DOCUMENT RESUME

ED 469 371 TM 034 474

AUTHOR	De Champlain, Andre F.
TITLE	An Overview of Nonlinear Factor Analysis and Its Relationship to Item Response Theory. Statistical Report. LSAC Research Report Series.
REPORT NO	LSAC-R-95-03
PUB DATE	1999-03-00
NOTE	23p.
PUB TYPE	Information Analyses (070) -- Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; Estimation (Mathematics); *Factor Analysis; *Item Response Theory; Law Schools; Models
IDENTIFIERS	Item Parameters; Law School Admission Test; *Nonlinear Models

ABSTRACT

This paper provides a brief overview of some of the research that has examined the relationship between item response theory (IRT) and nonlinear factor analysis (NFA), and outlines three NLFA models, emphasizing their major strengths and weaknesses for practical applications. The focus is on: (1) R. McDonald's polynomial approximation of a normal ogive model (1967, 1982); (2) A. Christoffersson's (1975) and B. Muthen's (1984) factor analytic model for dichotomous variables; and (3) the full information factor analytic model (R. Bock and M. Aitken, 1981; R. Bock, R. Gibbons, and E. Muraki, 1988). Items from two Law School Admission Test (LSAT) forms were calibrated using these three models to assess the degree of comparability of the IRT parameter estimates using these procedures. Findings show that the parameter estimates tended to be quite similar, regardless of the calibration procedure. Additional research should be conducted with a larger number of LSAT forms before reaching any conclusions about the comparability of the calibration procedures. (Contains 2 tables, 2 figures, and 55 references.) (Author/SLD)

TM

An Overview of Nonlinear Factor Analysis and Its Relationship to Item Response Theory

André F. De Champlain
Law School Admission Council

Law School Admission Council
Statistical Report 95-03
March 1999

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

LSAT®; *The Official LSAT PrepTest®*; *LSAT: The Official TriplePrep®*; and the Law Services logo are registered marks of the Law School Admission Council, Inc. Law School forum is a service mark of the Law School Admission Council, Inc. *LSAT: The Official TriplePrep Plus*; *The Whole Law School Package*; *The Official Guide to U.S. Law Schools*, and ***LSACD*** are trademarks of the Law School Admission Council, Inc.

Law School Admission Council fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, and other matters may change without notice at any time. To remain up-to-date on Law School Admission Council policies and procedures, you may obtain a current *LSAT/LSDAS Registration and Information Book*, or you may contact our candidate service representatives.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or ¨cy of the Law School Admission Council.

Table of Contents

Executive Summary

Item response theory (IRT) models have been used extensively to address educational measurement and psychometric concerns pertaining to a host of areas such as differential item functioning, equating, and computer-adaptive testing due to their many advantages, such as item and ability parameter invariance across test-taker subgroups and item pools, respectively. Another approach, which is gaining popularity in educational measurement, is the one that treats IRT as a special case of nonlinear factor analysis (NLFA). Several authors have shown that these models are mathematically equivalent (Goldstein & Wood, 1989; Knol & Berger, 1991; McDonald, 1967, 1989, 1994). It would therefore appear reasonable to make use of NLFA models to examine a multitude of educational measurement problems that had been, until quite recently, looked at solely from an IRT perspective.

The purpose of this paper is to provide a brief overview of some of the research that has examined the relationship between IRT and NLFA and to outline three NLFA models, emphasizing their major strengths and weaknesses for practical applications. More precisely, McDonald's (1967, 1982b) polynomial approximation to a normal ogive model, Christoffersson's (1975)/Muthén's (1984) factor analytic model for dichotomous variables, as well as Bock and Aitkin's (1981)/Bock, Gibbons, and Muraki's (1988) full-information factor analytic model, will be summarized. Also, the items from two LSAT forms will be calibrated using these three models in order to assess the degree of comparability of the IRT parameter estimates using these procedures.

Introduction

Over the past three decades, the educational measurement and psychometric literature have been replete with studies focusing on item response theory (IRT) models. The numerous textbooks that have been written centering primarily, and in some instances exclusively, on IRT attest to the importance of these models in the development and analysis of tests and items (Baker, 1992; Hambleton, 1983, 1989; Hambleton & Swaminathan, 1985; Hulin, Drasgow, & Parsons, 1983; Warm, 1978). The use of IRT models has been widespread in both testing organizations and departments of education for a variety of purposes such as item analysis (Baker, 1985; Mislevy & Bock, 1990; Thissen, 1993; Wingersky, Patrick, & Lord, 1991), score equating (Cook & Eignor, 1983; Lord, 1977, 1980, 1982; Petersen, Kolen, & Hoover, 1989; Skaggs & Lissitz, 1986), differential item functioning (Thissen, Steinberg, & Wainer, 1993), and computer adaptive testing (Hambleton, Zaal, & Pieters, 1993; Kingsbury & Zara, 1991; Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990). The many properties of IRT models, among them that "sample-free" item parameter estimates and "test-free" ability estimates can be obtained, have generated considerable interest in their use to solve a host of measurement-related problems.

Another approach, which is currently gaining popularity in educational measurement is the one that treats IRT as a special case of nonlinear factor analysis (NLFA). Several authors have shown that these models are mathematically equivalent (Balassiano & Ackerman, 1995a, 1995b; Goldstein & Wood, 1989; Knol & Berger, 1991; McDonald, 1967, 1989, 1994).

A considerable body of research has been dedicated to examining the relationship between common IRT models, e.g., logistic and normal ogive functions, and NLFA (Bartholomew, 1983; Goldstein & Wood, 1989; Knol & Berger, 1991; McDonald, 1967, 1989; Takane & De Leeuw, 1987). Muthén (1978, 1983, 1984) has demonstrated that commonly used models in IRT, for example the two-parameter normal ogive model, are specific cases of a more general factor analytic model for categorical variables with multiple indicators (i.e., response categories). McDonald (1982b), starting from Spearman's common factor model, also shows that IRT models are a special case of NLFA and provides a general framework which includes unidimensional/multidimensional, linear/nonlinear as well as dichotomous and polychotomous models.

Bartholomew (1983) proposed a general latent trait model on which several IRT as well as factor analytic functions for dichotomous variables are founded. The author states that common factor analytic models, such as those proposed by Christoffersson (1975) and Muthén (1978), are special cases of this general latent trait model. The model is of the form

$$G(\pi_i(y)) = \beta_i + \sum_{j=1}^{q} a_{ij} H(y_j), \; i = 1,2 \ldots, p. \qquad (1)$$

Bartholomew states that the models outlined by Christoffersson (1975) and Muthén (1978) use the probit function, $\{G(u) = \Phi^{-1}(u)\}$ for both G and H. Lord and Novick (1968), whose discussion on IRT models is restricted to the $q = 1$ (i.e., unidimensional) case, treat y_js as parameters and use the logit for G and the probit for H. Within the unidimensional IRT framework, the terms in Equation 1 would correspond to the following:

$G(\pi_i)$ = the response function indicating the probability of obtaining a correct response to item i;
y = a vector of ability (in this case, a scalar, given that $q = 1$);
β_i = a parameter related to the difficulty of item i;
α_{ij} = a parameter related to the discrimination of item i on latent trait j; and
$H(y_j)$ = the density function for a given latent trait j.

Takane and De Leeuw (1987) also established that IRT models are mathematically equivalent to NLFA. These authors provided a systematic series of proofs that show the equivalence of these models with dichotomous as well as polychotomous item responses. Finally, Knol and Berger (1991) compared Bock and Aitkin's (1981) full-information factor analysis (FIFA) model and McDonald's (1967) polynomial approximation to a normal ogive model to the two-parameter logistic IRT function and showed that they were equivalent.

