workers in this field have also been developing maximum likelihood MDS (e.g., Ramsay, 1982; Takane, 1980a, 1980b).

Basic Theory

MDS includes a set of spatial distance models. The basic idea of MDS is to represent the observed proximities among a set of objects in terms of latent distances embedded in a spatial configuration. A variety of indices such as correlations, judgments of similarity/disimilarity, or frequencies of co-occurrences of paired objects could serve as proximity measures. On the other hand, the word *distance* could refer to all kinds of Minkowski distances which can be expressed in the following function:

$$d_{ij} = (\sum_{k} |x_{ik} - x_{jk}|^{p})^{1/p}, \qquad (2.24)$$

where x_{ik} is the coordinate of stimulus x_i on dimension k and x_{jk} is the coordinate of stimulus x_j on dimension k. If p=2, then the above function

reduces to the Euclidean distance function. If p=1, then it reduces to the city-block metric.

Torgerson (1952) proposed one of the first algorithms for metric MDS. His basic assumption was:

$$\delta_{ij} \approx d_{ij} = [\sum_{k}(x_{ik} - x_{jk})^2]^{1/2}. \qquad (2.25)$$

This model attempts to represent the observed amount of proximity (δ_{ij}) between object x_i and x_j in terms of the Euclidean distance (d_{ij}) between points representing these two objects in a k-dimensional space. Later developments of MDS employ a weaker assumption, i.e. they aim to recover the order of, instead of the amount of, the proximity. Therefore, the basic model of non-metric MDS is as follows:

$$\delta_{ij} \approx f(d_{ij}) = f\{[\sum_{k}(x_{ik} - x_{jk})^2]^{1/2}\}, \qquad (2.26)$$

where f is a monotone function such that

$$d_{ij} < d_{i'j'} \Rightarrow f(d_{ij}) < f(d_{i'j'}). \qquad (2.27)$$

The basic notion underlying nonmetric MDS is that sufficient nonmetric constraints can act like metric constraints (Shepard, 1962). Given a dimensionality and a sufficient number of points, the solution configuration that complies with the set of inequalities becomes highly constrained (MacCallum, 1974). Therefore, it is possible to recover the metric information from nonmetric data (Young, 1970).

Ordered Preference Data

The distance models for preference data are often called unfolding models. Usually, unfolding models attempt to model the proximities between the subjects and the stimuli (e.g., items) instead of the proximities between the pairs of stimuli. Given one rectangular matrix of preference data, two classical unfolding models might be applicable: Coombs' (1964) internal unfolding (IMDU) and Carroll's (1972) external unfolding (EMDU) procedures. Coombs' model has the following form:

$$\delta_{is} \approx f_s(d_{is}) = f_s\{[\Sigma_k(x_{ik} - x_{sk})^2]^{1/2}\}, \qquad (2.28)$$

where δ_{is} refers to the degree of subject s' preference for stimulus i, f_s is the monotonic function specific to subject s, d_{is} is the estimated distance between the stimulus i and the subject s in the model, x_{ik} represents the location of stimulus i along dimension k, and x_{sk} represents the ideal point of subject s along dimension k. Evidently, this is a "point model" instead of a "vector model." It is assumed that the more the subject's ideal point is close to the location of the stimulus' scale value, the higher the subject's preference will be, regardless of whether the ideal point is at the positive or the negative side of the location of the scale value on the dimension.

Carroll's (1972) EMDU models have exactly the same assumptions except that the stimulus/item coordinates are assumed to be known from the theory or from a previous estimation. The EMDU procedures include the following steps: a) calculating dissimilarities among all pairs of items using squared Euclidean distances; b) subjecting the matrix of dissimilarity data to the classical MDS procedure for estimating item coordinates; and c) fixing the

item coordinates and obtaining estimates of person parameters with unfolding models. These procedures are called "successive unfolding" by Rodgers and Young (1981). It should be obvious that EMDU, like FA and IRT models, employs a two-stage procedure--namely, it first estimates the item parameters and then estimates the person parameters assuming that the item parameters are known.

The two classical MDU models may be applicable to the test data where the trace line of each item is a cumulative function, i.e., the relationships between the latent trait and responses to each item are monotonic. Some MDS theorists have argued that monotonic relationships between vectors are able to be modeled by positioning subjects' ideal points far away from the location of the items' ideal points in the direction of the latent dimension (Schiffman, Reynolds & Young, 1981, p.262). In other words, clusters of persons and items may be formed separately in opposite directions of the solution space. In this situation, the solution may look "degenerate" but the ideal points of the subjects and items may still be estimated appropriately.

In a simulation study, Fitzpatrick (1989) generated data according to the Rasch model, which assumes a monotonic relationship between the latent trait and the probability of a correct response to an item. Evidently, MDU analysis of this kind of data tends to position all ideal points of subjects at one side and all locations of items at the other side of the solution configuration. Therefore, Fitzpatrick found that the MDU procedure often produced what many researchers call "degenerate" solutions with high stress, but the resulting person and item coordinates still correlated highly with the true parameters. If a solution can recover the person and item parameters appropriately, as Fitzpatrick noted, then it should not be counted as a "degenerate" one. At least two criteria are necessary for judging an MDU solution to be degenerate: a) ideal points of subjects and items are positioned into two distinct sets; and b) "the number of distinct points in the solution configuration is small compared to the number of stimulus points" (Davison, 1983, p. 100). When the algorithm cannot differentiate the stimulus points

very well, the correlation between true and estimated ideal points cannot be high.

