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

Jason C. Chan

詹 志 禹*

摘 要

本研究比較三個統計模式從李克特式(Likert-Type)資料中估計單向度潛在特質的能力，這三個模式是「植基於多序類相關(polychoric correlations)的因素分析」(FA-PL)，「試題反應理論中的漸變反應模式」(IRT-GRM)，以及「加權多維展開法」(WMDU)。一般常用的方法(SSI) ——分派連續性整數給李克特式量尺中的一個反應類別（如「非常同意」），再將每一題得分加總——則做爲比較的基準線。本研究爲電腦模擬研究，操弄了樣本大小、測驗長度，以及試題反應分配的偏態程度等三個自變項，依變項則爲回復潛在特質的眞值的正確性，結果發現：IRT-GRM表現得最好，最不受偏態的影響；FA-PL只有在試題反應分配爲常態時，才能表現與IRT-GRM一樣好，而在試題反應分配爲高度偏態時，甚至表現得比SSI差；最後，WMDU只有在試題反應分配爲常態或輕微偏態時，才能表現得與SSI一樣好。本文也討論了這些發現對模式選擇的涵意。

關鍵字：李克特式量表、潛在特質、因素分析、試題反應理論、多維尺度法、電腦模擬研究。

Abstract

Three statistical models 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 polychoric correlations (FA-PL), the graded response model in item response theory (IRT-GRM), and the weighted multidimensional unfolding (WMDU). The common procedure of summing up successive integers assigned to response categories (SSI) served as the base-line procedure. 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-PL were competitive with IRT-GRM only when item responses were normally distributed. It performed even worse than did SSI when item responses were highly skewed. WMDU might be a rival alternative to SSI only when item responses were normally distributed or moderately skewed and sample size was large for MDU models (e.g., N=100).

Index terms: Likert-type data, latent trait, factor analysis, item response theory, multidimensional scaling, Monte-Carlo study.

*作者為本校教育系副教授

Estimating Latent Trait

Data collected from social/psychological research typically produces ordinal variables. 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 (θ), which is the variable of ultimate interest in most social/psychological researches. Therefore, the main objective of the current study is to compare three statistical models, not in all their aspects, but in terms of their ability to estimate the latent variable of interest.

The Assumed Response Processes

Two kinds of disturbance processes are assumed to be involved in measuring a latent dimension given Likert-type items: a stochastic process and a crude-classification process. First of all, it is usually assumed that the latent dimension (θ) and the continuous quantitative response (Y*) are linearly and probabilistically related. The basic mathematical form of this relationship is:

$$Y^* = \beta\,\theta + \epsilon, \tag{1}$$

where β is a weight and ϵ is the residual. The latent dimension (θ) is assumed to be stable across various replications, while the residual (ϵ) is assumed to be specific to replications. For estimation convenience, both θ and ϵ 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 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:

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

Where t_i (i=1, 2, 3, 4) are the threshold values or category boundaries. These values may be affected 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 θ instead of Y^*. The latent dimension, θ, is related to the manifest variable (Y) through both a linear/stochastic function and a step function.

Traditional Approach

When multiple Likert-type items as indicators of one latent variable are available, researchers usually assign successive integers to the response categories and then simply sum up the raw scores on each item 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 weights for all items and of equal intervals between ordinal response categories. Three alternative statistical procedures were proposed by the current study. They were the factor analysis based on polychoric correlations (FA-PL) (Olsson, 1979b), the graded response model from the item response theory (IRT-GRM) (Samejima, 1969), and the weighted multidimensional unfolding (WMDU) (Young, 1984). They have been growing out of three distinct areas but have converged in many aspects, which will be explicated later. Here, it is worth noting that the application of WMDU to Likert-type data is unique to this paper.

FA-PL

Factor analysis (FA) is usually performed on the matrix of Pearsonian correlations. Computation of Pearsonian correlations between Likert-type items assumes that successive integers assigned to the response categories are in equal-interval scales. It is also well known that the Pearsonian correlation is not free to range from –1 to 1 when the two correlated items are skewed highly in opposite directions (Carroll, 1961; Muthén, 1983). In a simulation study, Olsson (1979a) indeed found a substantial lack of fit of the true model and attenuated estimates of factor loadings when the

maximum likelihood FA was performed on the Pearsonian correlations. He suggested that researchers perform a FA on polychoric correlations when observed variables were obtained from a classification of some continuous latent variables. Olsson (1979b) presented two maximum likelihood (ML) estimation procedures for the polychoric correlations. One of the procedure was implemented in the LISREL program by Jîreskog and Sîrbom (1984, 1989).

