

ESTIMATING THE LATENT TRAIT FROM LIKERT-TYPE DATA:  
A COMPARISON OF FACTOR ANALYSIS, ITEM RESPONSE  
THEORY, AND MULTIDIMENSIONAL SCALING

APPROVED BY SUPERVISORY COMMITTEE:

William R. Koch

John C. Locke

William J. ...

W. Paul Kelley

Barbara B. Dodd

Copyright

by

Chihyu Chan

1991

To My Parents

and

Pichun and Hain-Ruey

and

Guy

ESTIMATING THE LATENT TRAIT FROM LIKERT-TYPE DATA:  
A COMPARISON OF FACTOR ANALYSIS, ITEM RESPONSE  
THEORY, AND MULTIDIMENSIONAL SCALING

by

CHIHU CHAN, B.ED., M.ED.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December, 1991

#### ACKNOWLEDGEMENTS

This dissertation was sponsored by Chiang Ching-Kuo Foundation for International Scholarly Exchange (USA). Costs for computer software and computer time were partially supported by the APA Student Dissertation Research Awards. An earlier version of this research was presented in the 1991 Annual Meeting of the American Psychological Association, San Francisco. Partial costs of that traveling and presentation were supported by The APA Student Travel Awards.

The author wishes to express his deep appreciation to the chair of the dissertation committee, Dr. William R. Koch, for his dedicated supervision, patient guidance, and numerous helpful suggestions.

The author would also like to thank the other committee members, Dr. Barbara G. Dodd, Dr. William L. Hays, Dr. H. Paul Kelley, and Dr. John C. Loehlin, for their stimulating criticisms and valuable feedbacks.

ESTIMATING THE LATENT TRAIT FROM LIKERT-TYPE DATA:  
A COMPARISON OF FACTOR ANALYSIS, ITEM RESPONSE  
THEORY, AND MULTIDIMENSIONAL SCALING

Publication No. \_\_\_\_\_

Chihyu Chan, Ph.D.  
The University of Texas at Austin, 1991

Supervising Professors: William R. Koch

Seven statistical procedures were compared with one another in terms of the ability to recover a unidimensional latent trait from Likert-type data. They are factor analysis based on either Pearson correlations (FA-PR) or polychoric correlations (FA-PL), the graded response model in item response theory (IRT-GRM), internal unfolding (IMDU), external unfolding (EMDU), weighted unfolding (WMDU), and the common procedure of summing up successive integers assigned to response categories (SSI). Sample size, test length, and skewness of item response distributions were manipulated in this simulation

study. Generally speaking, IRT-GRM performed the best and was most robust against skewness. FA-PR and FA-PL performed equally well across almost all conditions but were competitive with IRT-GRM only when item responses were normally distributed. SSI practice might be slightly worse than the two FA procedures when item responses were normally distributed, but it was better than them when item responses were highly skewed. WMDU performed as well as did SSI only when item responses were normally distributed or moderately skewed and sample size was large for MDS models (e.g.,  $N=100$ ). IMDU and EMDU were even worse than WMDU and appeared not appropriate for Likert-type data.

TABLE OF CONTENTS

<u>Chapter</u>		<u>Page</u>
I	Introduction .....	1
	The Postulated Scaling Situation .....	1
	Traditional Approaches .....	6
	Latent Variable Models .....	7
	Objectives .....	11
II	Literature Review .....	12
	Latent Variables .....	12
	Factor Analysis .....	17
	Basic Theories .....	17
	Ordered Polychotomous Indicators ..	22
	Item Response Theory .....	27
	Basic Theories .....	27
	Ordered Polychotomous Items .....	32
	Multidimensional Scaling .....	38
	Basic Theories .....	39
	Ordered Preference Data .....	41
	Comparative Analysis .....	51
	FA and IRT .....	51
	FA and MDS .....	57
	IRT and MDS .....	64
III	Statement of Problem .....	68



IV	Methodology .....	70
	The Simulated Situation .....	70
	Background Conditions .....	71
	Dependent Variables .....	80
	Independent Variables .....	81
	Programs for Estimation .....	86
V	Results .....	90
	Case I: N=1000 .....	91
	Case II: N=100 .....	95
	Case III: N=30 .....	100
	Conclusions Across Cases .....	104
VI	Discussion .....	106
	FA-PR vs. FA-PL .....	106
	WMDU vs. IMDU vs. EMDU .....	107
	IRT vs. FA vs. MDS .....	108
	SSI vs. Other Procedures .....	109
	Selection among Procedures .....	110
	Limitations .....	112
	References .....	116
	Vita .....	130

LIST OF TABLES

<u>Table</u>	<u>Page</u>
II-1	Latent Variable Models Utilizing the Principle of Local Independence ... 14
II-2	Threshold Values Defined in the Graded Response Model ..... 38
IV-1	Distributional Characteristics of the 12 Core Items in the Condition of Normal Distribution ..... 72
IV-2	Distributional Characteristics of the 12 Core Items in the Condition of Moderately Skewed Distribution .... 74
IV-3	Distributional Characteristics of the 12 Core Items in the Condition of Highly Skewed Distribution ..... 76
IV-4	Distributional Characteristics of the 12 Core Items in the Condition of Differentially Skewed Distribution 78
IV-5	Range, Mean, and SD of Skewness Values of 12 Items in Each Condition of Item Response Distributions ..... 86
V-1	Correlations between True and Recovered Person Parameters for Case I ..... 92