Typical degenerate solutions may occur when the MDU algorithm concentrates on recovering the rank order of individual subject's preferences for stimuli but tends to find a low-stress solution by making few distinctions among subjects within one set and few distinctions among items within the other set (Koch, 1984). This kind of solution is frequently obtained from row-conditional data where only the information about the proximities between the individual subject and each item is utilized while the information about the inter-subject and the inter-item proximities is ignored.

Fitzpatrick's (1989) findings were restricted to the situation where dichotomous items with equal discrimination parameters were used. Dichotomous items might be adequately represented by points because only one threshold value is involved for each item. In addition, the assumption of equal discrimination of items fits the classical MDU approaches, which explicitly model the differences

between difficulty parameters rather than discrimination parameters.

It is not clear how classical MDU approaches can model the differences among item difficulty levels and among item discrimination power simultaneously. For Likert-type data, it may be especially inappropriate for the two classical MDU approaches to analyze these polychotomous items having different discrimination parameters (or factor loadings). Every Likert-type item with n response categories can be decomposed into n-1 dichotomous items which attempt to measure a wide range of the latent trait. The two classical MDU models simply compress each Likert item into one geographical point and bypass the estimation of threshold values and of discrimination parameters. They estimate only one latent scale value (the item coordinate) for each item. Because each item coordinate represents the mean threshold value of each item, the two classical unfolding models should completely fail in the situation where all items have normal distributions and the same mean threshold values, which is a rather desirable situation for the FA and the graded

response models. If all item distributions are highly skewed and truncated, a large proportion of respondents should fall into few item response categories and very limited information can be gained through any statistical models including FA, IRT, or MDU models. Only in the situation where every item distribution has a different skewness value but none is highly skewed might the classical MDU models be applicable to the Likert-type data.

In theory, Coombs (1964) and Davison (1983) tended to argue against, while Coombs and Smith (1973), Lingoes (1972), and Young and Lewyckyj (1979) tended to argue for, the application of the classical MDU models to profile data. In the published literature, few applied studies can be found. In an empirical study where the internal MDU model was applied to two Likert format instruments, Koch (1984) found that degenerate solutions were paramount regardless of whether the ALSCAL (Young & Lewyckyj, 1979) or the MINISSA (Roskam & Lingoes, 1970) computer program was used, whether the row conditionality or the matrix conditionality assumption was made, and whatever number of

dimensions was used. Therefore, it seems desirable to search for a better alternative to the classical MDU models for the analysis of Likert-type data.

It is clear that both FA with polychoric correlations and the graded response model in IRT are able to estimate not only the item threshold values but also the relationship between each item and the latent dimension. Therefore, one promising direction for the MDU models to analyze Likert-type data is to regard each response category of each item as one stimulus and to take the relationship between each item and the latent dimension as the weight of each item on the latent dimension.

First of all, the assumption that the response categories of each item are in order must hold. Consequently, if a person i chooses response category j and j is the lowest category, then his/her preference about the response categories can be inferred to be: j, j+1, j+2, ..., to the highest category. If j is the highest category, then his/her preference order should be: j, j-1, j-2, ..., to the lowest category. If j is the middle category, then his/her preference order is: j, j±1, j±2, ..., to the

lowest or the highest category. If this data transformation were done for each item and each person, then a "three-way three-mode" data set would be obtained. This data set would contain multiple matrices of rectangular data with rows corresponding to subjects, columns corresponding to response categories, and matrices corresponding to items. If the weighted MDU model (WMDU) (Young & Lewyckyj, 1979; Young, 1984) is applied to this kind of data, coordinates of subjects and of item response categories and the weight for each item can be estimated for the latent dimension. Although only one set of coordinates of response categories is estimated for all items, this model preserves each Likert-type item as a vector and, therefore, is able to model the relationship between each item and the latent dimension.

The WMDU model can be viewed as the individual differences scaling model (Carroll & Chang, 1970) extended to preference data. For simplicity of expression, the model is shown with matrix algebra as follows:

$$\delta_{ijk} \approx f_i(d_{ijk})$$
$$= f_i\{[(\underline{y}_i - \underline{x}_j)'W_k(\underline{y}_i - \underline{x}_j)]^{1/2}\}, \qquad (2.29)$$

where δ_{ijk} is the proximity between row i and column j for matrix k, f_i is the monotonic function for row i, d_{ijk} is the estimated distance corresponding to δ_{ijk}, $\underline{y}_i$ is the vector for row i, $\underline{x}_j$ is the vector for column j, and the diagonal W_k is the weight for matrix k. In the current situation, subscripts i, j, and k correspond to subjects, response categories, and items. It can be seen that the weight for "individual difference" is applied to model the differences in item discriminations or factor loadings on the latent dimension.

In the MDS literature, few studies have performed any kind of MDU procedures on Likert-type data. Therefore, all three unfolding procedures including IMDU, EMDU, and WMDU were investigated by the current study in order to determine whether or not the WMDU model might be a viable alternative to the classical MDU models for analyzing Likert-type data. It was predicted that the WMDU model should outperform the two classical unfolding procedures in

most conditions except one where the response categories functioned very differently across items, i.e., the same category boundary had very different values across items.