Jöreskog and Sörbom (1988) showed 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 Jîreskog (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.

Based on the above findings, the current study decided to perform factor analysis on polychoric correlations (FA-PL) rather than on other types of correlations when Likert-type data were to be analyzed.

IRT-GRM

IRT typically has been developed for scaling dichotomous data onto an equal-interval scale. For dealing with ordinal 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). Only the last model was considered in the current study because discrimination parameters were assumed to vary across items.

Samejima (1969) 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-1 dichotomous items and the two-parameter model was applied to model the cumulative probability of an individual responding to a particúlar or higher category. This idea is expressed by the following equation:

$$\sum_{j=k}^{m} \pi_{nij} = \frac{\exp[a_i(\theta_n - b_{ik})]}{1 + \exp[a_i(\theta_n - b_{ik})]}, \quad k = 1, \ldots, m \tag{3}$$

where π_{nij} is the probability of person n responding to item i with response category j, 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 θn is the latent trait level of individual n.

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

$$\pi_{nij} = \sum_{j=k}^{m} \pi_{nij} - \sum_{j=k+1}^{m} \pi_{nij}, \quad k = 0, \ldots, m. \tag{4}$$

Equation (3) is the general form for drawing the operating characteristic curves for a graded response item.

Samejima's graded response model allows a parameter for each item to have a different discrimination power. In addition, it allows the category boundaries to vary across items. This model is conceptually similar to the FA-PL model. The applicability of the IRT-GRM to Likert-type scales has been demonstrated by several studies (e.g., Dodd, 1984; Dodd, Koch & DeAyala, 1988; Koch, 1983; Thissen & Steinberg, 1988).

WMDU

Traditionally, multidimensional scaling (MDS) has sometimes been applied to proximity measures including intercorrelations or squared euclidean distances between Likert-type items (e.g., Romney, Shepard, & Nerlove, 1972). These applications, however, usually concentrated on the configuration of items and neglected the estimation of subjects' coordinates. If MDS practitioners were interested in estimating subjects' coordinates, they might employ Coombs' (1964) internal unfolding or Carroll's (1972) external unfolding procedures. Unfortunately, the two classical unfolding procedures may not fit the response process of Likert data very well because they simply compress each Likert item into one geographical point and bypass the estimation of threshold values and of discrimination parameters. Inadequate matching between analytic models and response processes causes inappropriate estimation and/ or degenerate solutions (Chan, 1991; Koch, 1984).

Davison and Skay (1992) also discussed two MDS alternatives to FA for item responses. The first alternative was a vector model in which subject parameters reflected the significance of a dimension for the subject rather than the latent-trait level of the subject. The second alternative was still in the tradition of Coombs's (1964) internal

unfolding model. Consequently, neither alternatives can adequately model the response process of Likert-type data discussed previously.

If response categories of each Likert item were regarded as stimuli and each item was treated as a vector of response categories rather than a point, then the correlation between each item and the latent dimension can be modeled by WMDU (Young, 1984). WMDU 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:

$$\delta_{nji} \approx f_n(d_{nji}) = f_n\{[(\underline{y}_n - \underline{x}_j)'W_i(\underline{y}_n - \underline{x}_j)]^{1/2}\}, \tag{5}$$

where δ_{nji} is the proximity between row n and column j for matrix i, f_n is the monotonic function for row n, d_{nji} is the estimated distance corresponding to δ_{nji}, yn is the vector for row n, $\underline{x}_j$ is the vector for column j, and the diagonal W_i is the weight for matrix i. In the current situation, subscripts n, j, and i correspond to subjects, response categories (e.g., "strongly disagree" or "moderately agree"), and items. Therefore, δ_{nji} is the proximity measure between subject n and response category j on item i, d_{nji} the estimated distance between subject n and response category j on item i, $\underline{y}_n$ the estimated coordinates for subject n, $\underline{x}_j$ the estimated coordinates for response category j, and W_i the weight for item i. 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. It should be noted that δ_{nji} is a rescored proximity measure indicating preference order in response categories, which is to be demonstrated in the next paragraph.