V-2	RMSDs between True and Recovered Person Parameters for Case I ..... 93
V-3	Correlations between True and Recovered Person Parameters for Case II ..... 97
V-4	RMSDs between True and Recovered Person Parameters for Case II ..... 98
V-5	Correlations between True and Recovered Person Parameters for Case III ... 102
V-6	RMSDs between True and Recovered Person Parameters for Case III ... 103

LIST OF FIGURES

<u>Figure</u>		<u>Page</u>
I-1	Response Processes .....	2

CHAPTER I  
INTRODUCTION

The Postulated Scaling Situation

Data collected from social/psychological research typically produces ordinal manifest variables (Cliff, 1989; Hildebrand, Laing & Rosenthal, 1977). For example, about one half of all recorded observations in the 1975 General Social Survey were obtained through use of the Likert-type response format (Clogg, 1979). However, it is usually assumed that the ordinal manifest variable ( $Y$ ) is obtained through some crude classification of a continuous variable ( $Y^*$ ), which might have been obtained if an interval scale were available. In addition, the continuous response variable ( $Y^*$ ) is assumed to be related except for measurement error to an underlying latent dimension ( $\theta$ ), which is the variable of ultimate interest in most social/psychological researches. This situation is illustrated in Figure I-1.

Two kinds of disturbance processes are assumed to be involved in measuring a latent dimension given

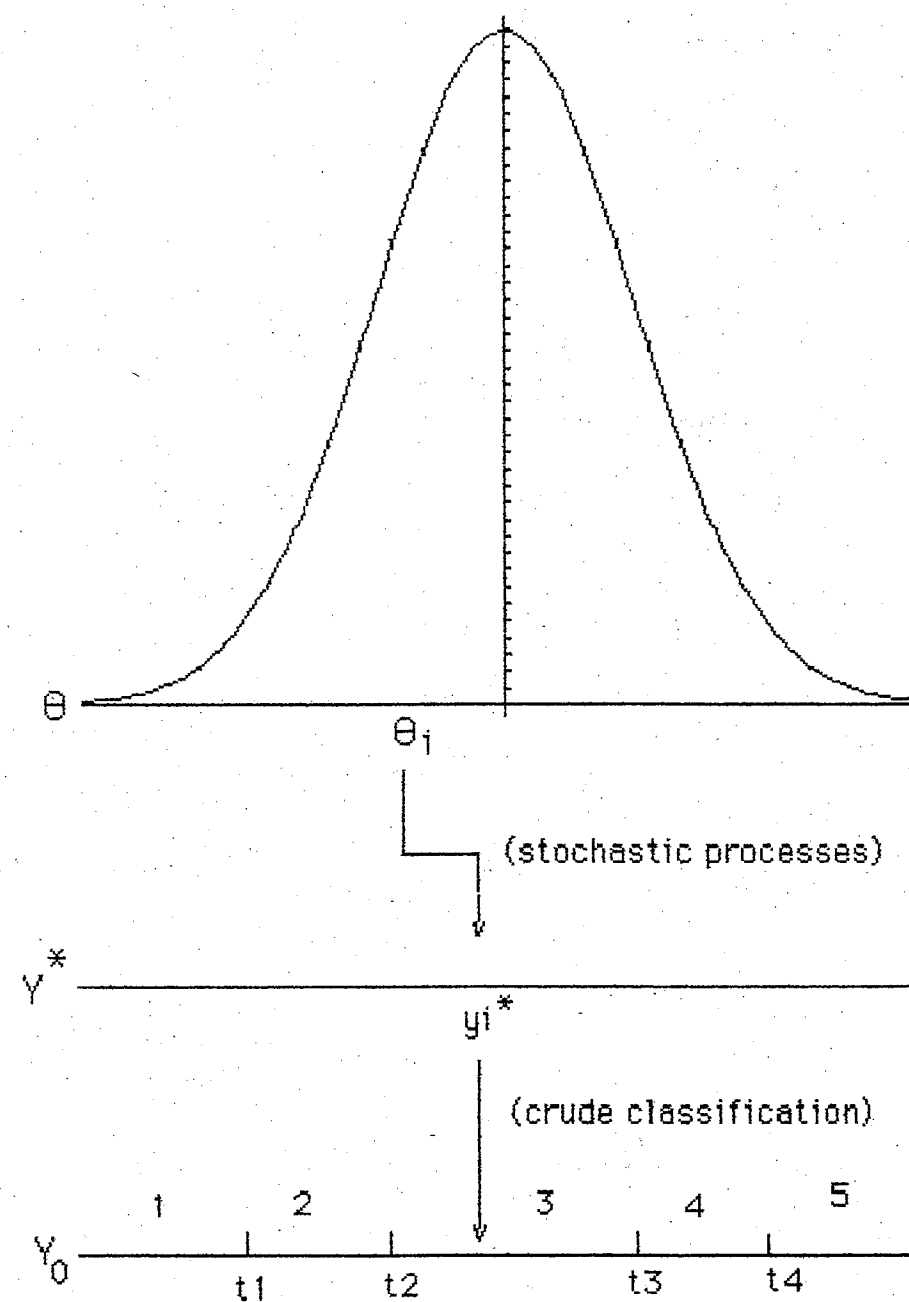


Figure I-1: Response Processes of Likert-Type Scales

Likert-type items: a stochastic process and a crude-classification process. First of all, it is usually assumed that the latent dimension ( $\theta$ ) and the continuous quantitative response ( $Y^*$ ) are linearly and probabilistically related. The basic mathematical form of this relationship is:

$$Y^* = W \theta + E, \quad (1.1)$$

where  $W$  is a weight and  $E$  is the residual. The latent dimension ( $\theta$ ) is assumed to be stable across various replicated observations, while the residual ( $E$ ) is assumed to be specific to replications. For estimation convenience, both  $\theta$  and  $E$  are frequently assumed to be normally distributed in the population. In addition, because of the limitations of the instrument, the continuous quantitative response ( $Y^*$ ) is unavailable and is classified into an ordinal scale ( $Y$ ). In terms of underlying psychological processes, it could be that a person compares his/her potentially quantitative response to the implicit threshold values ( $t_1$  to  $t_4$  in Figure I-1) on the ordinal scale and chooses one corresponding response

category. Therefore, the relationship between the manifest categorical variable ( $Y$ ) and the quantitative response variable ( $Y^*$ ) is an increasing step function. Supposing that five response categories are employed, the step function can be represented in the following scheme:

$$\begin{aligned}
 Y &= 1, \text{ if } Y^* < t_1 \\
 Y &= 2, \text{ if } t_1 \leq Y^* < t_2 \\
 Y &= 3, \text{ if } t_2 \leq Y^* < t_3 \\
 Y &= 4, \text{ if } t_3 \leq Y^* < t_4 \\
 Y &= 5, \text{ if } t_4 \leq Y^*
 \end{aligned}
 \tag{1.2}$$

Where  $t_i$  ( $i=1, 2, 3, 4$ ) are the threshold values or category boundaries. These values may be determined by the properties of the items as well as by the labels of the response categories. They are assumed to be stable across persons. It should be noticed that the variable of interest is  $\theta$  instead of  $Y^*$ . The latent dimension,  $\theta$ , is related to the manifest variable ( $Y$ ) through both a linear/stochastic function and a step function.



Replicated observations of the same latent dimensions are necessary for any scaling model because systematic variations cannot be differentiated from non-systematic fluctuations without multiple observations. The typical procedures for estimating reliability coefficients, including methods of test-retest, alternative forms, multiple raters, and multiple items, are just replicated observations at different levels or with different procedures which are sensitive to different sources of errors. These replicated observations can also be called "multiple indicators" of several latent dimensions. The present study, however, focused on the case where multiple items were indicators of one latent trait partly because that case is the fundamental basis of the multidimensional case and partly because most IRT models have been developed only for the unidimensional case.

The objective of a scaling model in the described situation is clear: to estimate the level of the underlying latent dimension given ordinal multiple indicators with disturbance from a stochastic process and a crude classification. The

assumptions are also clear, including one continuous latent dimension, multiple indicators, normal distribution of the latent dimension and of the residuals, and invariance of the response threshold values across subjects.

#### Traditional Approach

For estimating the latent trait underlying Likert-type items, the traditional approach is to assign successive integers to the response categories and then simply sum up the raw scores on each item to obtain a total score, which serves to estimate the true score of each person on the underlying dimension. This approach (SSI, Sum of Successive Integers) has been often criticized for its assumption of equal intervals between ordinal response categories. However, this approach was employed as a base-line for comparing the more sophisticated latent variable models because it is a common practice in the literature.

### Latent Variable Models

Due to methodological developments in the last three decades, three types of latent variable models growing out of three distinct areas have been used to estimate the underlying trait from ordinal data in a relatively sophisticated fashion. They are factor analysis (FA) of categorical data, item response theory (IRT), and non-metric multidimensional scaling (MDS). Despite their disparate traditions, the three approaches have been treated by a few researchers (e.g., Loehlin, 1987) under a more general concept, latent variable models, because they all attempt to reduce a large number of manifest variables into a few hypothetical latent variables constructed in the mathematical model.

Given a correlation or covariance matrix among manifest variables, the objective of FA is to estimate some latent variables which constitute the complete latent space so that the relationships among indicators disappear within a homogeneous subpopulation with respect to these latent variables. That is, all pairwise partial correlations among the manifest variables approach zero given that the

latent variables (factors) have been kept constant. After the factor structures have been determined, factor scores for individuals on the latent dimensions can be estimated. Because the observed data are assumed to be ordinal, to perform FA on the matrix of Pearson correlations computed from these data (FA-PR) may be criticized. A theoretically favored alternative is to first estimate polychoric correlations among ordered variables (Olsson, 1979b) and then to perform FA on the matrix of polychoric correlations (FA-PL). FA-PL makes it possible to start the estimation of latent variables with ordinal manifest variables but to end with results on equal-interval scales. The relative performances of FA-PL and FA-PR were the first focus of the present study.