Comparative Analysis

FA and IRT

In this section, it will be shown that the two-parameter normal ogive model in IRT is a nonlinear transformation of the Spearman single factor model. For simplicity of illustration, the dichotomous case and unidimensionality will be considered first. Let θ represent the latent variable and y_i the continuous response to item i. Then FA model states that

$$y_i = w_i \theta + e_i, \qquad (2.30)$$

where w_i is the correlation between y_i and θ, while e_i is the residual or error vector. However, y_i can only be measured dichotomously so that

$$x_i = 1, \text{ if } y_i > t_i,$$
$$x_i = 0, \text{ otherwise,} \qquad (2.31)$$

where t_i is the threshold value for item i.

It is possible to fit the common FA model (2.30) to the data without any distributional assumptions about θ and e_i (Bartholomew, 1987). The FA and IRT models may be connected, however, if e_i is assumed to be normally distributed. Under the normality assumption, it follows that the conditional distribution of y_i, given a fixed value of θ, is also normally distributed. If guessing is not a factor here, then the conditional probability of successfully passing item i is given by the cumulative probability from negative infinity to the threshold value under the conditional distribution of y_i (Bejar, 1983). That is,

$$P(x_i=1|\theta) = \int_{-\infty}^{t_i} \frac{1}{\sigma_i \sqrt{(2\pi)}} e^{(y_i-x_i)^2/2\sigma_i^2} dy_i, \qquad (2.32)$$

where $\sigma_i = \sqrt{1-w_i^2}$ is the standard deviation of the

error vector e_i. When this conditional probability is computed for every point of θ and plotted against θ, we obtain the item characteristic curve (ICC) of the two-parameter normal-ogive model (Lord, 1952). This ICC takes the following form:

$$P(x_i=1|\theta) = \int_{-\infty}^{a_i(\theta-b_i)} \frac{1}{\sqrt{(2\pi)}} e^{-z^2/2}\, dz, \qquad (2.33)$$

where a_i is the discrimination parameter and b_i is the difficulty (threshold) parameter. The two-parameter normal-ogive model and the common FA model are mathematically connected by the following equations:

$$a_i = w_i/\sigma_i, \qquad (2.34)$$

and $$b_i = t_i/w_i. \qquad (2.35)$$

These relationships have been noted by Lord and Novick (1968) and Bejar (1983), among others. These connections are readily extendable to the ordered categorical case because the IRT formulation for this

case usually decomposes the polychotomous item with n response categories into n-1 dichotomous items. The mathematical connections between common FA models for ordered categorical data and Samejima's (1969) normal ogive model for graded response data have been shown by Takane and de Leeuw (1987) to be as follows:

$$a_i = w_i/\sigma_i, \tag{2.36}$$

and $$b_{i(j)} = -t_{i(j)}/\sigma_i, \tag{2.37}$$

where the additional subscript letter j represents the jth response category of item i. Although the normal ogive function is usually replaced by the logistic function in practice, the above relationships between FA and IRT should be approximated very well because the logistic function approximates the normal ogive functions very well, especially when a scaling constant is used.

From equation (2.33), it seems clear that the two-parameter normal ogive model is a nonlinear transformation of the Spearman single factor model performed on the observed score x_i (McDonald, 1982,

1985). When the distribution of scores on an item is highly skewed, the relationship between the item and the latent trait is indeed nonlinear and may be better described by IRT models than by FA models.

The matrix of second-order cross-products analyzed by FA is a sufficient statistic for the multivariate normal distribution when all latent and manifest variables are approximately normally distributed (Mooijaart, 1985). In this situation, the relative merits between the limited and the full information approaches of estimation may disappear. The limited-information approach utilizes only the information in the lower order margins in the full k^p (p variables with k response categories) contingency table while the full-information approach utilizes all information in the table (Mislevy, 1986). FA with either the two-stage maximum-likelihood estimators of polychoric correlations or Pearsonian correlations is a limited information approach, but the marginal maximum likelihood (MML) estimation for the graded response model is a full information approach. Therefore, the performance of the IRT-GRM is predicted to be as good as both FA-PR and FA-PL

procedures in the normal-distribution situation but better than these two FA procedures in the skewed-distribution situation.

The above predictions are also made based on the consideration that the current study investigates only the unidimensional case. In discussing the MML estimation method based on the full information approach versus the GLS (Generalized Least Square procedure) based on limited information approach, Mislevy (1986) pointed out that the former is preferable for long tests with few factors. MML is better with few factors because it requires integration over the entire factor space, which implies geometric increases in computation load as the number of factors increases. For dichotomous items and in terms of ability to recover item parameters, Knol and Berger's (1989) findings were indeed consistent with this line of argument. Knol and Berger found in a simulation study that MML estimation for IRT models, implemented in the computer program TESTFACT (Wilson, Wood & Gibbons, 1984), performed slightly better than FA-PL with various estimation procedures in the unidimensional

case. FA-PL and IRT performed about equally well given two dimensions. The FA procedures outperformed the IRT models given more than two dimensions.

In conclusion, FA and IRT should perform equally well when item responses are normally distributed, where the correlation matrix is a sufficient statistic for the multivariate normal distribution. IRT should outperform both FA-PR and FA-PL when distributions of item responses are skewed, where the relationship between manifest responses and the latent trait became nonlinear and should be better modeled by IRT's nonlinear formulation.