Thus, it appears as though IRT and NLFA models represent two equivalent formulations of a more general latent trait model. Given the equivalence of IRT and NLFA, it would seem reasonable to make use of the latter models to examine a multitude of educational measurement problems that had been, until quite recently, looked at solely from an IRT perspective. In fact, several NLFA models, with potential applications to measurement and psychometric issues, have been proposed in the literature (Bock & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988; Bock & Lieberman, 1970; Christoffersson, 1975; McDonald, 1967, 1982b; Muthén, 1978, 1984).

Three NLFA models that have been used to address measurement issues will be presented in this paper. McDonald's (1967, 1982b) polynomial approximation to a normal ogive model, Christoffersson's (1975)/Muthén's (1978) factor analytic model for dichotomous variables, as well as Bock and Aitkin's (1981)/Bock, Gibbons, and Muraki's (1988) FIFA model, will be summarized. Also, the relationship that exists between these parameterizations and the normal ogive model will be emphasized. Finally, some of the strengths and weaknesses of the models for practical applications will be underscored.

The Normal Ogive Model

A common IRT model that outlines the probability that a randomly selected test taker of ability θ_j will correctly answer item i is the three-parameter normal ogive model. The item response function (IRF) for the model is given by

$$P_i(\theta_j) = c_i + (1 - c_i) \int_{-\infty}^{Z_{ij}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt, \qquad (2)$$

where c_i is the lower asymptote or "pseudo-guessing" parameter indicating the lowest probability of obtaining a correct response to item *i* and *t* is the variable of integration. Several parameterizations of Z_{ij} have been proposed in the literature (Bock & Aitkin, 1981; Christoffersson, 1975; Lord, 1952; McDonald, 1967). Four common formulations are presented in the next sections of the paper.

Lord's Parameterization

Lord (1952) proposed the following parameterization of Z_{ij} for item *i*,

$$Z_{ij} = \alpha_i(\theta_j - \beta_i), \tag{3}$$

where

- θ_j = the latent parameter estimate or ability value for test taker *j*;
- β_i = the item difficulty parameter estimate for item *i* or the θ value at the point of inflexion of the item response function (IRF); and
- α_i = the item discrimination parameter estimate for item *i* related to the slope of the IRF at its point of inflexion.

Based on this model, the probability of obtaining a correct response to item *i* is given by

$$P(Y_i = 1 \mid \theta_j) = c_i + (1 - c_i)N\,[\alpha_i(\theta_j - \beta_i)], \tag{4}$$

where, *a*, *b*, and *c* have been defined in Equation 3 and *N* corresponds to the normal ogive function given in Equation 2. Lord and Novick (1968) also proposed a logistic approximation to the normal ogive model which is computationally simpler and nearly identical given the relationship between the two functions,

$$L\,(Z_{ij}) = c_i + \frac{1 - c_i}{1 + e^{-z_{ij}}}, \tag{5}$$

where $Z_{ij} = \alpha_i(\theta_j - \beta_i)$ has been defined in Equation 3.

These logistic functions have been implemented in several computer programs, among them, BILOG (Mislevy & Bock, 1990) and LOGIST (Wingersky, Patrick, & Lord, 1991), that respectively utilize marginal and joint maximum likelihood estimation procedures to derive IRT item and ability parameter values. Currently, BILOG is used by the Law School Admission Council (LSAC) to calibrate Law School Admission Test (LSAT) items.

Bock and Aitkin's Parameterization

Bock and Aitkin's (1981) parameterization of Z_{ij}, based on an *m*-factor model is given by

$$Z_{ij} = \lambda_{i0} + \lambda_{i1}\theta_{1j} + \lambda_{i2}\theta_{2j}, \ldots, + \lambda_{im}\theta_{mj}. \tag{6}$$

That is, the unobservable response process for person *j* to item *i* is a linear function of *m* normally distributed latent variables $\theta_j = [\theta_{1j}, \theta_{2j}, \ldots, \theta_{mj}]$ and factor loadings $\lambda_i = [\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{im}]$. The latent response process, y_{ij} is related to the binary (observed) item response x_{ij} through a threshold parameter γ_i for item i, in the following fashion:

$$\text{if } y_{ij} \geq \gamma_i, \text{ then } x_{ij} = 1,$$
$$\text{if } y_{ij} < \gamma_i, \text{ then } x_{ij} = 0.$$

The probability that test taker j with abilities $\theta_j = [\theta_{1j}, \theta_{2j}, \ldots, \theta_{mj}]$ will correctly answer item i is given by the function

$$P(x_{ij} = 1 \mid \theta_j) = N\left(\left(\gamma_i - \sum_{k=1}^{m} \lambda_{ik}\, \theta_{kj}\right) / \sigma_i\right), \tag{7}$$

where N corresponds to the normal ogive function and σ_i is the standard deviation of the unobserved random variable $\varepsilon_{ij} \sim N(0,\sigma^2_i)$ in the common factor model

$$Y_{ij} = \lambda_{i0} + \lambda_{i1}\theta_{1j} + \lambda_{i2}\theta_{2j}, \ldots, + \lambda_{im}\,\theta_{mj} + \varepsilon_{ij}. \tag{8}$$

Bock and Aitkin (1981) proposed a marginal maximum likelihood (MML) procedure to estimate the parameters in the model based on Dempster, Laird, and Rubin's (1977) EM algorithm. The threshold and factor loadings are estimated so as to maximize the following multinomial probability function,

$$L_m = P(X) = \frac{N!}{r_1!\, r_2! \ldots r_s!} \tilde{P}_1^{\,r_1} \tilde{P}_2^{\,r_2} \ldots \tilde{P}_s^{\,r_s}, \tag{9}$$

where, r_s is the frequency of response pattern s and $\tilde{P}_s$ is the marginal probability of the response pattern based on the item parameter estimates. The function outlined in Equation 7, with the MML parameters estimated by means of the EM algorithm, is commonly referred to as *full-information item factor analysis* (Bock, Gibbons, & Muraki, 1988) and has been implemented in the computer program TESTFACT (Wilson, Wood, & Gibbons, 1987).

Advantages and Limitations of Bock and Aitkin's / Bock, Gibbons, and Muraki's Full-information Item Factor Analysis

The main advantage of FIFA is that it utilizes all available information in the estimation procedure. Contrary to most factor analytic models, which are restricted to lower-order marginals, FIFA is based on the estimation of item response vectors and hence uses all available information in the data.

Also, the procedure is implemented in the computer program TESTFACT (Wilson, Wood, & Gibbons, 1987). The output from a TESTFACT analysis contains, among other things, classical item statistics and factor analytic parameter estimates as well as their associated standard errors. In addition, a likelihood-ratio chi-square test is provided to help the user determine the fit of a model, or of alternative models. However, the use of all information contained in the 2^p item vectors by FIFA, where p is equal to the number of items, requires that there be no empty cells, which is usually not feasible unless some collapsing is done. In other words, the full information is never utilized when estimating parameters in practical testing situations. McDonald (1989) underscored this limitation of full-information methods when he stated:

> It is not impossible that to obtain acceptably precise estimates of the more flexible item response functions we would require a sample larger than the population for which the test is intended, at least for countries with smaller populations than that of China. (p.213)

In addition, as Mislevy (1986) and Berger and Knol (1990) have noted, the G^2 goodness-of-fit statistic computed by TESTFACT tends to be unreliable with data sets containing more than 10 items due to the small expected number of test takers per cell. More precisely, Mislevy (1986) states that the approximation to the chi-square distribution might be poor in these instances. Wilson, Wood, and Gibbons (1987) also caution against relying on the G^2 fit statistic when a large number of cells have expected frequencies near zero. As an alternative, the authors, based on work undertaken by Haberman (1977), recommend using the G^2 difference test to compare two competing models given that the statistic follows a chi-square distribution in large samples, even in the presence of a sparse frequency table. Also, the FIFA model, as currently implemented in TESTFACT, constrains the user to fit exploratory orthogonal solutions to item response data which may not adequately reflect some testing situations.

