values. The lack of equivalence between the LI and FI discrimination estimates under conditions involving a normal trait specification was not expected and suggests that there may be problems with the MML algorithm incorporated in TESTFACT in terms of recovering true item-factor relationships. Moreover, the lack of equivalence in parameter recovery may be partially attributable to the specific program implementations that were used (e.g., minimization algorithm, convergence criterion, number of quadrature points for FI FA). Nevertheless, given the marginal strength of the systematic underestimation of the

discrimination parameter and the equivalence of difficulty parameter estimates, it is doubtful whether test-calibration efforts based on FI FA would yield dramatically different item sets than those based on LI FA, at least under well-fitting model conditions.

It was also found that the estimates derived from FI estimation, where one attempts to extract more from the data, were comparably more accurate under conditions involving nonnormal trait specifications. However, similar to the conclusions of Collins et al. (1986), factor recovery was poor when compared to results that were obtained from experiments involving normal data sets. Furthermore, misestimation of the item parameters was magnified for extreme population values (i.e., a=1.5, b=-2,2). For FI FA, the poor parameter recovery was consistent with Zwinderman and van den Wollenberg (1990) who found that when the distribution of ability was skewed, MML estimators lose accuracy and efficiency.

Both LI and FI estimation produced spurious and unstable parameter estimates when the assumed trait distribution was not normal. Furthermore, the errors in parameter recovery increased as the trait distribution became more skewed. This result, in terms of MML estimation, was consistent with the findings from earlier research (e.g., Seong, 1990; Stone, 1992). For NOHARM II, the instability and inaccuracy of the estimates under nonnormal trait specifications was noteworthy. For conditions involving a nonnormal specification of ability the LI approach

produced estimates that were not only biased and unstable, but also prone to "infinite" values. For these specific situations it may not be appropriate to collapse the table over all items save two at a time and use only the pairwise information for estimation. More specifically, and contrary to McDonald (1982,1994), the harmonic analysis of the normal ogive model did not appear to be robust to the normality assumption. While the seriousness of the violation of the normality assumption will depend on the nature of the examinee sample and the intended application of the results, the estimates derived under the conditions modelled in this study suggest that the utility of NOHARM is suspect when the distribution of ability in the examinee sample is not normal. Although the parameter estimates for moderate difficulty and discrimination items (e.g., b=-1,0,1, a=.5, 1) were less subject to error, item recovery via NOHARM II still remained poor.

The improper estimates described in this study are neither plausible nor useful and indicate that, under certain conditions, practical applications of these FI and LI strategies may be problematic. Swaminathan and Gifford (1985) suggest that "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). From a factor analytic perspective, the existence of infinite item parameter estimates

indicates the occurrence of one or more unique variances approximately equal to zero. These are typically referred to as Heywood cases. For TESTFACT, the occurrence of some Heywood cases is not unexpected. By knowing the proportion of examinees that respond correctly to both items in any pair one can estimate the tetrachoric correlations for all distinct n(n-1)/2 pairs of items. To calculate these correlations TESTFACT uses Divgi's (1977) method. Collins et al. (1986) suggest that when item frequencies are low overall, as would be the case when the trait distribution is highly skewed, Divgi's method should never be used. Under these conditions Heywood cases are more likely to occur. As part of the TESTFACT program one can constrain the ML estimation of the slope and intercept parameters using a *beta* prior distribution on the uniquenesses and a normal prior distribution on the intercepts. Due to the unknown effect of the prior on the estimated factor loadings and a desire to maintain comparability with the NOHARM results, these options were not invoked. Nevertheless, if the parameters of the beta distribution are suitably chosen, the factor loadings will not approach one, thereby eliminating "improper" estimates of the discrimination parameters. Similarly, by specifying a normal prior with a specified mean and variance on the intercept parameter, excessively large or small difficulty estimates can be avoided.