A scheme of recoding raw data is required before the WMDU model could be applied to Likert items. The following demonstration assumes 5-point scales. Let Y represent the integer response score obtained by each subject, while C_1, C_2, C_3, C_4, and C_5 represents the five response categories across items. According to the following scheme, data recoding can be made for each subject:

If Y=1, then C_1=1, C_2=2, C_3=3, C_4=4, C_5=5;
Else if Y=2, then C_1=2, C_2=1, C_3=2, C_4=3, C_5=4;
Else if Y=3, then C_1=3, C_2=2, C_3=1, C_4=2, C_5=3; (6)
Else if Y=4, then C_1=4, C_2=3, C_3=2, C_4=1, C_5=2;
Else 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, a "three-way

three-mode" data set was formed with row stimuli corresponding to subjects, column stimuli corresponding to response categories, and matrices corresponding to items. The above derivation implied that Y is an interval variable. When the WMDU model was applied, however, data in each rectangular matrix were considered to be ordinal and row-conditional because of the existence of f_n in equation (5). This is a novel application of unfolding models to Likert-type data. With this application, however, coordinates of subjects and of item response categories and the weight for each item can be estimated with respect to 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.

Comparative Analysis

Although FA-PL, IRT-GRM and WMDU have been growing out of three distinct areas, they are similar in many aspects. First of all, the original data for FA-PL/IRT-GRM and the derived data for WMDU are assumed to be ordinal but the underlying dimension(s) is (are) assumed to be continuous. Second, they are able to estimate threshold values of response categories for each Likert item. Third, they are able to estimate the relationship between each item and the latent dimensions(s): FA-PL with factor loadings, IRT-GRM with discrimination parameters, and WMDU with weights for matrices. Finally, they all estimate the level of latent trait for each subject: FA-PL with factor scores, IRT-GRM with subject parameters, and WMDU with subject coordinates. Note that item parameters, including estimates of threshold values and of relationships between items and the latent dimension, were not variables of direct interest in the current study. They were not directly comparable because they had different metrics in the three models of interest. It was assumed by the current study that the inadequacy of modeling the item parameters would ultimately be reflected in the inaccuracy of estimating the subject parameters.

The three procedures are also different in many aspects. First of all, IRT-GRM is only applicable to unidimensional data while the other two procedure are able to deal with multidimensional data. Thus, the current study merely investigated the unidimensional case. Second, WMDU estimates one set of response-category coordinates for all items while the other two procedures estimates different sets of threshold values for different items. Therefore, WMDU is less flexible and should perform worse than do the other two procedures when threshold values vary across

items, which is the case assumed by the current study. Finally, FA-PL with the two-stage maximum-likelihood estimators of polychoric correlations requires that the latent variables are normally distributed, which in turn requires that the corresponding observed item responses are *approximately* normally distributed. The more severely the distributions of item responses depart from normal distributions, the less accurately the latent relationships will be estimated. By contrast, there is a well known property of "item-free person measurement" and "sample-free test calibration" in IRT if the particular model fits the data (Wright, 1967). Therefore, the performance of FA-PL is predicted to be as good as IRT-GRM only when item responses were normally distributed and to be worse than it when distributions of item responses were skewed.

The effect of the distributional property of the latent dimensions on the WMDU estimates is unknown. With the alternating least square approach (Young & Lewyckyj, 1979), MDS or MDU solutions are usually descriptive rather than inferential. Therefore, distributional assumptions are seldom discussed except in the emerging field of maximum likelihood MDS. However, no mention of distributional assumptions does not imply that the estimates from WMDU will not be affected by the distribution of the latent dimension. Gillespie (1989) have noticed that item response distributions may affect MDS results. The current study expected that the degree of distributional skewness should have effects on WMDU performances. In addition, since WMDU estimates one set of response-category coordinates for all items, it should perform worse than did FA-PL and IRT-GRM when distributions of item responses were differentially skewed, where response categories had highly varying threshold values across items (This condition is defined in the methodology section).

The current study was concerned about not only the practical implications of the compared statistical procedures but also the theoretical interest in pointing out the conceptual similarity of FA-PL, IRT-GRM, and WMDU, which were usually discussed in isolated traditions.

Methodology

The Simulated Situation

For each item, a continuous response variable (Y*) was generated by the RANNOR function of the SAS statistical package (SAS institute, Inc., 1985) according to the following formula:

$$Y^* = \beta\ \Theta + \sqrt{(1-\beta^2)}\,E, \qquad (7)$$

where Θ 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 Θ and E were generated by SAS RANNOR 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 a variance of one. The values of · were systematically chosen and randomly assigned to each item (see Table 1).