FA and MDS

Both FA and MDS are engaged in "data reduction", and both are able to yield a dimensional representation of the data structure. However, they are distinct from each other in many aspects. First, FA favors correlation coefficients or covariances while MDS accepts many proximity measures, such as direct judgments of similarity, correlations, or joint probability values. Second, FA customarily

derives correlations or covariances from direct response data while MDS accepts direct similarity judgments as well as indirectly derived proximity measures. Third, the coordinates estimated by FA are *multiplied* together and summed over dimensions to yield expected correlations while the coordinates estimated by MDS are *subtracted* from each other and transformed by a (Euclidean) distance function to yield expected distances. Fourth, the FA solution space is basically not for pictorial purposes but for reproducing the underlying relationships of a set of variables with respect to a sample, while the MDS solution space is for pictorial purposes because the interpoint distances correspond to the perceived relationships (MacCallum, 1974). Young (1987) also stated that "Multidimensional scaling (MDS) rests on the premise that a picture is worth a thousand numbers (p.3)." Finally, the interpretation of the FA solution is usually dimensional, while the interpretation of the MDS solution can be dimensional, regional, or in terms of facets.

MDS may be closer to principal components analysis (PC) than to common FA. For example,

MacCallum (1974) pointed out that there is no concept in MDS adequately corresponding to the communality concept in the general FA model. He analyzed the relationships of MDS and PC and concluded that the PC model was much richer than the classical metric MDS model because the former is able to describe individual differences in terms of component scores but the latter is not. Nonetheless, this conclusion does not hold for individual differences MDS and MDU models. Rodgers and Young (1981) claimed that applying metric MDS to a symmetric matrix of Euclidean distances computed from a rectangular matrix of preference data is equivalent to performing a PC on the preference data. Davison (1985) argued that, when both MDS and PC were performed on the same correlation matrix, MDS implicitly subtracted the standardized person mean from the solution while PC needed an additional component to represent the differences among person means, so that the first general component from PC would not appear in the MDS solution. If all person means were approximately equal, then the PC and the MDS solutions would yield the same number of dimensions. Davison's (1985)

predictions were supported by his own data, and they were found to be extendable to the relationships between common FA and MDS (Silverstein, 1987). Sometimes, when the first general component reflects a certain response set, such as acquiescence, which might not be interesting to the researcher, MDS can give a more parsimonious solution (Davison & Srichantra, 1988).

Although some researchers (e.g., Davison & Srichantra, 1988) regard PC as a special case of "factor analysis," it should be noted that PC and FA are different from each other in many aspects (Jöreskog & Sörbom, 1979). First, PC begins with the definition of the component as an observed combination of the manifest variables, whereas FA begins with the definition of the factor as the latent common predictor of the manifest variables. Thereafter, FA utilizes the principle of local independence, but PC does not. Second, PC aims to reproduce the total variance of the manifest variables, whereas FA aims to reproduce their intercorrelations (or common variance). Third, although PC can extract a large proportion of total

variance with a few components, it needs all components to reproduce the correlations exactly. On the other hand, the number of factors needed by FA to reproduce the correlations exactly can be substantially fewer than the number of the manifest variables (Jöreskog & Sörbom, 1979). Therefore, the relationships between MDS and PC are not necessarily applicable to the relationships between MDS and FA.

Most empirical studies which have compared MDS with PC/FA models typically have concentrated on the number of dimensions, the congruence of dimensions, and the interpretability of the constructs that emerged from the analyses. For example, Schlesinger & Guttman (1969) performed MDS on correlations of intelligence tests and compared their results to those of a previous study which used FA. They found that, combined with facet theory, MDS' results were more meaningful than FA's. Using correlation coefficients computed from marketing survey data, Lehmann (1974) made a similar comparison and concluded that MDS and FA gave comparable results. He argued that it was not the model but the similarity measures that affected the solutions. Levin, Montag &

Comrey (1983) performed FA, PC and MDS on a correlation matrix computed from personality inventories and found that four factors/components from FA and PC corresponded to four regions formed in a two-dimensional space obtained from MDS. Using Euclidean distances for MDS input and correlations for FA input, Koch (1984) concluded that both FA and MDS performed very well in terms of identifying meaningful underlying structures and inappropriate items. Dancer (1985) applied MDS to a correlation matrix of the Rosenberg Self-Esteem scale and compared her results to the previous results of a factor analysis. She found that MDS gave more dimensions than FA did and that not much correspondence existed between the MDS and the FA results. In a longitudinal study, Gillespie (1989) found that, in terms of interpretability and stability across time, FA performed better than MDS. She also noticed that item response distributions might affect both FA and MDS results.

Most of the reviewed studies ignored MDU and individual differences MDS and, therefore, ignored the person parameters. With a few exceptions, MDS

gave fewer dimensions than did FA/PC. However, researchers frequently employed facet interpretation for MDS but dimensional interpretation for FA/PC. Their results were not very consistent with one another because many issues could affect their conclusions. These intervening issues included number of stimuli/variables, distributions of responses, kinds of similarity measures, decision-making about the number of dimensions, and interpretation in terms of dimensions, facets, or regions. Some of these intervening issues can be manipulated systematically and investigated with Monte Carlo studies. In the current study, number of items, number of subjects, and distributions of item responses were manipulated. In addition, person parameters were the focused parameters.

To the knowledge of the author there existed no empirical study comparing FA to the WMDU model. Based on the facts that the WMDU model estimates one set of response-category coordinates for all items while FA-PL estimates different threshold values for different items, it was predicted that the WMDU model should perform as well as the FA-PL when response

categories had similar functions across items (e.g., when every item had an approximately normal distribution of responses). The WMDU model should perform worse than FA-PL when response categories functioned differentially across items (e.g., when every item had a differentially skewed distribution of responses). However, the relative performance of FA-PR and WMDU was treated as an open empirical question. Finally, it was predicted that both FA-PR and FA-PL should outperform IMDU and EMDU because the two classical unfolding models estimate neither threshold values nor discrimination parameters of items.