Christoffersson's (1975)/Muthén's (1978) Parameterization

Christoffersson's (1975) parameterization of Z_{ij}, is given by

$$Z_{ij} = (\gamma_i + \lambda_i \theta_j)/(1 - \lambda_i^2)^{1/2}, \tag{10}$$

where γ_i is a threshold parameter, λ_i is the factor loading on item i and θ_j corresponds to the ability parameter value previously defined in Equation 3.

Christoffersson (1975) proposed a factor analytic model for dichotomous variables in which it is also postulated that response variables X_i are accounted for by the latent continuous variables Y_i and threshold variables γ_i such that,

$$X_i = 1, \text{ if } Y_i > \gamma_i$$

$$X_i = 0, \text{ otherwise,}$$

where

$$Y = \Lambda\theta + E, \tag{11}$$

$\mathbf{Y} = (Y_i, \ldots, Y_n)'$, $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$ and $\Lambda = (\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{im})$. The model outlined in Equation 11 is identical to the common factor model with the exception that $\mathbf{Y}$ is unobserved. Assuming that $\theta \sim MVN(0,\mathbf{I})$, $\mathbf{E} \sim MVN(0,\Psi^2)$, that is multivariate normal, where Ψ^2 is a diagonal matrix of residual covariances, and $\text{cov}(\theta,\mathbf{E}) = 0$, the covariance matrix Σ among the $\mathbf{Y}$ latent variables can be expressed as

$$\sum = \Lambda\Phi\Lambda' + \psi^2. \tag{12}$$

where
- Λ = a matrix of factor loadings $[\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{im}]$;
- Φ = a matrix of factor correlations; and
- Ψ^2 = a matrix of residual covariances.

Also, assume that,

$$Y \sim MVN(0, \Lambda\Phi\Lambda' + \Psi^2). \tag{13}$$

The probability of a correct response based on Christoffersson's model is given by

$$P(y_i = 1) = \int_{\gamma_i}^{\infty} \frac{1}{(2\pi)^{1/2}} e^{-x^2/2}\, dx = H(\gamma_i). \tag{14}$$

The probability of correctly answering a pair of items is given by

$$P(y_i = 1, y_j = 1) = \int_{\gamma_i}^{\infty}\int_{\gamma_j}^{\infty} \frac{1}{2\pi\,|\Sigma_{ij}|^{1/2}} e^{-x'\Sigma_{ij}^{-1}x/2} dx. \tag{15}$$

Christoffersson (1975), using the tetrachoric expansion (Kendall, 1941) re-expresses Equation 15 as

$$P(y_i = 1, y_j = 1) = \sum_{s=0}^{\infty} \sigma_{ij}{}^{s}\, \tau_s(\gamma_i)\, \tau_s(\gamma_j), \tag{16}$$

where τ_s is the s-th tetrachoric function defined by

$$\tau_s(x) = H_{s-1}(x) f(x) / (s!)^{1/2}, \tag{17}$$

H_s is the s-th Hermite polynomial given by

$$H_s(x) f(x) = \left(-\frac{d}{dx}\right)^{s} f(x) \tag{18}$$

and $\sigma_{ij}{}^{s}$ is the i,j-th element of the latent covariance matrix Σ. Given the rapid convergence of the series, Christoffersson (1975) states that, in most applications, the expansion can be cut after 10 terms.

The parameters of Christoffersson's (1975) model are estimated using a generalized least-squares (GLS) estimation procedure that minimizes the fit function,

$$F = (p - P)' S_e^{-1} (p - P), \tag{19}$$

where

- **S_e** = a consistent estimator of Σ_e, the residual covariance matrix;
- **P** = a vector of expected proportions of test takers correctly answering P_j items and jointly answering P_{jk} items; and
- **p** = a vector of observed proportions of test takers correctly answering p_j items and jointly answering p_{jk} items.

Muthén (1978, 1983, 1984, 1988) proposed a GLS estimator that is equivalent to that outlined by Christoffersson (1975) but is computationally more efficient. According to Muthén (1978), the parameters of the factor analytic model for dichotomous variables can be estimated by minimizing the weighted least-squares fit function,

$$F = \frac{1}{2}(s - \sigma)' W_{\delta}^{-1}(s - \sigma), \tag{20}$$

where

- σ = population threshold and tetrachoric correlation values;
- s = sample estimates of the threshold and tetrachoric correlation values; and
- W_{δ} = a consistent estimator of the asymptotic covariance matrix of s, multiplied by the total sample size.

This approach, also referred to as GLS estimation using a full-weight matrix approach (Muthén, 1988), is asymptotically equivalent to Christoffersson's solution and slightly less demanding in terms of computational requirements. It is referred to as a full-weight matrix approach because, as Muthén (1988) states, the GLS estimator uses a weight matrix of size $p^* \times p^*$, where p^* corresponds to the total number of elements in the s vector.

Advantages and Limitations of Christoffersson's / Muthén's Factor Analytic Model for Dichotomous Variables

Muthén's solution is incorporated in the computer program LISCOMP (Muthén, 1988). LISCOMP, unlike the current version of TESTFACT, enables the user to fit both exploratory and confirmatory unidimensional or multidimensional models. As was the case with FIFA, statistical tests of model fit are readily available. The *F* functions minimized in the GLS solution (cf. Equations 19 and 20) asymptotically follow a chi-square distribution, with $df = k(k-1)/2 - t$, where *k* is equal to the number of items and *t* the number of parameters estimated in the model. Standard errors for the parameters estimated in the model can also be obtained quite easily. However, the GLS estimation procedure, unlike FIFA, utilizes terms solely from the one-way, two-way, three-way, and four-way margins, that is, the proportions of test takers correctly answering one to four items taken at the same time. In other words, the GLS estimator ignores higher-level interactions in the data and in that sense does not fully utilize all of the available information. Nonetheless, McDonald (1994) and Muthén (1978) have suggested that one should not lose too much information in the absence of higher-order marginals in most practical testing situations.

Also, GLS estimation can be computationally burdensome. Although Muthén's (1978) solution is more efficient than Christoffersson's (1975), the procedure, as implemented in LISCOMP (Muthén, 1988), is still impractical using a personal computer with tests containing more than 25 items (Mislevy, 1986; Muthén, 1988).

McDonald's Parameterization

McDonald (1981, 1982a, 1994) also examined the relationship between common IRFs and NLFA. McDonald's (1994) parameterization of Z_{ij} is given by

$$Z_{ij} = f_{i0} + f_{i1}\theta_j, \tag{21}$$

where, in the unidimensional case, $f_{i0} = -\alpha_i\beta_i$ and $f_{il} = \alpha_i$ in Lord's parameterization. The parameter f_{il} is equal to the factor loading of factor l on item i. McDonald (1967, 1994) states that the normal ogive model can be approximated by a polynomial function of the general form,

$$Z_{ij} = f_{i0} + f_{i1}\theta_j + f_{i2}\theta_j^2 + \ldots + f_{ik}\theta_j^k. \tag{22}$$

where $f_{ik}\theta_j^k$ = the factor loading of factor j on item i of polynomial degree k.

Specifically, the IRFs for this model are approximated by a third-degree Hermite-Tchebycheff polynomial. The probability of obtaining a correct response to item i based on this model is given by

$$P(y_i = 1 \mid \theta_j) = c_i + (1 - c_i)N[f_{i0} + f_{il}\theta_j], \tag{23}$$

where c_i corresponds to the lower asymptote parameter estimate outlined in Equation 3 and N is the normal ogive function. Function Equation 23, which is referred to by McDonald (1994) as the *latent trait model*, also generalizes to the multidimensional case as

$$P(y_i = 1 \mid \underline{\theta}) = c_i + (1 - c_i)N[f_{i0} + \underline{f}'_i\underline{\theta}], \tag{24}$$

where, f'_i is a vector of scale parameters. Fraser and McDonald (1988) and McDonald (1981, 1994) also demonstrated that the latent trait model shown in Equation 23 could be derived according to Christoffersson's (1975) parameterization in the form

$$P(Y_i = 1 \mid \underline{\theta}) = c_i + (1 - c_i)N[t_{i0} + \underline{h}'_i\underline{\theta}/m_i]. \tag{25}$$

The parameters modeled in Equations 24 and 25 are related by,

$$t_{i0} = f_{i0}/(1 + \underline{f}'_i \Phi \underline{f}_i)^{1/2},$$

$$h_i = f_i/(1 + \underline{f}'_i \Phi \underline{f}_i)^{1/2},$$

$$m_i^2 = 1/s_i^2,$$

and

$$s_i = (1 + \underline{f}'_i \Phi \underline{f}_i)^{1/2},$$

here ($i = 1, \ldots, m$) and Φ is the m x m matrix containing the correlations among the dimensions, assuming the latent traits have been standardized. A detailed discussion of this relationship is found in McDonald (1994).