IRT typically has been developed for scaling categorical response data onto an equal-interval scale. Traditional IRT requires that one single latent variable constitutes the complete latent space in the data so that local independence is achieved as a consequence. Unidimensionality may be seen as a possible result of FA, but it is a presupposition of most IRT models. Although FA and IRT seem to be

evolved from two separate traditions, they have been shown to converge in many aspects (McDonald, 1985). For dealing with ordinal data, the marginal likelihood of Samejima's (1969) normal ogive model and FA for ordered variables have been formally proven to be equivalent (de Leeuw, 1983; Takane, de Leeuw, 1987). In other words, they are different formulations of the same model. It should be noticed that in practice the logistic model is frequently used to approximate the normal ogive model because the logistic model is mathematically more tractable. The logistic model will approximate the FA model to the extent that the logistic model approximates the normal ogive model. The performance of Samejima's (1969) logistic IRT model for graded responses (GRM) was the second focus of the current study.

MDS utilizes the analogy between psychological proximities/preferences and geometric distances to represent stimuli in a perceptual solution space. Given the unidimensional situation, non-metric MDS attempts to represent the stimuli on a common dimension such that the distances among the stimuli on the underlying dimension have the same rank orders

as the observed proximity data. To deal with Likert-type data, Carroll's (1972) external unfolding (EMDU) and Coombs' (1964) internal unfolding (IMDU) models may be applicable. Unfortunately, these kinds of applications were few and were shown to be unsatisfactory by a few researchers (e.g., Koch, 1984). Likert-type data involves both a cumulative and an unfolding mechanism. The cumulative mechanism exists in the relationship between the item and the person, while the unfolding mechanism exists in the relationship between the item response categories and the person. This complex situation may be better modeled by the weighted multidimensional unfolding (WMDU) model (Young, 1984) than by the two classical unfolding models. With WMDU, coordinates for item response categories as well as for persons may be estimated. In addition, item discrimination power may be estimated with the weight for "individual differences." The relative performances of IMDU, EMDU, and WMDU models were the third focus of the present study.

Finally, latent variable models across different areas (FA, IRT, and MDS) were compared to

each other and to the traditional SSI procedure. This was the major focus of the current study.

#### Objectives

In terms of the ability to recover the latent trait from Likert-type data, six latent variable models (FA-PR, FA-PL, IRT-GRM, IMDU, EMDU, WMDU) in three statistical areas were compared to each other and to the common SSI procedure. These comparisons were made through many experimental conditions where sample sizes, test length, and distributions of item responses were manipulated. The estimates of the latent trait values are respectively called factor scores in FA, person parameters in IRT, and ideal points in MDS.

CHAPTER II  
LITERATURE REVIEW

Latent Variables

One objective of psychological research is to explore the nature and functions of some latent traits of individuals. Therefore, the concept of latent variables is central in psychometrics. Nevertheless, there are many statistical ways of modeling the latent variable. For example, classical true score test theory decomposes the observed variable into two latent parts, i.e., the true- and the error-score variables (Lord & Novick, 1968). The true score is defined as the expected value of the observed score or is regarded as the observed score of a person on a homogeneous test of infinite length. Under this conception, the observed score is taken as the unbiased estimator of the true score. Given Likert-type items, however, many researchers are reluctant to assign successive integers to the item response scales arbitrarily and then simply sum up the item scores to obtain a total score as the



estimator of the true score. This reluctance is due to the assumption that an interval scale exists in the arbitrarily assigned successive integers. For simplicity, however, this SSI approach is frequently employed by research practitioners. This practice is also justified by the results of some Monte Carlo studies (e.g., Bollen & Barb, 1981; Jenkins & Taber, 1977; Lissitz & Green, 1975), which showed that errors caused by regarding ordinal scales as interval ones were small enough to be ignored as long as a minimum of 5 response categories were employed. For the reason that "simple is better," Cohen (1990) recently also suggested that, for combining multiple indicators, unit weights instead of optimal weights coming from regression or FA be assigned to the individual items. Because this SSI approach is so widely employed, it served as the baseline for comparing the other latent variable models.

The second way to identify one or more latent variables in psychology is through the statistical concept of local independence, which underlies a body of latent variable models including factor analysis, structural equation modeling, item response theory,

Table II-1  
 Latent Variable Models Utilizing the Principle  
 of Local Independence

		Manifest variables	
		Metrical	Categorical
Latent variables	Metrical	Factor analysis	Latent trait analysis (IRT)
	Categorical	Latent profile analysis Analysis of mixtures	Factor analysis of categorical data Latent class analysis

Note: From Bartholomew (1987, p.4)

latent profile analysis and latent class analysis (Bartholomew, 1987)(see Table II-1). Because the last two analyses assume a non-metric latent variable, they will not be included in the current study.

A psychological interpretation of the term "local independence" is that an individual's systematic performance can be completely explained by some underlying traits so that, given his values on these traits, no more information can be learned from him/her. In other words, his/her performance should

be random once his/her values on these traits have been fixed (Anderson, 1959). Translated into statistical terms, local independence is the statistical independence of residuals once the latent variables in the model constitute the complete latent space and their values are fixed.

However, statistical independence may be defined in different ways in different models. For example, FA defines it in terms of correlations or covariances in the population, while latent class analysis and IRT define it in terms of probabilities. More specifically, FA tries to derive some common factors which constitute the complete latent space that can "explain away" the observed pairwise correlations between manifest variables. Latent class analysis attempts to search for some latent classes such that the relationships in the contingency table of the observed variables will disappear within each latent class. IRT requires the assumption of unidimensionality and utilizes its consequence of local independence to estimate person and item parameters in such a manner that "the probability of the response pattern for each examinee is equal to

the product of the probability associated with the examinee response to each item" at a fixed value of the latent trait (Hambleton & Swaminathan, 1985, p.23).