Secondly, the integer response score (Y) for each simulated subject was decided according to equation (2). Note that the threshold values (t_i) in equation (2) were systematically chosen to produce desired skewness values, which will be explicated later.

Dependent Variables

The dependent variable of interest was the absolute value of Pearson correlations between the recovered and the true person parameters. When necessary, the estimated latent trait continuum was reversed to have the same direction as the true latent trait continuum before the correlation was computed. Pearson correlations was employed as criterion in order to eliminate artificial scaling factors in different computer programs while keep the distributional shape of the estimated latent trait continuum unchanged.

Independent Variables

There were four independent variables: 1) Sample sizes (*30* vs *100* vs *1000* subjects); 2) Test lengths (*12* vs *24* items); 3) Distributional characteristics of item responses (*normally distributed* vs *moderately skewed* vs *highly skewed* vs *differentially skewed*); and 4) Four statistical procedures used to recover the true parameters (*FA-PL* vs. *IRT-GRM* vs. *WMDU* vs. *SSI*). The first three independent variables were used to form 24 (= 4 X 3 X 2) experimental conditions, within each of which five replications of the simulated data were generated. In each cell, five replications were among the lowest limit of number of replications found in literature and were adopted by the current study simply due to limitation of computer time, which is expensive for FA-PL, IRT-GRM, and WMDU.

Because a sample size of 1000 subjects (Case I) was too large for most current

MDU procedures, only FA-PL, IRT-GRM, and SSI procedures were compared with each other in this case. A sample size of 100 subjects (Case II) was very special because all four 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 MDU procedures. Given this case, a comparison of FA's with IRT's robustness against small sample size was available. Finally, a sample size of 30 subjects (Case III) was evidently too small for FA and IRT, so that only WMDU and SSI were compared with each other. The last case was investigated because MDU models are frequently employed to deal with small sample sizes.

The distributions of item responses are a function of item threshold values. In terms of standardized Z-scores, the population threshold values were systematically chosen so that each item had a distribution with desired skewness value in the population (see Table 1). Skewness 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 formulas, $\mu = E(X) = \sum P_i X_i$, and $\sigma^2 = E(X^2) - \mu^2$, 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.

Skewness values were deliberately chosen to produce normal, moderately skewed, highly skewed, and differentially skewed distributions of item responses (see Table 1 and Table 2). Within each distributional condition, twelve β weights (.35, .40, .45, .50, .55, .60, .65, .70, .75, .80, .85, .90) were randomly assigned to the 12 items. The twelve items serve as "core items" and were duplicated in order to obtain the condition of 24 items.

The expected proportion of subjects responding in each response category was computed as the area between threshold values under the normal curve. An effort was made to avoid the expected number of cases in each response category being zero when extreme threshold values were to be selected.

Programs and Procedures for Estimation

For applying the FA-PL procedure, the LISREL VI computer program (Jöreskog & Sîrbom, 1984) was employed. In general, ML estimation was adopted. However, when the matrix of polychoric correlations was not positive definite, the unweighted least square estimation was used. Starting values for all parameters were set at 0.5.

Table 1
Population Threshold Values in Terms of Z-Scores

Dist.	Item No.	Threshold Values t_1	t_2	t_3	t_4	Skewness	β
(1)	1.	−2.076	−.716	716	2.076	.000	.55
	2.	−2.052	−.712	712	2.052	.000	.80
	3.	−1.836	−.676	676	1.836	.011	.35
	4.	−1.812	−.672	672	1.812	.011	.75
	5.	−1.740	−.660	660	1.740	.010	.50
	6.	−1,692	−.652	652	1.692	.000	.45
	7.	−1.668	−.648	648	1.668	.000	.90
	8.	−1.620	−.640	640	1.620	.000	.65
	9.	−1.596	−.636	636	1.596	.000	.40
	10.	−1.548	−.628	628	1.548	−.010	.85
	11.	−1.500	−.620	620	1.500	−.010	.70
	12.	−1.476	−.616	616	1.476	−.010	.60
(2)	1.	−1.645	−.862	−.331	1.100	−.720	.50
	2.	−1.345	−.942	−.415	.900	−.778	.75
	3.	−1.645	−1.002	−.478	.750	−.835	.80
	4.	−1.645	−1.062	−.541	.600	−.878	.65
	5.	−1.645	−1.122	−.604	.450	−.989	.45
	6.	−1.645	−1.142	−.625	.400	−1.021	.90
	7.	−1.645	−1.162	−.646	.350	−1.083	.70
	8.	−1.645	−1.202	−.688	.250	−1.129	.40
	9.	−1.645	−1.222	−.709	.200	−1.172	.85
	10.	−1.645	−1.242	−.730	.150	−1.213	.35
	11.	−1.645	−1.262	−.751	.100	−1.261	.55
	12.	−1.645	−1.282	−.772	.050	−1.307	.60