Lord (1983) suggests that in practical work the ML estimates of the discrimination parameter sometimes tends to become

infinite. Although this suggests a positive bias in some data sets, for certain items, the results of the present investigation revealed a consistent negative bias in the discrimination parameter when estimated via MML. When the effects of outlier values were controlled, the mean of the discrimination estimates tended to be lower than the mean of the population values. This result, while unexpected, may be associated with the bias in the threshold values. When the population difficulty value was large and positive the estimated difficulty value was positively biased; when the population value was large and negative the estimated value tended to be negatively biased. Given that the data were generated using the two-parameter logistic model, there may be a number of possible a,b pairs that represent the data equally well. Therefore, there may be a number of different slope and intercept parameters that maximize the likelihood function. Finally, the bias evidenced in the discrimination and difficulty parameters was magnified by data sets that incorporated an ability distribution that was not normally distributed. The distortion in the factor loading matrix due to shape of the population distribution results in estimation problems which, in turn, may account for the bias in the parameter estimates.

The effects of sample size, test length, and the ratio of sample size to test length on parameter recovery were also investigated for various initial population item parameters. Although these factors were not important in explaining

differences in the accuracy of parameter recovery between LI and FI, they did have some effect on the error, or spread, of the estimates. In general, more stable estimates were produced as the sample size increased. This property was not unexpected in that a good estimator should do better when it is based on a large sample than when it is based on a small sample. Also, as with common linear FA, one would expect that the increased information afforded by larger samples in nonlinear FA would allow for a more precise estimation of the factor loadings and item thresholds. Thus, the estimates of the item discriminations and difficulties would also be more stable.

For some conditions, especially with sufficiently large numbers of items, it would be expected that numerical integration used in the MML approach would be heavy, thereby affecting the accuracy and stability of the parameter estimates. For a test comprised of 45 items, a total of 2^{45} (35,184,372,088,832) distinct response patterns are available. Thus, even if many of these patterns will not occur in a given sample, hundreds of thousands of distinct pieces of data must be maintained to produce fully efficient MML estimates of the item parameters. Also, in many situations, especially when trait distributions are not normally distributed, numerous combinations of these response vectors will not be present. Only when the sample size is sufficiently large will all the 2^p possible response patterns have expected values greater than one or two. Therefore, using FI FA with missing response patterns results in the collapsing of

cells, which would appear to impact on the error of the parameter estimates. While Bock et al. (1988) comment that the computations involved in MML estimation increase linearly with the number of items, Reise and Yu (1990) suggest that with MML estimation test length is not a crucial variable because a distribution is being estimated at each cycle rather than a potentially error-prone point estimate of θ for each person. For the present investigation, the accuracy of the FI difficulty and discrimination estimates was not overly affected by the test length. However, increases in test length were generally associated with increases in the RMSEs for the MML difficulty and discrimination estimates. This trend was consistent with Mislevy (1986) who suggested that the use of a greater amount of information in the estimation process should be advantageous for tests that contain fewer items. For the LI approach, which incorporates a simpler ULS fitting function, test length did not appear to be related to the goodness-of-fit of the parameter estimates. This was to be expected in that the sampling instabilities inherent with using information in the higher joint moments of the binary responses should not be a factor in deriving parameter estimates through LI FA approaches. Therefore, at least for well-fitting model conditions where parameter recovery is unbiased, the LI approach would appear to be advantageous for extended test lengths.