The third way to define a latent variable is in terms of spatial distance. Typical examples of spatial models can be found in the family of MDS. Like exploratory FA, MDS attempts to search for a small number of latent variables which can explain the observed relationships among a much larger number of stimuli (Loehlin, 1987). In other words, it tries to construct a spatial configuration which can represent the quantity or rank order information of all observed pairwise proximities. Unlike common FA, however, MDS does not utilize the principle of local independence. Instead, it (especially classical metric MDS) employs algorithms similar to the ones applied by principal components analysis (Davison, 1985; Davison & Srichantra, 1988; Rodgers & Young, 1981).

In summary, there are at least three mathematical ways of defining and estimating a latent variable: a) classical true score theory, b) a body

of models which utilize the principle of local independence, and c) the distance models of MDS. Because the second and the third ways of estimation utilize sophisticated procedures for recovering metric information from non-metric data, they will be explored further in the following sections.

#### Factor Analysis

FA originated from the work of Galton, Pearson and particularly Spearman. The maximum likelihood approach to the estimation of the parameters in the FA model, a statistically respectable approach, was introduced by Lawley forty years ago (Everitt, 1984; Lawley & Maxwell, 1971). However, the computational difficulties were not solved until Jöreskog's work in the late 1960s. More recently, Jöreskog's algorithms were implemented in the LISREL computer package (e.g., Jöreskog & Sörbom, 1984).

#### *Basic Theories*

FA is basically founded on the following two procedures: a) conditional independence, and b) linear least squares regression. Without loss of

generality, the two procedures will be explicated in terms of a correlation matrix and standardized variables.

Given a matrix of correlations between any pair of manifest variables,  $x_1, x_2, \dots, x_k$ , FA tries to estimate an underlying factor  $f_1$  such that:

$$\begin{aligned} x_1 &= w_1 f_1 + e_1 \\ x_2 &= w_2 f_1 + e_2 \\ &\dots \\ x_k &= w_k f_1 + e_k, \end{aligned} \tag{2.1}$$

where  $w_i$  and  $e_i$  ( $i=1, 2, \dots, k$ ) are respectively the regression weight and the residual. Each equation in (2.1) represents the linear least squares regression of a manifest variable on the latent variable  $f_1$  so that the residual  $e_i$  is uncorrelated with the predictor,  $f_1$  (Jöreskog & Sörbom, 1979). If  $f_1$  constitutes the complete latent space, then  $f_1$  is sufficient to explain the observed correlations so that the correlation between any pair of residuals  $e_i$  and  $e_j$  is zero, i.e., conditional independence is upheld. In other words,

$$r(e_i, e_j) = 0 \quad (2.2)$$

$$\text{or } r(x_i, x_j | f_1) = 0, \quad i \neq j. \quad (2.3)$$

In this situation, it can easily be verified that

$$r_{ij} = w_i w_j, \quad i \neq j. \quad (2.4)$$

If the reproduced correlations from the model are significantly different from the corresponding observed correlations, then conditional independence may not be upheld and more than one factor may be required. Therefore, a second factor is introduced as in:

$$\begin{aligned} x_1 &= w_{11} f_1 + w_{12} f_2 + e_1 \\ x_2 &= w_{21} f_1 + w_{22} f_2 + e_2 \\ &\dots \\ x_k &= w_{k1} f_1 + w_{k2} f_2 + e_k. \end{aligned} \quad (2.5)$$

Each equation in (2.5) still represents a linear least squares regression of a manifest variable on

the latent variables,  $f_1$  and  $f_2$ . The aim is still to accomplish conditional independence so that

$$r(x_i, x_j | f_1, f_2) = 0, \quad i \neq j. \quad (2.6)$$

If this aim is achieved, then the two factors span the complete latent space and it can be verified that

$$r(x_i, x_j) = w_{i1} w_{j1} + w_{i2} w_{j2}. \quad (2.7)$$

If the reproduced correlations are significantly different from the corresponding observed correlations, then a third factor may be introduced, and so forth. Certain indices of goodness of fit, such as chi-square tests, are needed to decide whether the differences between the reproduced and the observed correlations are statistically significant or not.

The above basic theories of FA can be easily summarized in matrix form. Let  $x$  represent the vector of manifest variables,  $W$  the matrix of factor loadings,  $f$  the vector of common factor scores, and  $e$  the vector of residuals, then



$$x = W f + e. \quad (2.8)$$

The covariance matrix  $\Sigma$  of  $x$  is then given by

$$\Sigma = W \Phi W' + s, \quad (2.9)$$

where  $\Phi$  is the covariance matrix of  $f$ , and  $s$  is the vector of unique variances of  $x$ . If the distributions of the residual  $e$  are assumed to be multivariate normal, then the conditional distribution of  $x$  will also be multivariate normal. In addition, if latent factors,  $f_s$ , follow multivariate normal distributions, then the marginal distributions of the manifest variables,  $x_s$ , can be derived and will also be multivariate normal. This marginalization process is usually employed to facilitate the estimation of latent variables given discrete data (Mislevy, 1986).