IRT and MDS

According to Andrich (1988, 1989), Thurstone's (1927, 1931) Law of Comparative Judgment involved two response mechanisms, i.e., the cumulative and unfolding mechanisms. In the cumulative mechanism, subjects were asked to compare two statements irrespective of their own attitudes. Therefore, the stronger a statement was than the other statement, the more likely it was to be chosen. In the unfolding

mechanism, subjects were asked to compare their own attitudes to the statements, so that the closer the statement was to the location of the subject's attitude on the continuum, the more likely it was to be endorsed by the subject. Therefore, an item with the cumulative mechanism would have a monotonically increasing trace line, while an item with the unfolding mechanism would form a single-peaked trace line. With this differentiation and the analogy between the paired comparison task and the dichotomous item, Andrich (1988, 1989) was able to formulate a logistic model closely related to the Rasch model for the unfolding data. Because the Rasch model implies a cumulative mechanism, Andrich's model and the Rasch model have an identical form. However, in Andrich's model the probability of positive endorsement depends on the square of the distance between the person and the item locations whereas in the Rasch model it depends on the directed distance. In practice, it is possible that an unknown response mechanism underlies some attitude items with "agree" vs. "disagree" response categories. Then, a test of the goodness of fit of the two models may

help clarify the situation. Unfortunately, the connection between the field of unfolding models and of IRT models does not go beyond the case of dichotomous items with equal discrimination power.

It is interesting to see that both cumulative and unfolding mechanisms are involved in Likert-type data. It is clear that the relationship between the item and the latent trait implies a cumulative mechanism. It may be less clear that the relationship between the response category and the latent trait contains both cumulative and unfolding mechanisms. On the one hand, the response mechanism for the two extreme categories is a cumulative one. On the other hand, the response mechanism for the middle categories is an unfolding one. This mixed mechanism is modeled by IRT through taking the difference between the adjacent cumulative probabilities. It can be seen that, in the operating characteristic curves for the graded response model, the partial credit model and the rating scale model, the probability of responding to either of the extreme categories is a monotonic function, while the probability of

responding to any of the middle categories is a single peaked function.

The current study predicted that the WMDU model might not perform as well as the IRT-GRM because it did not differentiate the two extreme categories from the middle categories. An additional reason for this prediction was that WMDU gave only one set of estimates of threshold values for all items, while the IRT-GRM gave different sets of estimates of threshold values for different items. These relative advantages of IRT-GRM to WMDU may not be evident when all items had approximately normal distributions of responses, where response categories had similar functions across items.

Since the two classical unfolding models, IMDU and EMDU, were predicted to perform worse than WMDU, it follows that these two models should also perform worse than the IRT-GRM in most conditions.

CHAPTER III

STATEMENT OF PROBLEM

Problem 1: In terms of the ability to recover the true latent trait parameters from ordered Likert-type data, what would be the relative merits of FA-PL and FA-PR? How would their relative performances depend on test lengths, sample sizes, and distributions of item responses?

Problem 2: In terms of the ability to recover the true latent trait parameters from ordered Likert-type data, what would be the relative merits of WMDU, IMDU, and EMDU? How would their relative performances depend on test lengths, sample sizes, and distributions of item responses?

Problem 3: In terms of the ability to recover the true latent trait parameters from ordered Likert-type data, what would be the relative merits of the three latent variable approaches (the FA approach, the IRT approach, and the MDS approach)? How would their

relative performances depend on test lengths, sample sizes, and distributions of item responses?

Problem 4: In terms of the ability to recover the true latent trait parameters from ordered Likert-type data, would the three latent variable approaches perform better than the common SSI practice? How would their relative performances depend on test lengths, sample sizes, and distributions of item responses?

CHAPTER IV

METHODOLOGY

The Simulated Situation

For each item, a continuous response variable (Y^*) was generated according to the following formula:

$$Y^* = \beta X + \sqrt{(1-\beta^2)}\, E, \qquad (3.1)$$

where X was the set of standardized Z-scores on the latent variable, β was a regression weight conceptually corresponding to the relationship between the item and the latent variable, and E was the error or residual vector. Both X and E were generated according to the normal distribution and had a mean of zero and a variance of one. Therefore, Y^* also had a mean of zero and variance of one because

$$\text{Mean}(Y^*) = \beta * 0 + \sqrt{(1-\beta^2)} * 0 = 0,$$

and $\text{Var}(Y^*) = \beta^2 * \text{Var}(X) + (1-\beta^2) * \text{Var}(E) = 1.$

The values of β were systematically chosen and randomly assigned to each item because each item was

systematically assigned with a skewness value (see Table IV-1 to IV-4).

Secondly, the integer response score (Y) for each simulated subject was decided according to the following scheme:

$$
\begin{aligned}
&Y = 1, \text{ if } \quad Y^* < t_1 \\
&Y = 2, \text{ if } t_1 \leq Y^* < t_2 \\
&Y = 3, \text{ if } t_2 \leq Y^* < t_3 \\
&Y = 4, \text{ if } t_3 \leq Y^* < t_4 \\
&Y = 5, \text{ if } t_4 \leq Y^*.
\end{aligned}
\tag{3.2}
$$

Note that t1, t2, t3, and t4 were threshold values (to be described in a later section), which were manipulated in order to obtain desired distributional characteristics of item responses.