(continued on next page)

Table 1
(continued)

Dist.	Item No.	Threshold Values t_1	t_2	t_3	t_4	Skewness	β
(3)	1.	−1.815	−1.316	−.959	−.460	−1.753	.60
	2.	−1.830	−1.336	−.982	−.490	−1.808	.35
	3.	−1.845	−1.356	−1.005	−.520	−1.858	.85
	4.	−1.860	−1.376	−1.028	−.550	−1.916	.50
	5.	−1.875	−1.396	−1.051	−.580	−1.911	.80
	6.	−1.890	−1.416	−1.074	−.610	−2.035	.45
	7.	−1.905	−1.436	−1.097	−.640	−2.095	.70
	8.	−1.920	−1.456	−1.120	−.670	−2.159	.40
	9.	−1.935	−1.476	−1.143	−.700	−2.140	.55
	10.	−1.950	−1.496	−1.166	−.730	−2.282	.75
	11.	−1.965	−1.516	−1.189	−.760	−2.261	.90
	12.	−1.980	−1.536	−1.212	−.790	−2.330	.65
(4)	1.	.790	1.212	1.536	1.980	2.330	.40
	2.	.610	1.074	1.416	1.89-	2.035	.75
	3.	.460	.959	1.316	1.815	1.753	.90
	4.	−.050	.772	1.282	1.645	1.307	.55
	5.	−.350	.646	1.162	1.645	1.083	.65
	6.	−1.100	.331	.862	1.645	.720	.45
	7.	−1.645	−.862	−.331	1.110	−.720	.80
	8.	−1.645	−1.162	−.646	.350	−1.083	.50
	9.	−1.645	−1.282	−.772	.050	−1.307	.35
	10.	−1.815	−1.316	−.959	−.460	−1.753	.85
	11.	−1.890	−1.416	−1.074	−.610	−2.035	.60
	12.	−1.980	−1.536	−1.212	−.790	−2.330	.70

Note: Dist. (1) = Normally distributed; Dist. (2) = Moderately skewed; Dist. (3) = Highly skewed; Dist. (4) = Differentially skewed.

Table 2. Range, Mean, and SD of Skewness Values of 12 Items in Each Condition of Item Response Distributions

Dist.	Range of Skewness	Mean of Skewness	SD of Skewness
(1)	–1.010 to .011	.000	.007
(2)	– .720 to – 1.307	–1.032	.187
(3)	–1.753 to – 2.330	–2.046	.187
(4)	–2.330 to 2.330	.000	1.640

Note: Dist. (1) = Normally distributed; (2) = Moderately skewed; (3) = Highly skewed; (4) = Differentially skewed.

In LISREL VI, the matrix A of regression weights to compute factor scores were obtained by the formula (Lawley & Maxwell, 1971, p.109):

$$A = \phi \Lambda' \Sigma^{-1},$$

where ϕ is the matrix of estimated factor correlations, Λ is the matrix of estimated factor patterns, and Σ is the matrix of reproduced correlations.

For applying the IRT-GRM procedure, the MULTILOG computer program (Thissen, 1986) was employed. ML estimation was used to estimate item while conditional marginal ML was used to estimate person parameters. By default, the starting value for discrimination parameter 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 procedure, the SAS PROC ALSCAL computer program (Young & Lewyckyj, 1979) was employed. Before analysis, the usual item-person data matrix was converted into k (the number of items) category-person rectangular matrices according to equation (9). 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.

Results

Means and standard deviations (SDs) of correlations between true and recovered person parameters were presented in Table 3 thru Table 5. Since correlations in each

cell were rather close, their means and SDs were computed without transforming correlations to Fisher's Z because the transformation made negligible differences in the resultant mean correlations (after the 2nd digit of the decimal point.)