*Ordered Polychotomous Indicators*

Given ordered polychotomous indicators, the question arises as to whether or not it is appropriate to perform FA on a Pearson correlation matrix (FA-PR). Labovitz (1967, 1970) provided very influential justification for using Pearson correlations with ordinal-level variables. In his simulation studies, he demonstrated that "monotonic random scoring" systems, which indicated randomly stretched scales, are highly related with "equal distance scoring" systems. He argued that the errors due to treatment of ordinal variables as interval were small enough to be ignored. However, his studies were limited to the following conditions: a) The underlying latent variable was uniformly distributed; and b) As many as thirty-one categories of the ordinal variable were used. Given the situation where the underlying variable was normally or uniformly distributed, O'Brien (1979) found that the correlations between stretched scales and the equal distance scoring system were quite high and increased with the number of categories (C) when C was greater than four. His results were dramatically different

when the underlying continuous variable was quite skewed (log-normal): Pearson  $r$ 's were substantially smaller than those based on uniform and normal distributions and decreased as the number of categories increased. Bollen and Barb (1981) examined differences in the Pearson  $r$ 's computed on two normally distributed continuous variables compared to the same two variables equal-intervally collapsed into a few ordered categories. They found that these differences were basically small and that the greatest differences occurred when the continuous variables correlated highly and only a few (less than five) categories were used for collapsing. Generally speaking, the appropriateness of the use of Pearson's  $r$  with ordinal data seems to depend on a) the distributions of the two variables correlated; b) the number of the collapsed categories; and c) the equality of the intervals between collapsed categories.

Given discrete data, it is well known that the Pearsonian correlation is not free to range from  $-1$  to  $1$  when the two correlated variables are skewed highly in opposite directions (Carroll, 1961; Muthén,

1983). In other words, the Pearsonian correlation coefficient will generally underestimate the latent relationship between two normally distributed continuous variables which are categorized into two manifest variables with opposite skewness. In a simulation study, Olsson (1979a) indeed found that the maximum likelihood FA may create a substantial lack of fit of the true model when it was performed on the Pearsonian correlations computed from successive integers assigned to ordinal categories. His findings were especially true when the observed variables were skewed in opposite directions and the true loadings were high. He also found that the classification of continuous scores into categorical scores attenuated the estimates of factor loadings. The attenuation increased when the variation in skewness of manifest variables increased and the number of scale steps decreased. Given these results, Olsson suggested that researchers perform a FA on polychoric correlations (FA-PL) when observed variables were obtained from a classification of some continuous latent variables. This suggestion was also

made by some other researchers (e.g., Carroll, 1961; Muthén, 1983; Muthén, 1984).

The polychoric correlation coefficient is a generalization of the tetrachoric correlation coefficient to the polychoric case (Olsson, 1979b; Olsson, Drasgow & Dorans, 1982). It estimates the relationships between two latent continuous variables both of which are assumed to be normally distributed and are measured by ordinal scales. Various methods of estimating the population polychoric correlation can be traced back to Pearson (1913). For the tetrachoric case, the Maximum Likelihood (ML) estimation of the threshold values and the latent correlations is just-determined and simpler. For the polychoric case, the ML estimation is over-determined and time-consuming. Olsson (1979b) presented two ML estimation procedures for the polychoric case. The first procedure estimates the threshold values and the polychoric correlation simultaneously. The second procedure has two steps: a) to estimate threshold values with the inverse of the normal distribution function evaluated at the cumulative marginal proportions of each variable, and b) to obtain the

maximum likelihood (ML) estimate of the polychoric correlation, given the threshold values.

There are pros and cons for using polychoric correlations as the input to factor analysis. Polychoric correlations are not stable. The matrix of the polychoric correlations may not be positive definite so that maximum likelihood FA may not be applicable. The estimated standard deviations of the polychoric correlations in LISREL VI are inflated so that the chi-square test of the goodness-of-fit is incorrect. On the other hand, Jöreskog and Sörbom (1988) have shown in a simulation study that: a) polychoric correlations were not sensitive to the marginal distributions of the observed variables; b) compared to Spearman's rank correlations, Kendall's tau-b correlations, and product-moment correlations, polychoric correlations were the best estimators of the true latent relationships; and c) polychoric correlations appear to be the only consistent estimators of the true latent relationships. In another simulation study, Babakus, Ferguson and Jöreskog (1987) also found that, compared to product-moment, Spearman's rho, and Kendall's tau-b

correlations, polychoric correlations gave the most accurate estimates of the true latent correlations and factor loadings but produced the worst goodness-of-fit values. Given these results, the current study predicted that FA-PL should have a better performance than FA-PR in terms of the accuracy of estimating the underlying traits from multiple indicators.

#### Item Response Theory

Although the genesis of IRT can be traced back to the 1930s, the foundational work of IRT was done by Frederic M. Lord (1952; 1953a; 1953b). Birnbaum (1958, 1968, 1969) substituted the more mathematically tractable logistic models for the normal-ogive models developed by Lord and stimulated substantial progress in IRT. In addition, Rasch's (1960) independent work in Denmark also encouraged numerous studies of the one-parameter logistic model.

#### *Basic Theory*

Since the latent traits are not treated as random variables in IRT, Bartholomew (1987) considered this theory to be non-practical and

outside the mainstream of theoretical development. However, many psychometricians consider it to be "magic" (Thissen, 1986). It is magical because, given that the model fits the data, the calibration of items is independent of the ability distribution of those individuals who happen to be used for calibration, and the measurement of individuals is independent of the items that happen to be selected for measuring (Rasch, 1960; Wright, 1967). These two properties are known as "specific objectivity."