Background Conditions

The following conditions were held constant across experimental conditions: 1) the observed variableswere ordinal and had five discrete response categories; 2) there was only one latent dimension

Table IV-1

Distributional Characteristics of the 12 Core Items in the Condition of Normal Distribution

	Item#					
	1	2	3	4	5	6
t_1	-2.076	-2.052	-1.836	-1.812	-1.740	-1.692
t_2	-.716	-.712	-.676	-.672	-.660	-.652
t_3	.716	.712	.676	.672	.660	.652
t_4	2.076	2.052	1.836	1.812	1.740	1.692
C_1	.018	.019	.031	.033	.039	.043
C_2	.218	.218	.217	.216	.214	.213
C_3	.528	.526	.503	.501	.493	.488
C_4	.218	.218	.217	.216	.214	.213
C_5	.018	.019	.031	.033	.039	.043
SK	.000	.000	.011	.011	.010	.000
KT	.008	.020	-.013	-.006	-.012	.039
β	.55	.80	.35	.75	.50	.45

(continued on next page)

Note: t1 - t4 are threshold values; C1 - C5 are percentages of population in each response category; SK=Skewness; KT=Kurtosis; β=Factor loadings.

Table IV-1 (Continued)

	Item#					
	7	8	9	10	11	12
t_1	-1.668	-1.620	-1.596	-1.548	-1.500	-1.476
t_2	-.648	-.640	-.636	-.628	-.620	-.616
t_3	.648	.640	.636	.628	.620	.616
t_4	1.668	1.620	1.596	1.548	1.500	1.476
C_1	.045	.050	.053	.058	.064	.067
C_2	.212	.210	.208	.206	.203	.201
C_3	.486	.480	.478	.473	.467	.465
C_4	.212	.210	.208	.206	.203	.201
C_5	.045	.050	.053	.058	.064	.067
SK	.000	.000	.000	-.010	-.010	-.010
KT	.033	.004	-.007	.017	-.030	-.050
β	.90	.65	.40	.85	.70	.60

Note: t1 - t4 are threshold values; C1 - C5 are percentages of population in each response category; SK=Skewness; KT=Kurtosis; β=Factor loadings.

Table IV-2

Distributional Characteristics of the 12 Core Items in the Condition of Moderately Skewed Distribution

	Item#					
	1	2	3	4	5	6
t_1	-1.645	-1.645	-1.645	-1.645	-1.645	-1.645
t_2	-.862	-.942	-1.002	-1.062	-1.122	-1.142
t_3	-.331	-.415	-.478	-.541	-.604	-.625
t_4	1.100	.900	.750	.600	.450	.400
C_1	.050	.050	.050	.050	.050	.050
C_2	.144	.123	.108	.094	.081	.077
C_3	.176	.166	.158	.150	.142	.139
C_4	.494	.477	.457	.431	.401	.389
C_5	.136	.184	.227	.274	.326	.345
SK	-.720	-.778	-.835	-.878	-.989	-1.021
KT	-.170	-.037	.067	.100	.333	.389
β	.50	.75	.80	.65	.45	.90

(continued on next page)

Note: t1 - t4 are threshold values; C1 - C5 are percentages of population in each response category; SK=Skewness; KT=Kurtosis; β=Factor loadings.

Table IV-2 (Continued)

	Item#					
	7	8	9	10	11	12
t_1	-1.645	-1.645	-1.645	-1.645	-1.645	-1.645
t_2	-1.162	-1.202	-1.222	-1.242	-1.262	-1.282
t_3	-.646	-.688	-.709	-.730	-.751	-.772
t_4	.350	.250	.200	.150	.100	.050
C_1	.050	.050	.050	.050	.050	.050
C_2	.073	.065	.061	.057	.053	.050
C_3	.137	.131	.128	.126	.123	.120
C_4	.378	.353	.340	.327	.314	.300
C_5	.363	.401	.421	.440	.460	.480
SK	-1.083	-1.129	-1.172	-1.213	-1.261	-1.307
KT	.551	.604	.697	.792	.909	1.015
β	.70	.40	.85	.35	.55	.60

Note: t1 - t4 are threshold values; C1 - C5 are percentages of population in each response category; SK=Skewness; KT=Kurtosis; β=Factor loadings.

Table IV-3

Distributional Characteristics of the 12 Core Items in the Condition of Highly Skewed Distribution.

	Item#					
	1	2	3	4	5	6
t_1	-1.815	-1.830	-1.845	-1.860	-1.875	-1.890
t_2	-1.316	-1.336	-1.356	-1.376	-1.396	-1.416
t_3	-.959	-.982	-1.005	-1.028	-1.051	-1.074
t_4	-.460	-.490	-.520	-.550	-.580	-.610
C_1	.035	.034	.033	.031	.030	.029
C_2	.059	.057	.055	.053	.051	.049
C_3	.075	.072	.070	.068	.065	.063
C_4	.154	.149	.144	.139	.134	.130
C_5	.677	.688	.698	.709	.719	.729
SK	-1.753	-1.808	-1.858	-1.916	-1.911	-2.035
KT	2.080	2.291	2.492	2.740	2.751	3.255
β	.60	.35	.85	.50	.80	.45

(continued on next page)

Note: t1 - t4 are threshold values; C1 - C5 are percentages of population in each response category; SK=Skewness; KT=Kurtosis; β=Factor loadings.