Descriptive rather than inferential statistics were used for analysis because of the following reasons: a) In Table 3, significance tests of mean differences would be redundant due to extremely low variation across replicated samples; b) In Table 4 and 5, significance tests of mean differences seemed inappropriate because the homogeneity assumption was violated; and c) Practical significance was of greater concern than was statistical significance in the current situation. The following analyses would focus on systematic patterns and practical significance. Practical significance was roughly judged by the ratio of each mean difference to the standard deviation of the combined sample. For being judged as "practical significance" a mean difference of at least one standard deviation was considered. When the two standard deviations involved in comparing means were highly heterogeneous, a more conservative conclusion was drawn.

Case I: N=1000

In this case, IRT-GRM, FA-PL, and SSI were compared with one another. Generally speaking, IRT-GRM performed best among the three procedures in terms of recovering the latent trait parameters (Table 3). However, a few interactions between the recovering procedures and the distributions of item responses could also be observed. First, compared to FA-PL, IRT-GRM performed just as well as did FA-PL when the item responses were normally distributed but it tended to outperform FA-PL when the distributions of item responses became skewed. The advantage of IRT-GRM was especially evident when the distributions of item responses were highly skewed. Second, compared to the common SSI procedure, IRT-GRM performed slightly better when the distributions of item responses were normal, moderately skewed, or differentially skewed but much better when the distributions were highly skewed. It seemed that IRT-GRM was more robust against skewness than the other two procedures involved. These results were true across the conditions of test length.

FA-PL performed slightly better than SSI when item responses were normally distributed. This advantage of FA-PL tended to disappear when distributions of item responses were moderately or differentially skewed. Moreover, FA-PL performed slightly worse than SSI when the distributions of item responses were highly skewed.

It seemed that FA-PL was more sensitive to the disturbance of skewness than was SSI. These results were true across the conditions of test length.

It is not surprising that both test length and distributions of item responses have a main effect. All three procedures performed better in the condition of 24 items than in the condition of 12 items. They performed best when the distributions of item responses were normal, next best when the distributions were moderately or differentially skewed, and worst when the distributions were highly skewed.

All recovering procedures were pretty stable across replicated samples according

Table 3
Means and S.D.s of Correlations between True and Recovered Person Parameters for Case I (N=1000)

#Items	Dist.[a]	IRT-GRM	FA-PL	SSI
12	(1)	.95[b] (.00)[c]	.95 (.00)	.93 (.00)
	(2)	.94 (.00)	.92 (.00)	.91 (.00)
	(3)	.88 (.00)	.80 (.01)	.81 (.01)
	(4)	.93 (.00)	.93 (.01)	.91 (.00)
24	(1)	.98 (.00)	.97 (.00)	.96 (.00)
	(2)	.97 (.00)	.94 (.00)	.94 (.00)
	(3)	.92 (.00)	.83 (.01)	.85 (.01)
	(4)	.97 (.00)	.96 (.00)	.95 (.00)

a Dist.(1) = Normally distributed; (2) = Moderately skewed; (3) = Highly skewed; (4) = Differentially skewed.
b Means over replications.
c The number in the parenthesis is the standard deviation of five replicated correlations.

to the fact that all standard deviations of the five correlations were less than or equal to 0.01.

Case II: N=100

In this case, all four procedures were compared with one another. Generally speaking, IRT-GRM performed best among the four procedures in terms of recovering the latent trait parameters (Table 4). This conclusion depended somewhat on the condition of item response distributions. When the item responses were normally distributed, IRT-GRM performed just as well as did FA-PL, WMDU, and SSI.

However, it outperformed all other three procedures when the distributions of item responses were moderately or highly skewed. When the distributions of item responses were differentially skewed, IRT-GRM performed slightly better than did FA-PL and SSI but much better than did WMDU. It seemed that IRT-GRM was more robust against skewness than were the other three procedures.

FA-PL and SSI had similar performances across most conditions. However, a slight tendency might be observed: although FA-PL performed as well as or even slightly better than did SSI when the distributions of item responses were normal or moderately skewed, it performed worse than did SSI when the distributions were highly skewed. These results implied that FA-PL might be more severely disturbed by a high degree of skewness than was SSI.

WMDU performed as well as did IRT-GRM, FA-PL, and SSI when distributions of item responses were normal. It performed worse than IRT-GRM when distributions of item responses were moderately or highly skewed. It performed much worse than all other three procedures when item responses were differentially skewed.