How do IRT models achieve the properties of objectivity? For simplicity, dichotomous items are assumed in the following illustration. Imagine a unidimensional latent variable as a straight line on which individuals can be differentiated with different trait levels and on which items can be differentiated with different scale (item difficulty) values. Georg Rasch modeled the interaction between the person and the item with the following function:

$$\pi_{mi} = \exp(\theta_m - b_i) / [1 + \exp(\theta_m - b_i)], \quad (2.10)$$



where  $\pi_{mi}$  is the probability of person  $m$  succeeding on item  $i$ ,  $\theta_m$  is the latent trait level of person  $m$ , and  $b_i$  is the scale/threshold/difficulty value of item  $i$ . Similarly, the probability of person  $n$  succeeding on item  $i$  is modeled as:

$$\pi_{ni} = \exp(\theta_n - b_i) / [1 + \exp(\theta_n - b_i)]. \quad (2.11)$$

The above two functions compare the latent trait level,  $\theta$ , to the scale value,  $b$ . When  $\theta = b$ , the probability of this person succeeding on this item is .50. As  $\theta > b$ , the probability of success is greater than .50 and approaches 1.0. As  $\theta < b$ , the probability of success is less than .50 and approaches 0. This is a probability model for item response processes.

Given one item  $i$ , if persons  $m$  and  $n$  both succeed or both fail on this item, then no information is available to differentiate  $m$  from  $n$  on the latent trait. However, if one of them succeeds and the other fails, then information is available. The probability of  $m$  succeeding but  $n$  failing on item  $i$  ( $\pi_{10i}$ ) is

$$\pi_{10i} = \pi_{mi}(1 - \pi_{ni}) = \exp(\theta_m - b_i)/K, \quad (2.12)$$

where  $K = [1 + \exp(\theta_m - b_i)][1 + \exp(\theta_n - b_i)]$ .  
Similarly, the probability of m failing but n succeeding on item i ( $\pi_{01i}$ ) is

$$\pi_{01i} = (1 - \pi_{mi})\pi_{ni} = \exp(\theta_n - b_i)/K. \quad (2.13)$$

The "magic" occurs when (2.12) and (2.13) are combined to produce the following conditional probability:

$$\frac{\pi_{10i}}{\pi_{01i} + \pi_{10i}} = \frac{\exp(\theta_m - \theta_n)}{1 + \exp(\theta_m - \theta_n)} \quad (2.14)$$

The above function can be rewritten as

$$\theta_m - \theta_n = \ln(\pi_{10i}/\pi_{01i}), \quad (2.15)$$

where  $\pi_{10i}$  can be estimated with the number of items answered correctly by person m but failed by person n, while  $\pi_{01i}$  can be estimated with the number of

items which are failed by person  $m$  but answered correctly by person  $n$  (Masters, 1988). Note that in (2.14) and (2.15) the item parameter ( $b_i$ ) disappeared. It implies that the estimation of the difference between the two person parameters ( $\theta_m, \theta_n$ ) does not depend on the item parameter ( $b_i$ ). As a consequence, this result permits individuals to be scored on the same scale, even though they do not respond to the same set of items.

By the same logic, it can be shown that the estimation of the difference between the scale values of item  $i$  and  $j$  does not depend on person parameters. The conditional probability of person  $n$  answering item  $i$  correctly but failing item  $j$  ( $\pi_{n10}$ ), given that he/she answers only one of the items correctly, is

$$\frac{\pi_{n10}}{\pi_{n01} + \pi_{n10}} = \frac{\exp(b_i - b_j)}{1 + \exp(b_i - b_j)}, \quad (2.16)$$

which can be rewritten as

$$b_i - b_j = \ln(\pi_{n01}/\pi_{n10}), \quad (2.17)$$

where  $\pi_{n01}$  can be estimated with the number of persons who fail item  $i$  but answer item  $j$  correctly, while  $\pi_{n10}$  can be estimated with the number of persons who answer item  $i$  correctly but fail item  $j$ .

#### *Ordered Polychotomous Items*

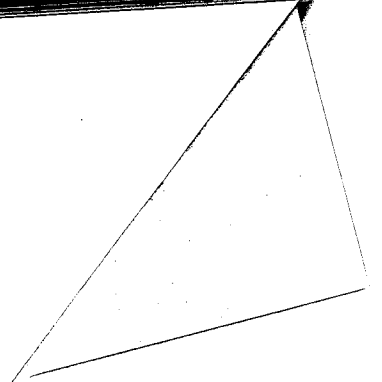
When the IRT models for dichotomous data are extended to polychotomous data, there are at least three models available: a) the partial credit model (Masters, 1982); b) the rating scale model (Andrich, 1978); and c) the graded response model (Samejima, 1969). In the following discussion, the term "threshold" will refer to a general concept of difficulty in the polychotomous items because it has been widely employed under the contexts of FA-P, IRT, and even latent class analysis. Thresholds defined by the partial credit model and the rating scale model will be called "steps," and by the graded response model, "category boundaries."