The differences that were found in both the accuracy and error of the two NLFA approaches can be better understood through

the use of an IRT framework. From an IRT perspective, the ICC, which can take different mathematical forms, is simply a nonlinear function for the regression of item scores on the ability or trait measured by the test. It is therefore quite reasonable that for accurate item parameter estimation one would require a heterogeneous distribution of examinees on the ability measured by the test. Without data distributed along the ability continuum it would be very difficult to estimate the ICC. Furthermore, it follows that any combination of population parameters that yields an ICC that is practically flat on the part of the range of ability where examinees are concentrated should yield less precise, unstable, parameter estimates. For the 2-parameter IRT model, item difficulty, b, is simply the value of the ability score (θ) when the slope of the item characteristic curve is at a maximum (inflexion point). In terms of ability tests, the higher the difficulty parameter, the greater the ability that is required before the examinee has a 50% probability of a correct response. Item discrimination, a, is the value of the slope of the ICC at the inflexion point. High values lead to very steep ICCs, indicating that the item discriminates well over a narrow range of abilities. A low value indicates a flat ICC which suggests that the item discriminates poorly over a wide range of abilities. It would therefore be reasonable to expect that the intercept parameter would be difficult to estimate accurately if the number of examinees at the inflection point for a particular item is small. Similarly,

if the slope of the ICC is difficult to estimate, the location along the ability scale at which an examinee has a probability of 0.5 of correctly answering the item would also be difficult to determine. Overall, when items do not match the ability of the examinees one would expect less accurate, unstable, parameter estimates. In the present investigation, especially for conditions involving a nonnormal trait distribution and extreme population parameter values, a heterogeneous distribution of scores along the ability continuum was unlikely. Therefore, estimation problems occurred. The number and pattern of improper difficulty and discrimination estimates highlights this point.

Not surprisingly, even under the nonnormal specification, increases in sample size resulted in fewer discrepant discrimination estimates. It would be expected that the specific distribution for trait would not be very crucial if the sample size was extremely large. Under such conditions the trait distribution should only have a marginal effect on the parameter estimates. As either sample size, or the ratio of sample size to the number of items increases, the homogeneity of examinee sample, in terms of ability, will likely decrease. As a result, there will be more data distributed along the continuum where the estimate is made. For the location parameter a substantial proportion of the improper estimates occurred under a nonnormal trait specification and initial conditions corresponding to a highly discriminating (a=1.5) easy items (b=-2). The ICC suggested by these initial values, combined with the distribution

of ability, also implies that there will be few, if any, examinees in the range of ability with which to estimate the location parameter. Under the nonnormal specification of the latent variable, LI estimation resulted in more improper difficulty estimates for larger sample sizes. While this result is unexpected, it suggests that, for the LI estimator, the additional heterogeneity of the examinee sample does little to compensate for the estimation difficulties associated with a nonnormal trait distribution. For FI, this trend was not evident. In fact, even under the skewed ability distribution, there were very few discrepant difficulty estimates when the sample size reached 10,000. In terms of test length, holding other factors constant, both LI and FI produced comparably fewer improper difficulty estimates when the number of items was small. This result, combined with trends noted above, suggests that, when the ability distribution is not normal, the additional information utilized in the FI estimation process may yield superior parameter estimates, provided that the sample size is sufficiently large.

As stated above, tests that are very easy or hard for the calibration sample of examinees can result in biased and unstable item parameter estimates. However, for TESTFACT, Bayesian priors can be placed on any item parameters that are difficult to measure. Hambleton (1993) suggests that the feasibility of Bayesian procedures should be investigated more fully. While the incorporation of prior information in the estimation procedure

will likely provide improved parameter estimates, especially for short tests and small sample sizes, the use of an inappropriate specification may detract from estimation. Seong (1990) suggested that the user should specify the prior ability distribution based on theoretical or empirical considerations. Unfortunately, the true population distribution of ability is often impossible to determine. In these situations, the default normal prior ability distribution is recommended. At present, the LI estimation incorporated in NOHARM does not provide a means for placing constraints on item parameters or, if possible, utilizing prior information concerning the distribution of the latent variable in the population of persons responding to the items. The relatively large number of improper estimates, especially for conditions involving nonnormal ability distributions, suggests that the ability to place place constraints on the item parameter estimates may facilitate the estimation process. It should be noted, however, that setting such constraints may mask serious problems with the data.