Table IV-3 (Continued)

	Item#					
	7	8	9	10	11	12
t_1	-1.905	-1.920	-1.935	-1.950	-1.965	-1.980
t_2	-1.436	-1.456	-1.476	-1.496	-1.516	-1.536
t_3	-1.097	-1.120	-1.143	-1.166	-1.189	-1.212
t_4	-.640	-.670	-.700	-.730	-.760	-.790
C_1	.028	.027	.026	.026	.025	.024
C_2	.047	.045	.043	.042	.040	.038
C_3	.061	.059	.057	.054	.052	.050
C_4	.125	.120	.115	.111	.106	.102
C_5	.739	.749	.758	.767	.776	.785
SK	-2.095	-2.159	-2.140	-2.282	-2.261	-2.330
KT	3.531	3.829	3.807	4.414	4.394	4.759
β	.70	.40	.55	.75	.90	.65

Note: t1 - t4 are threshold values; C1 - C5 are percentages of population in each response category; SK=Skewness; KT=Kurtosis; β=Factor loadings.

Table IV-4

Distributional Characteristics of the 12 Core Items in the Condition of Differentially Skewed Distribution

	Item#					
	1	2	3	4	5	6
t_1	.790	.610	.460	-.050	-.350	-1.100
t_2	1.212	1.074	.959	.772	.646	.331
t_3	1.536	1.416	1.316	1.282	1.162	.862
t_4	1.980	1.890	1.815	1.645	1.645	1.645
C_1	.785	.729	.677	.480	.363	.136
C_2	.102	.130	.154	.300	.378	.494
C_3	.050	.063	.075	.120	.137	.176
C_4	.038	.049	.059	.050	.073	.144
C_5	.024	.029	.035	.050	.050	.050
SK	2.330	2.035	1.753	1.307	1.083	.720
KT	4.759	3.255	2.080	1.015	.551	-.170
β	.40	.75	.90	.55	.65	.45

(continued on next page)

Note: t1 - t4 are threshold values; C1 - C5 are percentages of population in each response category; SK=Skewness; KT=Kurtosis; β=Factor loadings.

Table IV-4 (Continued)

	Item#					
	7	8	9	10	11	12
t_1	-1.645	-1.645	-1.645	-1.815	-1.890	-1.980
t_2	-.862	-1.162	-1.282	-1.316	-1.416	-1.536
t_3	-.331	-.646	-.772	-.959	-1.074	-1.212
t_4	1.100	.350	.050	-.460	-.610	-.790
C_1	.050	.050	.050	.035	.029	.024
C_2	.144	.073	.050	.059	.049	.038
C_3	.176	.137	.120	.075	.063	.050
C_4	.494	.378	.300	.154	.130	.102
C_5	.136	.363	.480	.677	.729	.785
SK	-.720	-1.083	-1.307	-1.753	-2.035	-2.330
KT	-.170	.551	1.015	2.080	3.255	4.759
β	.80	.50	.35	.85	.60	.70

Note: t1 - t4 are threshold values; C1 - C5 are percentages of population in each response category; SK=Skewness; KT=Kurtosis; β=Factor loadings.

underlying the observed variables; 3) the latent variable was continuous; 4) the observed and the latent variable had a monotonic non-decreasing relationship; 5) the latent variable and the measurement error were normally distributed; and 6) the response data were interpersonally comparable (c.f., Brady, 1989), that is, the function which related the latent variable to the response variable was the same for all subjects.

Dependent Variables

Two dependent variables of interest were: 1) the absolute value of Pearson correlations between the recovered and the true person parameters, and 2) the root mean squared differences between the recovered and the true person parameters.

The scaled latent trait continuum had been standardized to have a mean of 0 and a standard deviation of 1 before the two dependent variables were computed. When necessary, it also had been reversed to have the same direction as the true latent trait continuum before it was standardized. This linear transformation was employed in order to

eliminate artificial scaling factors while keeping the distributional shape of the estimated latent trait continuum unchanged.

Independent Variables

There were four independent variables: 1) The number of cases (30 vs. 100 vs. 1000 cases); 2) The number of items (12 vs. 24 items); 3) The distributional characteristics of item responses (normally distributed vs. moderately skewed vs. highly skewed vs. differentially skewed); and 4) The seven statistical procedures used to recover the true parameters (FA-PL vs. FA-PR vs. IRT-GRM vs. WMDU vs. IMDU vs. EMDU vs. SSI). The first three independent variables were used to form (4 X 3 X 2 =) 24 experimental conditions, within each of which five replications of the simulated data were generated.

Because a sample size of 1000 cases (Case I) is too large for most current MDS procedures, only FA-PR, FA-PL, IRT-GRM, and SSI procedures were compared with each other in this case. A sample size of 100 cases (Case II) is very special because all seven statistical procedures were applicable in this

case. This size of sample might be insufficiently large for FA and IRT but was considered very large for MDS procedures. Given this case, a test of FA's and IRT's robustness against small sample size was available. Given 12 or 24 items, a sample size of 30 cases (Case III) was clearly too small for FA and IRT, so that only WMDU, IMDU, EMDU, and SSI were compared with each other in this case.

The distributions of item responses are a function of item threshold values. In terms of standardized Z-scores, the population threshold values in Tables IV-1, IV-2, IV-3, and IV-4 were systematically chosen so that each item had a desired distribution in the population. In Table IV-1, both skewness and kurtosis values were systematically chosen and slightly varied to produce approximately normal distributions. Therefore, the skewness values for the 12 items were between -.010 and .011, and the kurtosis values were between -.050 and .039. Because all items in Table IV-1 are symmetrically distributed, it can be inferred that the variation of the latent scale values of the 12 items should be zero in the population.