The partial credit model is simply the Rasch model applied to polychotomous items. Therefore, all items are assumed to have the same discrimination parameter, 1.0. In addition, the step difficulty

values in an item are defined locally by the probability of responding in either of two adjacent categories. In other words, the step values correspond to the intersections of adjacent probability curves. Consequently, the step values for an item are not necessarily ordered, although the response categories must be ordered. This model implies that, given an item, it can be more difficult moving from response category 1 to category 2 than from category 2 to category 3. The partial credit model is not appropriate for the current simulation study because the items will be designed to have different discrimination parameters.

The rating scale model assumes that the functioning of the response categories (e.g., *never/sometimes/often/always*) is the same across items, although each item may have a different scale value (location along the trait continuum). Given a set of polychotomous items with response categories 0, 1, ..., m, this model incorporates the following constraint into the partial credit model:

$$b_{ik} = b_i + \delta_k, \quad (k = 1, \dots, m), \quad (2.18)$$



where  $b_{ik}$  is the step value for item  $i$  at scale step  $k$ ,  $b_i$  is the scale value for item  $i$  and is usually defined as the mean of the step values ( $b_{ik:s}$ ), and  $\delta_k$  is the distance between each step value and the scale value. It can be seen that, in this model, scale values ( $b_i:s$ ) can vary across items while  $\delta_k$ 's are kept constant across items. When compared to the partial credit and the graded response model, the rating scale model has the advantage of simplicity because it involves fewer parameters. This model is suitable when the interaction between response categories and items does not occur. However, it is not suitable for the data of the present study because: a) the present simulation allows the distances between step values and the scale value to vary across items; and b) the present simulation allows items to have different discrimination parameters.

Given the simulated situation in the present study, the most appropriate model from IRT may be the graded response model (GRM) (Samejima, 1969). Samejima developed a two-stage procedure to derive

the probability of an individual selecting a particular response category in an polychotomous item. In the first stage, an item with response category 0, 1, ..., m was viewed as a combination of m dichotomous items and the two-parameter model was applied to model the cumulative probability of an individual responding to a particular or higher category. This idea is expressed by the following equation:

$$\sum_{j=k}^m \pi_{nij} = \frac{\exp[a_i(\theta_n - b_{ik})]}{1 + \exp[a_i(\theta_n - b_{ik})]}, \quad k = 1, \dots, m \quad (2.19)$$

where  $\pi_{nij}$  is the probability of person n responding to item i with response category j or higher,  $a_i$  is the discrimination parameter for item i,  $b_{ik}$  is the category boundary between response category k and k-1 of item i, and  $\theta_n$  is the latent trait level of individual n. Note that the probability of responding to category 0 or higher covers the complete probability space and was set to be one, that is,

$$\sum_{j=0}^m \pi_{nij} = 1. \quad (2.20)$$

In the second stage of Samejima's procedure, the probability of an individual responding to a particular response category  $j$  is given by:

$$\pi_{nij} = \sum_{j=k}^m \pi_{nij} - \sum_{j=k+1}^m \pi_{nij}, \quad k = 0, \dots, m. \quad (2.21)$$

The above equation is the general form for drawing the operating characteristic curves for a graded response item. From (2.18), (2.19) and (2.20), it can be derived that the probability of responding to the first and the last category is given by:

$$\pi_{ni0} = 1 - \sum_{j=1}^m \pi_{nij}, \quad (2.22)$$

$$\text{and } \pi_{nim} = \frac{\exp[a_i(\theta_n - b_{im})]}{1 + \exp[a_i(\theta_n - b_{im})]}, \quad (2.23)$$

respectively. Equation (2.23) is equivalent to equation (2.19) when  $k=m$ .



The GRM allows a parameter for each item to have a different discrimination power. In addition, it allows the distance between response scale thresholds to vary across items. Certainly, the cost of this flexibility is the increased number of item parameters to be estimated and, therefore, the accompanying estimation problems and computer time needed. This model also differs from the partial credit model in its definition of the threshold values. This model defines threshold values globally in terms of cumulative probabilities of every response category within an item. For example, given an item with 4 response categories, the threshold values are defined in the way shown in Table II-2. A consequence of this definition of category boundaries is that the threshold values are, by definition, ordered. This consequence is, however, consistent with the assumption of the simulation process used in the present study.

The applicability of the graded response model to Likert-type scales has been demonstrated by several studies (e.g., Dodd, 1984; Dodd, Koch & DeAyala, 1988; Koch, 1983a; Thissen & Steinberg,

1988). This model is much closer to the FA approaches than is either the partial credit or the rating scale model. The relationships between the FA and IRT models will be explored further in the last section of this chapter.

Table II-2

Threshold Values Defined in the Graded Response Model

Threshold Value	Location
$t_{i1}$	$\pi_{ni0} = \pi_{ni1} + \pi_{ni2} + \pi_{ni3}$
$t_{i2}$	$\pi_{ni0} + \pi_{ni1} = \pi_{ni2} + \pi_{ni3}$
$t_{i3}$	$\pi_{ni0} + \pi_{ni1} + \pi_{ni2} = \pi_{ni3}$

#### Multidimensional Scaling

According to Young (1987), the first decade of MDS was initiated by Torgerson's (1952, 1958) metric MDS, following which Shepard's (1962) and Kruskal's (1964) nonmetric MDS methods opened the second decade. The beginning of the third decade was due to the development of individual differences or weighted MDS methods (Carroll & Chang, 1970). Currently,