As in Case I, all procedures performed best when item responses were normally

Table 4
Means and S.D.s of Correlations between True and Recovered Person Parameters for Case II (N=100)

Dist.[a]	IRT-GRM	FA-PL	SSI	WMDU
		12 items		
(1)	.95[b] (.01)[c]	.95 (.01)	.95 (.01)	.94 (.01)
(2)	.95 (.01)	.92 (.01)	.91 (.02)	.92 (.01)
(3)	.89 (.02)	.79 (.06)	.82 (.03)	.81 (.04)
(4)	.92 (.02)	.92 (.02)	.90 (.03)	.83 (.02)
		24 items		
(1)	.97 (.00)	.97 (.00)	.96 (.00)	.97 (.00)
(2)	.96 (.00)	.94 (.01)	.94 (.01)	.94 (.01)
(3)	.90 (.01)	.84 (.03)	.85 (.03)	.81 (.05)
(4)	.97 (.01)	.95 (.03)	.95 (.01)	.89 (.03)

a Dist.(1) = Normally distributed; (2) = Moderately skewed; (3) = Highly skewed; (4) = Differentially skewed.
b Means over replications.
c The number in the parenthesis is the standard deviation of five replicated correlations.

distributed and worst when distributions were highly skewed. Besides, increasing test length improved estimation for all procedures regardless of conditions of item response distributions.

Case III: N=30.

In this case, WMDU and SSI were compared with each other. The common SSI procedure performed as well as or better than the WMDU procedure in all conditions of test length and of response distributions (Table 5).

Not surprisingly, both SSI and WMDU performed best when item responses were normally distributed and performed worst when item responses were highly skewed. However, SSI performed better when item responses were differentially skewed than when item responses were moderately skewed, while WMDU had the opposite pattern of performances. The above outcomes became even clearer when number of items became larger.

Conclusions across Cases

Table 5
Means and S.D.s of Correlations between True and Recovered Person Parameters for Case III (N=30)

#Items	Dist.[a]	SSI	WMDU
12	(1)	.93[b] (.02)[c]	.92 (.05)
	(2)	.89 (.05)	.88 (.05)
	(3)	.79 (.08)	.76 (.04)
	(4)	.91 (.03)	.84 (.07)
24	(1)	.96 (.01)	.96 (.02)
	(2)	.94 (.01)	.92 (.02)
	(3)	.87 (.04)	.81 (.04)
	(4)	.94 (.02)	.91 (.02)

a Dist.(1) = Normally distributed; (2) = Moderately skewed; (3) = Highly skewed; (4) = Differentially skewed.
b Means over replications.
c The number in the parenthesis is the standard deviation of five replicated correlations.

1) From the best to the worst, the general performances of the four procedures could be ordered as follows: a) IRT-GRM, b) FA-PL and SSI, and c) WMDU.

2) IRT-GRM was more robust against skewness than the other three procedures even when sample size was as small as N=100.

3) FA-PL was competitive with IRT-GRM only when item responses were normally distributed.

4) The WMDU procedure was generally not a better alternative to the common SSI practice. It performed as well as did SSI only when item responses were normally distributed or moderately skewed and when sample size was large for MDU (e.g., N=100).

Discussion

The predicted relative merits of FA-PL, IRT-GRM, and WMDU were generally confirmed. IRT's non-linear formulation of the relationship between responses and latent traits seemed to have some advantages, especially when distributions of item responses were skewed. The current study suggests that when distributions of item responses are moderately or highly skewed, IRT-GRM is the favored choice for estimating the latent trait. When item responses are normally distributed, IRT-GRM, FA-PL, and WMDU are all reasonable choices, which may be decided according to other objectives.

With small samples and "few" items (e.g., 100 subjects and 12 items), current results seemed to suggest that the recovery for IRT-GRM was better than what researchers have generally experienced with 2-parameter IRT models. The reason is probably not in the particular computer program used but in the particular IRT models used. In IRT-GRM, one Likert item with m response categories was viewed as a combination of m-1 dichotomous items. Therefore, 12 Likert items with 5-point scales are potentially able to provide information from 48 dichotomous items. Increasing number of homogeneous items normally improves estimation of person parameters.

Although distributional assumptions are infrequently discussed in the MDS tradition, the current study found that item response distributions did affect WMDU performances. WMDU was not more robust against skewness than any other procedures investigated. Besides, WMDU performed evidently worse than all other three procedures when item responses were differentially skewed. It is surprising that WMDU performed even worse than did SSI in this condition.