In Tables IV-2, IV-3, and IV-4, the skewness values were manipulated without considering kurtosis because the major distributional characteristic of item responses in the current study was skewness. The variation of kurtosis is a consequence of different values of skewness because the principle, kurtosis > skewness^2 - 2, must hold (Kendall & Stuart, 1977, pp. 88, 95). Items in Table 3 were considered to be moderately skewed in practical situations. Their skewness values ranged from -0.720 to -1.307 with a mean of -1.032 and a standard deviation of 0.187. Nearly 50% of the respondents fell into either item response category 4 or 5. By contrast, items in Table IV-3 were considered to be highly skewed in practice. Their skewness values ranged from -1.753 to -2.330, with a mean of 2.046 and a standard deviation of 0.187. In this situation, 67.7% to 78.5% of the respondents were classified in item response category 5. The skewness values in Tables 3 and 4 were deliberately chosen to have different means but the same standard deviation so that, although the degree of skewness was different, the variation of the latent scale values of the 12 items may be considered

approximately equal across the two conditions. Finally, the items presented in Table IV-4 had the highest variation of skewness and, therefore, the highest variation of latent scale values among the four conditions of item responses. Their skewness values ranged between -2.330 and 2.330.

Within each condition from Tables IV-1 to IV-4, twelve ß weights (.35, .40, .45, .50, .55, .60, .65, .70, .75, .80, .85, .90) were randomly assigned to the 12 items. Randomization was employed to avoid correlation between ß weights and skewness values. The twelve items serve as "core items" and were duplicated in order to obtain the condition of 24 items.

The proportion of cases in each response category (C_1 to C_5) was computed as the area between threshold values under the normal curve. An effort was made to avoid the number of expected cases in each response category being zero when extreme threshold values were to be selected. Skewness and kurtosis were computed through the following procedures: 1) successive integers, 1 thru 5, were assigned to the five response categories; 2)

population means and standard deviations of these response scores for each item were obtained with the following two formulas:

$$\mu = E(X) = \Sigma P_i X_i \qquad (3.3)$$

and $$\sigma = E(X^2) - \mu^2, \qquad (3.4)$$

where X_i was the set of integer scores ranging from 1 to 5 and P_i was the probability of each integer score; 3) the assigned integer scores were standardized with the obtained mean and standard deviation; 4) skewness was computed as the third moment of the standardized scores; and 5) kurtosis was computed as the fourth moment minus 3.0 (Kendall & Stuart, 1977, p.88).

The four conditions of the item response distribution were summarized in Table IV-5.

Table IV-5

Range, Mean,, and SD of Skewness Values of 12 Items in Each Condition of Item Response Distributions

Condition	Range of Skewness	Mean of Skewness	SD of Skewness
Normally Distributed	-.010 to .011	.000	.007
Moderately Skewed	-.720 to -1.307	-1.032	.187
Highly Skewed	-1.753 to -2.330	-2.046	.187
Differentially Skewed	-2.330 to 2.330	.000	1.640

Programs for Estimation

For applying the FA-PR and the FA-PL procedures, the LISREL VI computer program (Jöreskog & Sörbom, 1984) was employed. In general, ML estimation was adopted. When item responses were moderately/highly skewed, sample size was 100, and number of items were 24, however, matrices of polychoric correlations were sometimes not positive definite. This situation was encountered 8 times and

unweighted least squares estimation was used. Starting values for all parameters were set at 0.5.

For applying the IRT-GRM procedure, the MULTILOG computer program (Thissen, 1986) was employed. Marginal ML estimation was used to estimate item parameters while conditional ML was used to estimate person parameters. By default, the starting value for descrimination parameters was 1.0, while the starting values for the threshold parameters were -1.39, -0.405, 0.405, and 1.39 respectively.

For applying the WMDU, IMDU, and the EMDU procedures, the SAS PROC ALSCAL computer program (Young & Lewyckyj, 1979) was employed. Because Young and Lewyckyj suggested that results of the analysis might be less valid with similarities data, all similarities data were converted into disimilarities data.

A scheme of recoding raw data was required before the WMDU model could be applied. Let Y represent the integer response score, while C_1, C_2, C_3, C_4, and C_5 represent the five response categories as column stimuli. According to the following scheme, data recoding was made for each subject:

If Y=1, then C_1=1, C_2=2, C_3=3, C_4=4, C_5=5;

If Y=2, then C_1=2, C_2=1, C_3=2, C_4=3, C_5=4;

If Y=3, then C_1=3, C_2=2, C_3=1, C_4=2, C_5=3; (3.5)

If Y=4, then C_1=4, C_2=3, C_3=2, C_4=1, C_5=2;

If Y=5, then C_1=5, C_2=4, C_3=3, C_4=2, C_5=1.

After the above recoding procedure was applied to every subject, multiple rectangular matrices were formed with row stimuli corresponding to subjects, column stimuli corresponding to response categories, and matrices corresponding to items. Data in each rectangular matrix were considered to be ordinal and row-conditional. Tied data were set to be untied. Number of dimensions was set at two, which was the minimum number required by the individual differences MDS. The subject coordinates on the first dimension were taken as estimates of the latent trait. The second dimension represents a residual vector.

The usual subject-by-item rectangular data were directly submitted to the ALSCAL program for IMDU analysis. The data were considered to be ordinal, and matrix-conditional. Tied data were set