The unweighted least squares (ULS) function that is minimized in the estimation of the pairwise probabilities $\hat{\pi}_{ij} = P(X_i = 1, X_j = 1)$ is

$$F = \sum_{i<j}\sum [p_{ij} - \hat{\pi}_{ij}]^2, \tag{26}$$

where p_{ij} corresponds to the proportion of test takers correctly answering both items i and j.

Advantages and Limitations of McDonald's Polynomial Approximation to a Normal Ogive Model

As was previously stated, McDonald's (1967) approach to NLFA employs ULS estimation of the model parameters. ULS estimation is quite economical as compared to generalized least-squares and maximum likelihood procedures and hence has the practical advantage of allowing for the analysis of tests with a fairly large number of items and/or dimensions.

Also, McDonald's model has been implemented in the computer program NOHARM (Fraser & McDonald, 1988). The program enables the user to fit confirmatory or exploratory unidimensional and multidimensional models to item response matrices. The NOHARM output includes the results from the latent trait parameterization, the common factor model reparameterization as well as, in the unidimensional case, Lord's parameterization. In addition, a residual joint-proportions matrix is included in the output which can be useful to assess the fit of a given model. However, the greater degree of computational efficiency associated with the ULS estimation procedure is achieved at the sacrifice of information (Mislevy, 1986). That is, only the information in the one-way marginals (proportion of test takers who correctly answer each item) and two-way marginals (proportion of test takers who correctly answer pairs of items) is utilized by NOHARM in the estimation of parameters, thus explaining why it is often referred to as a "limited" or "bivariate" factor analytic method. Knol and Berger (1991) compared NOHARM parameter estimates to those obtained based on FIFA (i.e., using TESTFACT) and generally found only slight differences between the two procedures with respect to their ability in recovering (simulated) factor analytic parameters. However, these findings were based on a limited number of replications (10) and should be interpreted cautiously. Nonetheless, from a practical perspective, it would seem that there might not be much to be gained in using full-information methods. Balassiano and Ackerman (1995b) have also shown that the overall performance of NOHARM, with respect to recovering simulated item parameter values, was satisfactory, even with small sample sizes ($N = 200$).

Another limitation of the model, again attributable to the ULS estimation procedure, is the absence of standard errors for the parameter estimates and a fit statistic for the given model. However, McDonald (1994) and Balassiano and Ackerman (1995b) have suggested criteria (e.g., the inverse of the square root of the sample size) that may be used as approximate standard errors for the parameters of the model. Also, an approximate χ^2 statistic, based on the residuals obtained after fitting a NLFA (NOHARM) model to an item response matrix, was proposed and investigated by De Champlain (1992) and Gessaroli and De Champlain (1996). Results obtained with a variety of simulated data sets showed that the approximate χ^2 statistics were quite accurate in correctly determining the number of factors underlying simulated item responses. This would suggest that these procedures might be useful as practical guides for the assessment of model fit, even though they are perhaps not the theoretically preferred statistics due to the ULS estimation method on which they're based. Nonetheless, further research needs to be undertaken in order to evaluate the behavior of these approximate χ^2 statistics in a larger number of conditions before making any definite statements about their usefulness.

Finally, some authors have noted that another problem with McDonald's model is the absence of an index that would indicate the appropriate number of polynomials to retain in a series (Hambleton & Rovinelli, 1986). Findings pertaining to this question, however, seem to indicate that terms beyond the cubic can generally be ignored (McDonald, 1982b; Nandakumar, 1991).

Illustration

In order to determine the degree of similarity between the BILOG, NOHARM, and TESTFACT IRT item discrimination and difficulty parameter estimates for LSAT datasets, separate calibrations were undertaken for two forms. More precisely, item difficulty and discrimination parameters were estimated for 101 October 1992 and 101 October 1994 LSAT items using the three above mentioned procedures. Pseudo-guessing parameters that are usually estimated during LSAT equatings were not estimated here. The October 1992 LSAT form was administered to 45,918 test takers while the October 1994 LSAT form was given to 42,361 test takers. Both datasets excluded test takers who required an accommodated testing situation. Item parameter descriptive statistics are presented for both test forms by estimation procedure in Table 1.

TABLE 1
IRT parameter descriptive statistics by form and estimation procedure

	October 1992 LSAT			October 1994 LSAT		
Statistics	BILOG	NOHARM	TESTFACT	BILOG	NOHARM	TESTFACT
Mean *a*	0.773	0.661	0.546	0.632	0.652	0.557
Standard deviation *a*	0.177	0.163	0.117	0.177	0.195	0.153
Minimum *a*	0.407	0.352	0.295	0.259	0.277	0.238
Maximum *a*	1.383	1.390	0.922	1.188	1.604	1.047
Mean *b*	0.187	0.029	0.046	-0.001	0.026	0.043
Standard deviation *b*	0.961	1.156	1.145	1.156	1.168	1.354
Minimum *b*	-2.133	-2.425	-3.328	-2.514	-2.524	-2.877
Maximum *b*	2.044	2.246	2.837	2.476	2.483	3.135

For both LSAT forms that were examined, the parameter estimates tended to be similar, irrespective of the procedure employed. The mean absolute difference between BILOG and NOHARM item discrimination parameter estimates was equal to 0.116 for the October 1992 administration and 0.031 for the October 1994 administration. The mean absolute difference between BILOG and TESTFACT item discrimination parameter estimates was equal to 0.227 for the October 1992 administration and 0.075 for the October 1994 administration. The mean absolute difference between BILOG and NOHARM item difficulty parameter estimates was equal to 0.213 for the October 1992 administration and 0.036 for the October 1994 administration. Finally, the mean absolute difference between BILOG and TESTFACT item difficulty parameter estimates was equal to 0.374 for the October 1992 administration and 0.173 for the October 1994 administration. Plots of the BILOG and NOHARM item discrimination and difficulty parameter estimates are provided in Figures 1a, 1b, 1c, and 1d.