Unlike much of the previous research concerning parameter recovery from various estimation strategies, the accuracy and error of the estimates at different levels of the population parameters was investigated here. For extreme levels of population difficulty, one would expect that the parameter estimates would generally be less accurate and less stable. These estimates would be based on the information (responses) provided by a relatively small number of examinees. It would

also be anticipated that, provided the distribution of the latent trait was normal, conditions involving extreme population difficulty values and high population discrimination values would also provide unstable estimates for both LI and FI estimators. Under these conditions, the range of ability where the examinees are concentrated may not correspond with the area where the ICC has an inflection point. Therefore, due to potential range restrictions involved in estimating item-factor correlations, it would be difficult to determine where and how sharply the curve turns. From an IRT perspective we are attempting to describe the proportion of correct responses in a certain range of ability. When this range is small (e.g., the items are too difficult or easy for a given group), there will be a variety of discrepant (a,b) pairs that capture the data equally well. Therefore, one would expect unstable estimates. The incorporation of a nonnormal trait distribution, especially one that is positively skewed, compounds this problem.

CHAPTER VI: CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE PRACTICE

It remains that "the issues and technology associated with item response theory are neither fully developed nor without controversy (Hambleton, 1993, p. 149). In a review article covering major developments in factor analysis Mulaik (1986) suggested that further developments in terms of generalizing particular FA models will be an important ongoing research activity. Furthermore, given the equivalence between NLFA and the more traditional IRT model specifications, applications of NLFA to existing measurement problems such as test dimensionality, model-data fit, local item dependence and item bias need further exploration. The results of the present study add to the synthesis of common FA and latent trait models by providing empirical evidence to suggest that test calibrations can be accomplished using nonlinear FA. That is, for some conditions, NLFA is appropriate for the analysis of binary data. Furthermore, for well-fitting model conditions, little information is lost by not using higher-order relationships among the items. Full-information FA, which is based on the estimation of item response vectors, did provide a marginally better calibration of the 2PM model under conditions involving a nonnormal trait distribution. Nevertheless, given the general instability of the FI estimator under the ill-fitting conditions

modelled in this study, the use of all of the information in the 2^p item vectors may still be insufficient for practical applications, at least for small sample sizes and tests that consist of items with extreme difficulties.

The results of this study raise a number of concerns regarding the use of NLFA methods and estimation strategies. In addition, the restricted scope of this investigation suggests that there are a number of additional areas where studies of the use of NLFA for item analysis are warranted.

First, the comparison of two nonlinear FA models was based on their estimation of the threshold and slope values for the two-parameter normal ogive model. The lower asymptote parameter was set at zero and not estimated. In many testing situations it would be unrealistic to assume that guessing is not a relevant factor. For example, one would expect some low ability examinees to correctly answer some test items due to guessing. Nevertheless, the present investigation does provide valuable information regarding the relative utility of FI and LI estimation methods in recovering item parameters under the 2-parameter normal ogive model. One would expect that the choice of another model may have some effect on item recovery under the conditions investigated. For example, the 3-parameter model, which has three unconstrained parameters per item, may require proportionally larger numbers of examinees to calibrate successfully. Unfortunately, while both TESTFACT and NOHARM

provide a mechanism for inputting fixed guessing values, there are currently no procedures available to estimate 3PM models.

Second, although the simulated conditions (number of items, number of examinees, item parameter values, ability distributions) were selected to represent a realistic set of population values, they do not cover all combinations that would be of interest to educational testing specialists. They do, however, provide a substantive array of conditions in which to compare parameter recovery for LI and FI methods. Furthermore, given the extremely heavy computational resources required by TESTFACT, the existence of simulated conditions such as those involving 60 items and 10,000 examinees provides valuable information on the practical utility of LI and FI methods. Nevertheless, there may be certain population conditions that need further exploration. For example, although two nonormal trait distributions were modelled in this study, it would be informative to delimit more specifically the conditions (i.e., degree of skewness of the ability distribution) under which NOHARM produces discrepant parameter estimates.

Third, although unidimensional, nonlinear, dichotomous response models are still mainly used by researchers who wish to analyse a set of test items and estimate examinee ability (De Champlain, 1992; Hambleton, 1993), there is a growing concern regarding tenability that one latent trait is sufficient to explain examinee performance and the interrelationships among test items. Moreover, fitting unidimensional models to

multidimensional data is rarely appropriate. As a result, numerous research efforts are now being directed at formulating and testing the utility of multidimensional IRT models (e.g., Ackerman, 1994; Reckase & McKinley, 1991). The present investigation, while still based fundamentally on the arguably unrealistic assumption of unidimensionality, provides valuable data on the relative strengths of competing estimation strategies. As both McDonald (1994) and Muthen (1978) suggested, very little information is lost by not using higher-order relationships among the items, provided that the model assumptions are not violated (e.g., for a unidimensional model the underlying trait is normally distributed). However, additional studies are needed in order to ascertain whether this premise is sustainable when the data are multidimensional in nature. It would be expected, however, that the lack of robustness of the parameter estimates would extend to multidimensional item response theory (MIRT) models. From a MIRT perspective, the Item Characteristic Surface (ICS), which can take on different mathematical forms, is simply a nonlinear function for the regression of item scores on the traits measured by the test. It is therefore quite reasonable that for appropriate item parameter estimation one would require a heterogeneous distribution of examinees on the abilities measured by the test. This situation would be less likely to occur when the underlying latent response variables are not normally distributed. Fortunately, both the LI and FI methods used in

this study allow for the incorporation of models that are not unidimensional. Therefore, parameter estimation comparisons between LI and FI could be performed for MIRT models.

Fourth, many indices and statistics that are used to test dimensionality are based on LI and FI models. Both TESTFACT and NOHARM provide, either directly or indirectly, information that can be used to assess departure from the assumption of unidimensionality. For NOHARM, the analysis of the residual covariance matrix after fitting a nonlinear 1-factor model can be used. Unidimensionality of the latent trait would theoretically imply zero residual covariances among all pairs of items at fixed ability levels. Indices based on this premise have recently been shown to be quite promising (De Champlain, 1992; Gessaroli, 1995). For TESTFACT, a chi-square approximation for the likelihood ratio test of fit of the model relative to the general multinomial alternative is used. In terms of assessing dimensionality using indices based on nonlinear FA models, it would be important to ascertain the comparative accuracy of the estimates from LI and FI estimators. It would be expected that the validity of tests of dimensionality that are based on either the weak or strong principles of local independence would depend, in part, on the degree to which FI and LI methods adequately calibrate the test items.

Finally, only two NLFA strategies were investigated in this study. Limited-information estimates can also be derived via the GLS estimator available in the software package LISCOMP (Muthén,

1987). However, unlike the ULS estimator, GLS is computationally very demanding and utilizes not only terms from the one-way and two-way margins but also from the three-way and four-way margins. Muthen (1989) suggests that GLS estimation probably should not be attempted when the number of items exceeds 30. Muraki and Englehard (1985) also advise that this approach to item FA is of little practical use for analyzing achievement and cognitive tests since these tests consist of a sizable number of items. Finally, in order to estimate the GLS weight matrix properly with many items, numerous subjects may be required. The asymptotically distribution free (ADF) estimator proposed by Browne (1984) can also be used in the factor analysis of dichotomous variables. Unfortunately, this solution is also computationally burdensome. Nevertheless, the rapid growth in the availability of powerful computing facilities suggests that the cost of test calibration using GLS or WLS(ADF) may not be that prohibitive. Therefore, where adequate conditions exist (i.e., adequate sample sizes, restricted test lengths, small item to factor ratios), contrasts of NLFA parameter estimates using these alternative estimation strategies should be undertaken.