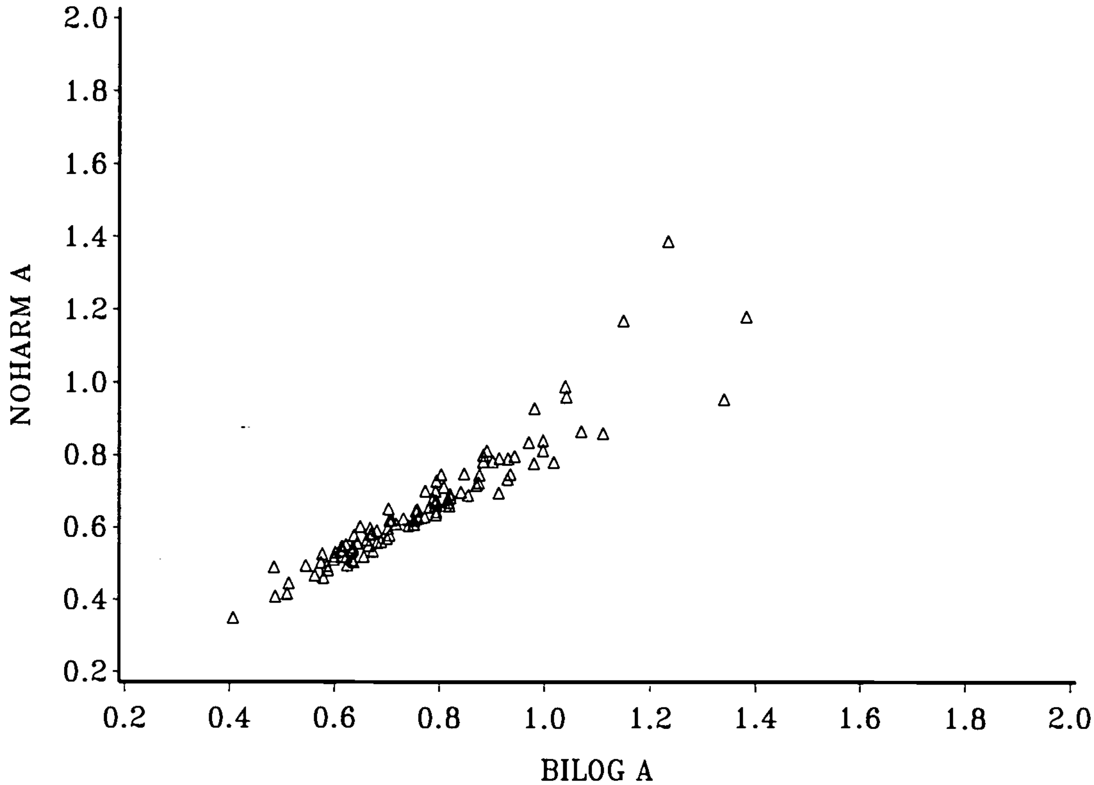

FIGURE 1a. *Comparing BILOG and NOHARM item discrimination parameter estimates, October 1992 LSAT administration*

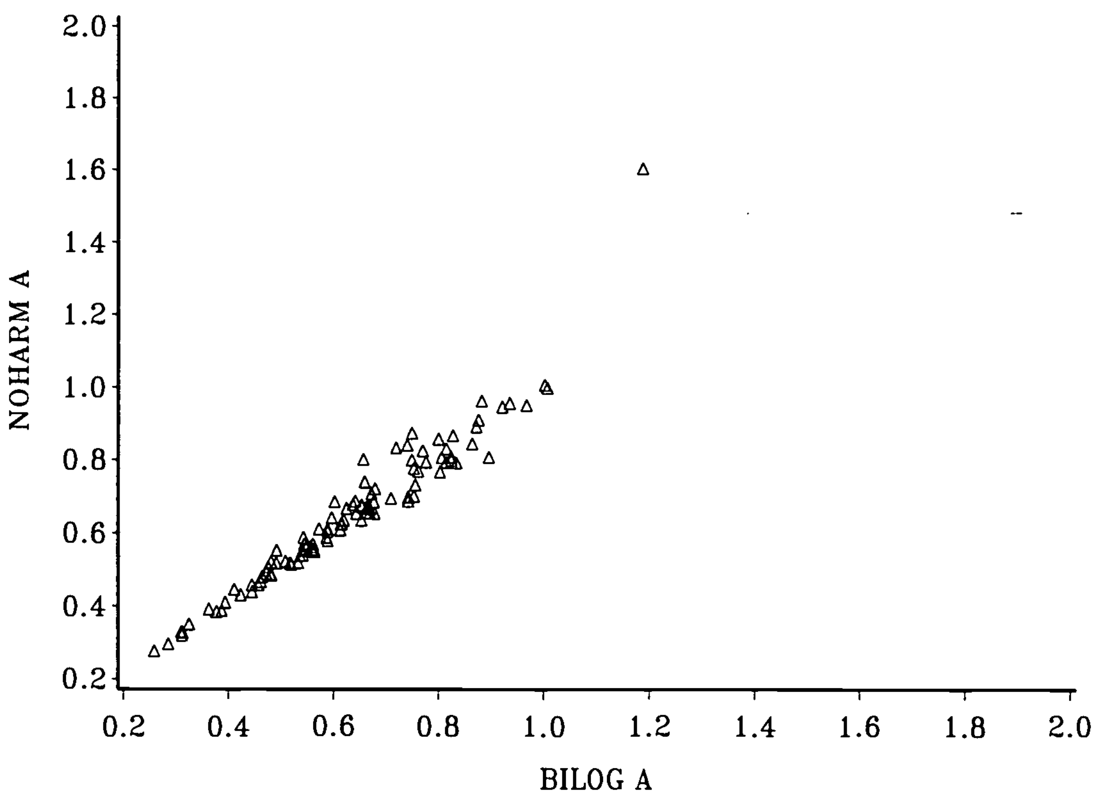

FIGURE 1b. *Comparing BILOG and NOHARM item discrimination parameter estimates, October 1994 LSAT administration*

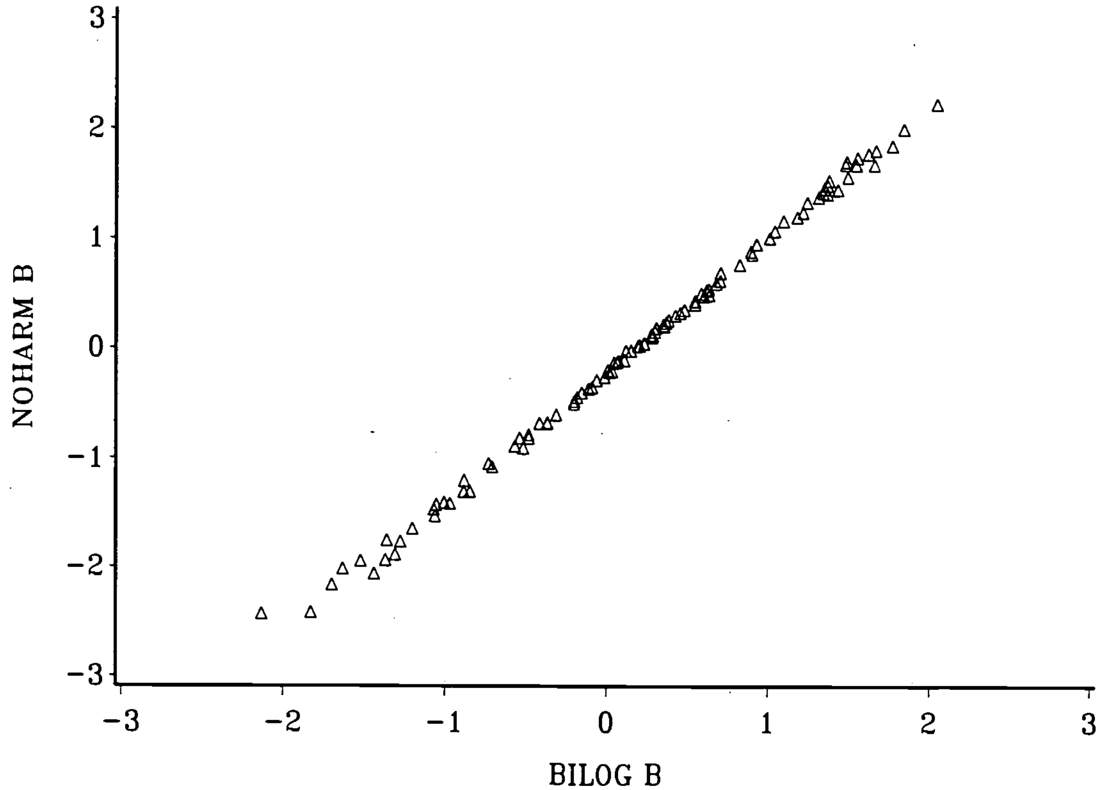

FIGURE 1c. *Comparing BILOG and NOHARM item difficulty parameter estimates, October 1992 LSAT administration*

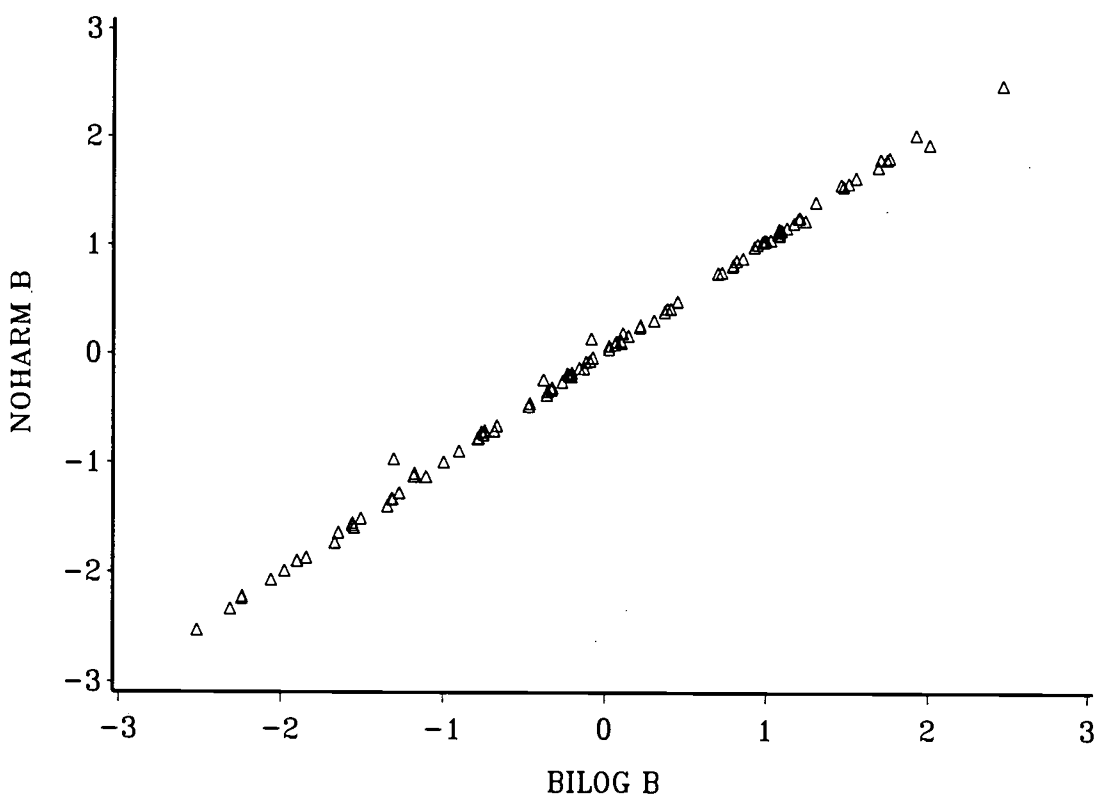

FIGURE 1d. *Comparing BILOG and NOHARM item difficulty parameter estimates, October 1994 LSAT administration*

The relationship between the parameter estimates obtained using BILOG and NOHARM was very strong. The correlation between the BILOG and NOHARM item discrimination parameter estimates was high for both the October 1992 ($r = 0.942$) and October 1994 ($r = 0.963$) LSAT forms. In addition, the correlation between the BILOG and NOHARM item difficulty parameter estimates was nearly perfect for the October 1992 ($r = 0.999$) as well as October 1994 ($r = 0.999$) LSAT forms. Plots of the BILOG and TESTFACT item discrimination and difficulty parameter estimates are provided in Figures 2a, 2b, 2c, and 2d.

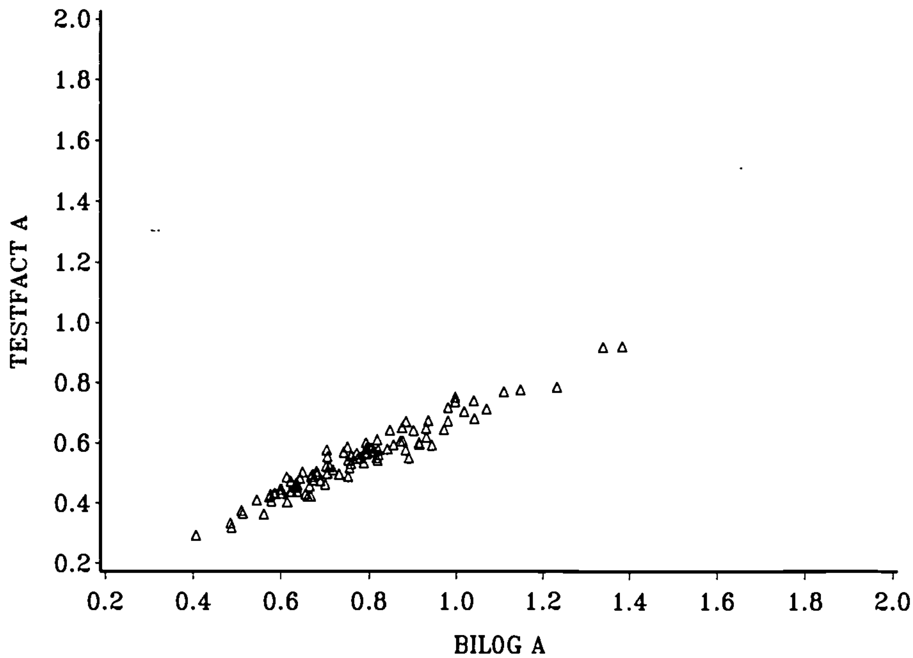

FIGURE 2a. *Comparing BILOG and TESTFACT item discrimination parameter estimates, October 1992 LSAT administration*

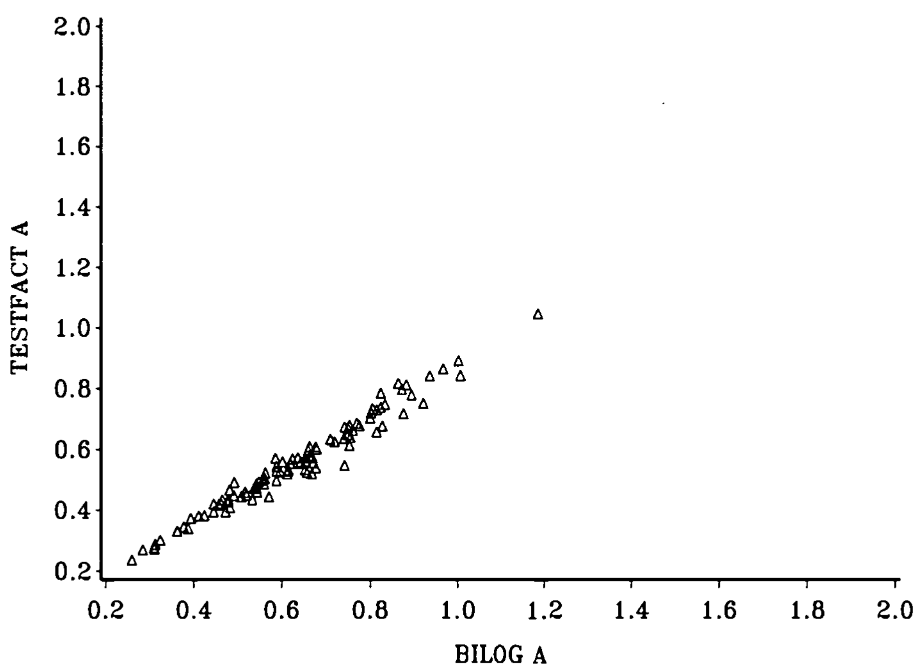

FIGURE 2b. *Comparing BILOG and TESTFACT item discrimination parameter estimates, October 1994 LSAT administration*