The establishment of nonlinear FA as an appropriate methodology for analyzing latent trait models will be of benefit to many researchers and practitioners. Nonlinear FA is a more general model and, provided that a suitably large and appropriate sample of examinees is available to facilitate parameter estimation, allows for analysis of response data under various

conditions typically encountered in educational testing. However, the properties of the item parameter estimation techniques are intertwined with the computer program used to estimate them. Therefore, additional theoretical and empirical studies are needed in order to delimit the specific conditions where NLFA, and the associated estimation strategies, can be best used to solve practical measurement problems.

REFERENCES

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

Ackerman, T. A. (1985). M2PLGEN: A computer program for generating thetas and response strings corresponding to the M2PL model. Iowa City, Iowa: The American College Testing Program.

Baker, F. B. (1987). Methodological review: item parameter estimation under the one-, two-, and three-parameter logistic models. Applied Psychological Measurement, 11, 111-141.

BMDP Statistical Software Manual (1990). University of California Press: Los Angeles.

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika, 46, 443-459.

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

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

Boulet, J. & Gessaroli, M. (1992). A monte carlo comparison of the accuracy of parameter estimates using full-information and bivariate-information models. Paper presented at the annual meeting of the Canadian Educational Researchers' Association, Charlottetown, P.E.I. Canada.

Browne, M. W. (1984). Asymptotic distribution free methods in the analysis of covariance structure. British Journal of Mathematical and Statistical Psychology, 37, 62-83.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (second edition). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. Psychological Bulletin, 1, 155-159.

Collins, L. M., Cliff, N., McCormick, D. J., & Zatkin, J. L. (1986). Factor recovery in binary data sets: a simulation. Multivariate Behavioral Research, 21, 377-391.

De Champlain, A. (1995). An overview of nonlinear factor analysis and its relationship to item response theory. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Divgi (1979). Calculation of the tetrachoric correlation coefficient. Psychometrika, 44, 169-172.

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

Etezadi-Amoli, J., & McDonald, R. P. (1983). A second generation of nonlinear factor analysis. Psychometrika, 48, 315-342.

Everitt, B. S. (1987). Introduction to optimization methods and their application in statistics. New York: Chapman and Hall.

Fang, T. & Gessaroli, M. (1995). Assessing local item dependence using nonlinear factor analysis. Paper presented in A.F. De Champlain (chair), The utility of nonlinear factor analysis in addressing common measurement issues: Applications to a national testing program. Symposium conducted at the annual meeting of the American Educational Research Association.

Fraser, C. (1983). Noharm: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. Armidale: University of New England, Centre for Behavioural Studies.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. Multivariate Behavioral Research, 23, 267-269.

Gessaroli, M. & De Champlain, A. (in press). Using an approximate Chi-square statistic to test the number of dimensions underlying the responses to a set of items. Journal of Educational Measurement.

Hambleton, R. K. (1993). Principles and selected applications of item response theory. In R.L. Linn (ed.). Educational measurement: third edition. Phoenix, AZ: National Council on Measurement in Education, Oryx Press.

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. Applied Measurement in Education, 7, 171-186.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. Journal of Educational Measurement, 30, 143-155.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: principles and applications. Boston, MA: Kluwer-Nyjhoff.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). Fundamentals of item response theory. Sage: Newbury Park, CA.

Hsu, T-C, & Yu, L. (1989). Using computers to analyse item response data. Educational Measurement: Issues and Practice, Fall, 21-28.

Kirk, D. B. (1973). On the numerical approximation of the bivariate normal(tetrachoric) correlation coefficient. Psychometrika, 38, 259-268.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparisons between factor analysis and multidimensional item response models. Multivariate Behavioral Research, 26, 457-477.

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

Lord, F. M. (1953). The relationship of the test score to the trait underlying the test. Educational and Psychological Measurement, 13, 517-548.