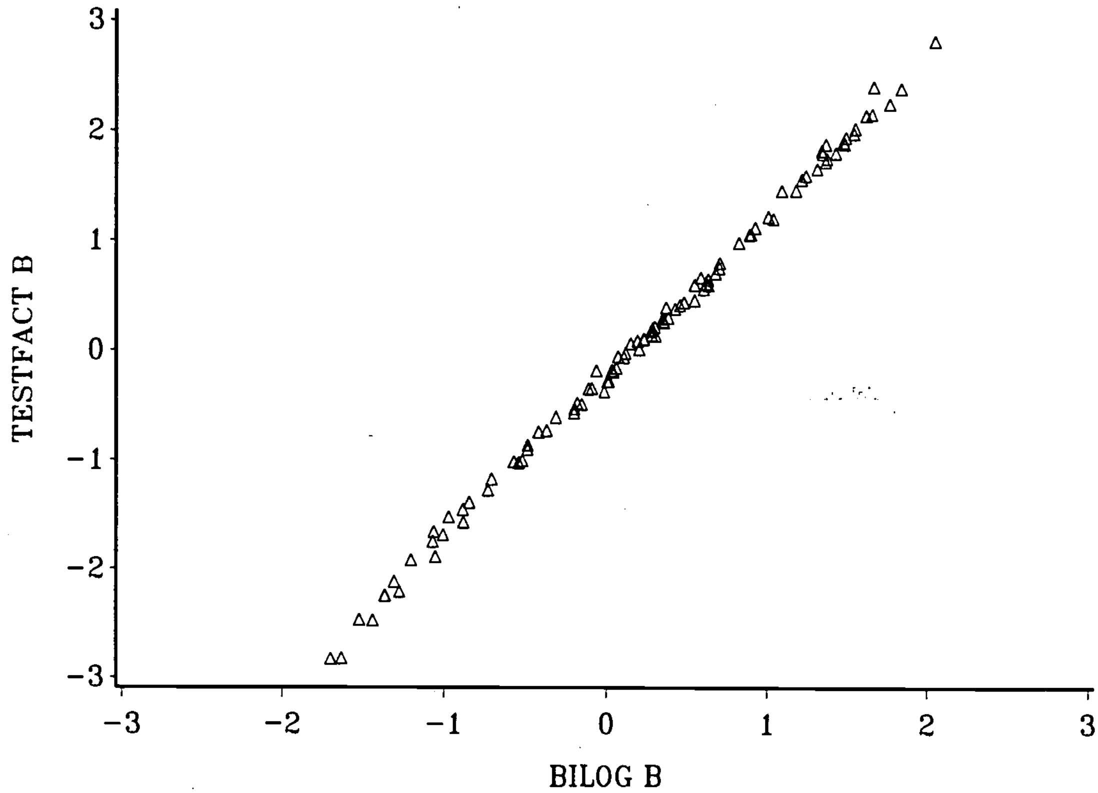

FIGURE 2c. *Comparing BILOG and TESTFACT item difficulty parameter estimates, October 1992 LSAT administration*

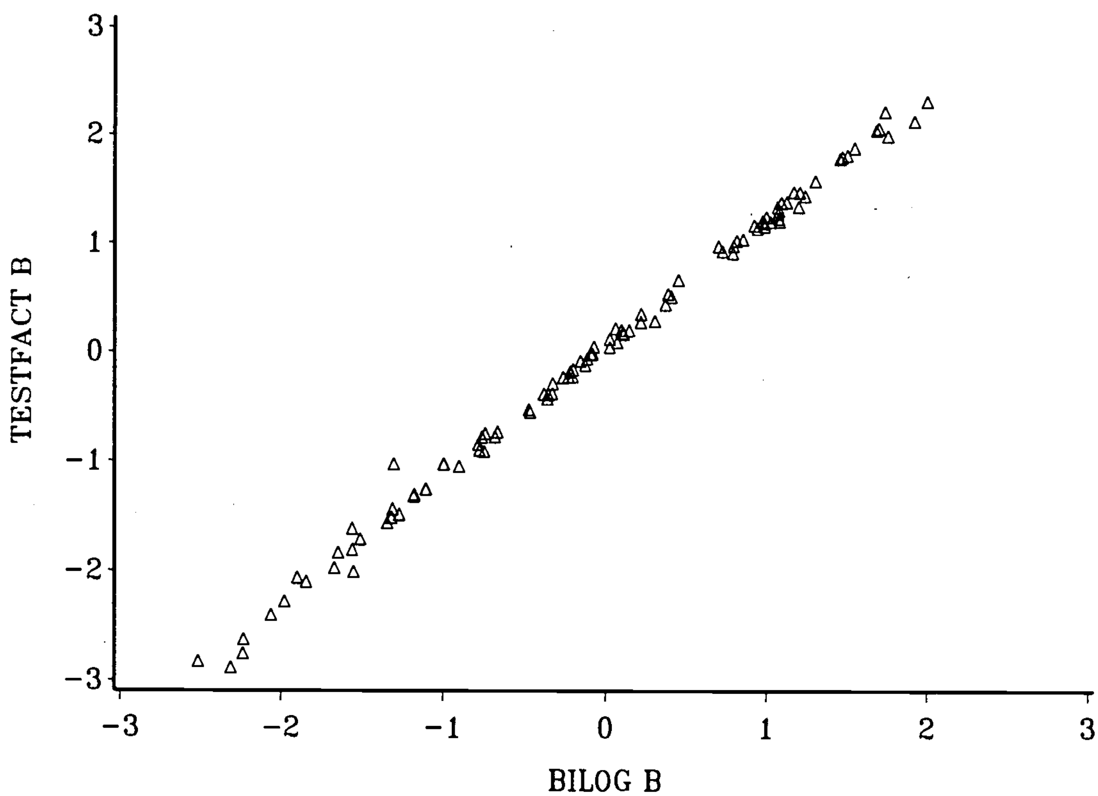

FIGURE 2d. *Comparing BILOG and TESTFACT item difficulty parameter estimates, October 1994 LSAT Administration*

The relationship between the parameter estimates obtained using BILOG and TESTFACT was also very strong. The correlation between the BILOG and NOHARM item discrimination parameter estimates was high for both the October 1992 ($r = 0.969$) and October 1994 ($r = 0.983$) LSAT forms. In addition, the correlation between the BILOG and TESTFACT item difficulty parameter estimates was nearly perfect for the October 1992 ($r = 0.999$) as well as October 1994 ($r = 0.998$) LSAT forms.

Although these results are preliminary, as they are based on only two data sets, some tentative conclusions can be drawn based on the analyses. Beforehand, it is important to point out that the criterion utilized against which to compare NOHARM and TESTFACT item parameter estimates (i.e., BILOG) is by no means infallible and should be viewed as such. BILOG seemed to be the most appropriate yardstick given that it is an extensively used and studied calibration procedure and is the one currently being used by LSAC staff. Having said this, the evidence gathered would seem to indicate that the item difficulty and discrimination parameter estimates, on the whole, differ only slightly across procedures. However, the scatterplots do reveal the presence of a few outlying pairs of item discrimination parameter estimates that should be examined in future research in order to better understand the limitations of each calibration method. Also, these preliminary findings suggest that the item discrimination parameters are underestimated by TESTFACT, as compared to BILOG, which is consistent with past simulation studies (Boulet, 1995). Finally, it appears as though the IRT item parameters were more similar for the October 1994 LSAT form. Again, a larger number of analyses should be undertaken across multiple LSAT forms and simulated data sets before any definite conclusions are made with respect to the usefulness of these three calibration procedures with LSAT data sets.

Conclusion

IRT models have been used extensively in the past few decades not only in the development and analysis of educational test items but also in a host of other applications, such as for the equating of alternate test forms and the detection of differentially functioning items. Several researchers have suggested, however, that common IRT models are really specific cases of more general NLFA models (Goldstein & Wood, 1989; Knol & Berger, 1991; McDonald, 1967; Takane & De Leeuw, 1987). The research conducted by the last mentioned authors clearly shows that common IRT models, such as those based on the normal ogive and logistic functions, can easily be expressed with factor analytic parameterizations. The findings obtained in these studies would therefore seem to suggest that NLFA might provide a useful framework with which to address measurement-related issues that had been primarily investigated using IRT models.

Three factor analytic models were briefly outlined. More precisely, McDonald's (1967, 1982b) polynomial approximation to a normal ogive model, Christoffersson's (1975)/Muthén's (1978) factor analysis model for dichotomous variables and Bock and Aitkin's (1981)/Bock, Gibbons, and Muraki's (1988) full-information factor analytic model, were described. In addition, the major strengths and weaknesses of each model were delineated. Table 2 provides a comparison of the main features of these models.

TABLE 2
A comparison of three nonlinear factor analytic models

	Polynomial approximation (McDonald, 1967)	Factor analytic model for dichotimized variables (Christoffersson, 1975)	Full-information factor analysis (Bock & Aitkin, 1981)
Estimation procedure	ULS	GLS	MML
Computer program	NOHARM (Fraser & McDonald, 1988)	LISCOMP (Muthén, 1988)	TESTFACT (Wilson, Wood, & Gibbons, 1987)
Fit confirmatory analyses?	Yes	Yes	No
Information used	Lower-order marginals	Lower-order marginals	Higher-order marginals
Standard errors for parameter estimates?	No	Yes	Yes
Tests of model fit?	No	Yes	Yes

Based on this information, are there any conditions that might dictate the use of one model rather than another as they are currently implemented in the various software packages? As shown in Table 2, the fit of confirmatory factor analytic models to item response matrices can, at present, be estimated using either NOHARM or LISCOMP. As previously stated, the current version of TESTFACT restricts users to fitting exploratory models. It is important to point out, however, that Muraki (1991) has undertaken research aimed at incorporating prior information in TESTFACT estimation procedures. Also, the estimation of factor correlations cannot be undertaken with TESTFACT given the orthogonal structure of the factor analytic model.

Another issue that might be considered by practitioners interested in fitting a factor analytic model to item response data is the amount of information utilized by the estimator. With respect to this point, the FIFA model implemented in TESTFACT would seem to be the preferred choice given its use of higher-order marginals in the estimation process. However, in practice, the use of the full information is usually not feasible. For example, in order to utilize all of the information contained in a 30-item test when fitting a FIFA model, 2^{30} or 1,073,741,824 test takers are required. Hence, in most situations, FIFA reduces to a limited-information model.

Also, the availability of standard errors and fit statistics might be a factor to consider when selecting one of the factor analytic models. For example, the ULS estimation procedure implemented in NOHARM does not allow for a valid chi-square test to aid the practitioner in selecting the best fitting model. In addition, standard errors for the parameters estimated in a given model are unavailable with ULS estimation. On the other hand, the output from both TESTFACT (Wilson, Wood, & Gibbons, 1987) and LISCOMP (Muthén, 1988) contains chi-square fit statistics as well as standard errors. It must be noted, however, that approximate chi-square fit statistics (De Champlain, 1992; Gessaroli & De Champlain, 1996) and standard errors (Balassiano & Ackerman, 1995b; McDonald, 1994) have been proposed to accompany the NOHARM (Fraser & McDonald, 1988) output.

To illustrate similarities and differences in IRT item difficulty and discrimination parameter estimation, two LSAT forms were analyzed using the procedures implemented in BILOG, NOHARM, and TESTFACT. Findings show that the parameter estimates tended to be quite similar, regardless of the calibration procedure. However, additional research should be conducted with a larger number of LSAT forms before reaching any definite conclusions as to the degree of comparability of the calibration procedures.

In summary, the purpose of this paper was to underscore the usefulness of the NLFA framework in addressing common educational measurement problems, as well as to provide practitioners with an overview of the strengths and limitations of three factor analytic models. Also, empirical analyses were undertaken using LSAT item response data in order to provide evidence to support the claim that the calibration procedures provide very similar item parameter estimates. Hopefully, this overview and these preliminary analyses will help the practitioner in arriving at a more informed decision when contemplating the selection of a NLFA model and foster future research with respect to the potential application of these models in educational measurement and more specifically, with the LSAT.

References

Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker Inc.

Balassiano, M., & Ackerman, T. (1995a). *An in-depth analysis of the NOHARM estimation algorithm and implications for modeling the multidimensional latent ability space*. Unpublished manuscript, University of Illinois at Urbana-Champaign, Faculty of Education, Urbana-Champaign.

Balassiano, M., & Ackerman, T. (1995b). *An evaluation of NOHARM estimation accuracy with a two-dimensional latent space*. Unpublished manuscript, University of Illinois at Urbana-Champaign, Faculty of Education, Urbana-Champaign.

Bartholomew, D. J. (1983). Latent variable models for ordered categorical data. *Journal of Econometrics, 22*, 229-243.

Berger, M. P. F., & Knol, D. L. (1990, April). *On the assessment of dimensionality in multidimensional item response theory models*. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 4*, 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261-280.

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

Boulet, J. (1995, April). *A comparison of parameter estimates using limited-information and full-information nonlinear factor analysis*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*(1), 5-32.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 175-195). Vancouver, BC: Educational Research Institute of British Columbia.

De Champlain, A. (1992). *Assessing test dimensionality using two approximate chi-square statistics*. Unpublished doctoral dissertation, University of Ottawa, Ottawa, Ontario, Canada.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

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

Gessaroli, M. E., & De Champlain, A. (1996). Using an approximate chi-square statistic to test for the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement, 33*, 157-179.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

Haberman, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Annals of Statistics, 5*, 1148-1169.

Hambleton, R. K. (1983). *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational Measurement* (pp. 147-200). New York: American Council on Education and Macmillan Publishing Company.

Hambleton, R. K., & Rovinelli, R. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

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

Hambleton, R. K., Zaal, J., & Pieters, J. P. M. (1993). Computer adaptive testing: Theory, applications, and standards. In R. K. Hambleton and J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (pp. 341-366). Boston: Kluwer Academic Publishers.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Homewood, IL: Dow-Jones Irwin.

Kendall, M. G. (1941). Relations connected with the tetrachoric series and its generalization. *Biometrika, 32,* 196-198.

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4,* 241-261.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison 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. (1977). Practical applications of characteristic curve theory. *Journal of Educational Measurement, 14,* 117-138.

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

Lord, F. M. (1982). Item response theory and equating—A technical summary. In P. W. Holland & D. B. Rubin (Eds.), *Test Equating* (pp. 141-148). New York: Academic Press.

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

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometrika Monograph No. 15, 32*(4, Pt. 2).

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

McDonald, R. P. (1982a). Some alternative approaches to the improvement of measurement in education and psychology: Fitting latent trait models. In D. Spearitt (Ed.), *The improvement of measurement in education and psychology* (pp. 213-233). Hawthorn, VI: Australian Council for Educational Research.

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

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, ON: Edumetrics Research Group, University of Ottawa.

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. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software, Inc.

Muraki, E. (1991, April). *Confirmatory full-information factor analysis of the NAEP data.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

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

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

Muthén, B. (1988). *LISCOMP*. Mooresville, IN: Scientific Software, Inc.

Nandakumar, R. (1991, April). *Assessing the dimensionality of a set of items—Comparison of different approaches.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.), *Educational Measurement* (pp. 221-262). New York: American Council on Education and Macmillan Publishing Company.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 4*, 495-529.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Thissen, D. (1993). *MULTILOG 6*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R., Steinberg, L., & Thissen, D. (1990). *Computerized Adaptive Testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Warm, T. A. (1978). *A primer of item response theory* (Tech. Rep. No. OG-941278). Oklahoma City, OK: U.S. Coast Guard Institute.

Wilson, D., Wood, R., & Gibbons, R. D. (1987). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Chicago: Scientific Software International

Wingersky, M. S., Patrick, R., & Lord, F. M. (1991). *LOGIST VI user's guide*. Princeton, NJ: Educational Testing Service.

NOTICE

Reproduction Basis

X This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").