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

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

McDonald, R. P. (1962). A general approach to nonlinear factor analysis. Psychometrika, 27, 397-415.

McDonald, R. P. (1965). Difficulty factors and non-linear factor analysis. The British Journal of Mathematical and Statistical Psychology, 18, 11-23.

McDonald, R. P. (1967a). Nonlinear factor analysis. Psychometric Monograph, No. 15. (a).

McDonald, R. P. (1967b). Factor interaction in nonlinear factor analysis. The British Journal of Mathematical and Statistical Psychology, 20, 205-215.

McDonald, R. P. (1979). The simultaneous estimation of factor loadings and score. British Journal of Mathematical and Statistical Psychology, 32, 212-228.

McDonald, R. P. (1980). A simple comprehensive model for the analysis of covariance structures: some remarks on applications. The British Journal of Mathematical and Statistical Psychology, 33, 161-183.

McDonald, R. P. (1981). The dimensionality of tests and items. The British Journal of Mathematical and Statistical Psychology, 34, 100-117

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. Applied Psychological Measurement, 6, 379-396.

McDonald, R. P. (1986). Describing the elephant: structure and function in multivariate data. Psychometrika, 51, 513-534.

McDonald, R. P. (1989). Future directions for item response theory. International Journal of Educational Research, 13, 205-220.

McDonald, R. P. (1994). Testing for approximate dimensionality. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.). Modern theories in measurement: problems and issues, pp. 63-86, Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. The British Journal of Mathematical and Statistical Psychology, 27, 82-99.

Miller, T. R. (1991). Empirical estimation of standard errors of compensatory MIRT model parameters obtained from the NOHARM estimation program. ACT Research Report Series, ONR91-2, August.

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

Mislevy, R. J., & Bock, R. D. (1984). BILOG: Maximum likelihood item analysis and test scoring with logistic models. Mooresville, IN: Scientific Software.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). Introduction to the theory of statistics (3rd edition), New York: McGraw Hill.

Mulaik, S. A. (1986). Factor analysis and Psychometrika: major developments. Psychometrika, 51, 23-33.

Muraki, E., & Englehard, G. (1985). Full information item factor analysis: applications of EAP scores. Applied Psychological Measurement, 9, 417-430.

Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

Muthén, B. O. (1983). Latent variable structural equation modelling with categorical data. Journal of Econometrics, 22, 43-65.

Muthén, B. O. (1984). A general structural equation model with dichotomous ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.

Muthén, B. O. (1985). LISCOMP. Mooresville, IN: Scientific Software.

Muthén, B. O. (1989). Dichotomous factor analysis of symptom data. Sociological Methods & Research, 18, 19-65.

Muthén, B.O., & Leehman, J. (1985). Multiple group IRT modelling: applications to item bias analysis. Journal of Educational Statistics, 10, 133-142.

Olsson, U. (1979). On the robustness of factor analysis against crude classifications of the observations. Multivariate Behavioral Research, 14, 485-500.

Parry, C. D. H., & McArdle, J. J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. Applied Psychological Measurement, 15, 35-46.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. Psychological Bulletin, 1, 160-164.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. Applied Psychological Measurement, 15, 361-373.

Reise, S. P., & Yu, J. (1990). parameter recovery in the graded response model using MULTILOG. Journal of Educational Measurement, 27, 133-144.

Rigdon, E. E., & Ferguson, C. E. jr. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. Journal or Marketing Research, XXVIII, 491-497.

Rubin, D. B. (1991). EM and beyond. Psychometrika, 56, 241-254.

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

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

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation of 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. (1986). MULTILOG [Computer program]. Mooresville, IN: Scientific Software.

Wilson, D., Wood, R., & Gibbons, R. (1991). TESTFACT: Test scoring, item statistics, and item factor analysis [computer program]. Scientific Software, Inc.: Mooresville, IN.

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

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

Zwinderman, A. H., & van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. Applied Psychological Measurement, 14, 73-81.