The common SSI practice for estimating the latent trait from Likert-type data has

usually been criticized for its strong assumption of equal intervals in the ordinal scales. The present study found that this strong assumption was not very harmful given that the number of response categories for each item was five. In the present study, the SSI procedure might be slightly worse than the IRT-GRM but was comparable with the sophisticated FA procedure. The SSI procedure was even more robust against skewness than was the FA procedure. In addition, SSI was as good as or better than the WMDU procedure. If computer time and computational simplicity were the major consideration, the common SSI practice became more attractive than the other three procedures. This implication is limited in the conditions currently investigated but is consistent with Cohen's (1990) suggestion.

Based on the results of the current study, SSI might continue to be the procedure of choice because of its simplicity. Then, under what conditions would a researcher prefer an estimation based on mathematical models? The answer depends on the objectives of the researcher. Sometimes, researchers would sacrifice simplicity for the mathematical properties of the three models involved. IRT models have the well known property of "item-free person measurement" and "sample-free test calibration" (Wright, 1967). Therefore, IRT models are very suitable for dealing with test equating and adaptive testing. FA models are able to handle large numbers of latent variables and, if embedded in the context of structural equation modeling, are able to statistically test hypotheses about latent structure, i.e., the relationships among latent variables. Finally, MDS transforms psychological relationships to spatial relationships and can satisfy researchers who prefer pictures to numbers for revealing the meaning of data.

Limitations

Regarding the latent space, the current study investigated only the unidimensional case. Some of the current findings may not be generalizable to multidimensional cases. For example, IRT-GRM which utilizes marginal ML estimation based on the full information approach is preferable for long tests with few factors. Marginal ML 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. Therefore, the merit of IRT-GRM over the other procedures considered may disappear in the multidimensional cases.

Although Likert-type items are often based on 5-point response scales, general attitude measurements may also frequently be implemented with non 5-point scales. It is unknown whether or not the current findings would generalize to non 5-point

scales. Although the current study had no reason opposing this kind of generalization, a further study is needed for providing persuasive evidences. It is possible that, due to the increasing or decreasing number of parameters to be estimated, the relative merits of the three models investigated may change with differing numbers of response categories.

There are many kinds of item-response distributions that are qualitatively different from one another. For example, U-shaped, uniform, or other irregular distributions are occasionally encountered in practice. The current study, however, investigated only normal and skewed distributions. It is still unknown how accurate the four statistical procedures would be in estimating the latent trait from Likert-type data given those response distributions which were not investigated.

Invariance of the response threshold values across subjects is a major and neglected assumption of psychological research (Brady, 1989). It is also assumed by the current simulation processes. The effects of violating this assumption on the performances of the statistical procedures discussed above need to be explored in a further study. If this assumption is violated seriously, performances of all procedures except WMDU may be affected. WMDU may be appropriate for interpersonally incomparable data because in the "three way three mode" data set input for WMDU, the row stimuli which represent subjects/persons can be set to be row-conditional.

Although the current study tended to assume that the performance differences among investigated procedures were mainly due to mathematical modeling, estimation algorithms could be a minor confounding factor when comparisons were made across IRT, FA, and MDU. Further studies are needed to isolate the effects of estimation from those of modeling.

Finally, the number of replications in the current study was small though the replicated estimations seemed to be stable. Besides, the criterion for assessing the four procedures was insufficient because Pearson correlations can reflect only the stability of relative positions of latent traits. Further studies were encouraged to utilize more replications and more assessing criteria.

Acknowledgements

This paper was part of a dissertation which was sponsored by the Chiang Ching-Kuo Foundation for International Scholarly Exchange (USA). The author also would like to thank Dr. William R. Koch, Dr. Barbara G. Dodd, Dr. William L. Hays, Dr. H. Paul Kelley, Dr. John C. Loehlin, and several anonymous reviewers for their

constructive criticisms.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Babakus, E., Ferguson, C. E., Jr. & Jîreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, 24, 222-228.

Brady, H. E. (1989). Factor and ideal point analysis for interpersonally incomparable data. *Psychometrika*, 54, 181-202.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika, 26*, 347-371.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35*, 238-319.

Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, & S. Nerlove (Eds.), *Multidimensional scaling: Theory and applications for the behavioral sciences. Vol. I: Theory.* New York: Seminar Press.

Chan, J. C. (1991). Estimating the latent trait from Likert-type data: A comparison of factor analysis, item response theory, and multidimensional scaling. Paper presented at the Annual Meeting of the American Psychological Association, San Francisco.

Clogg, C. C. (1979). Some latent structure models for the analysis of Likert-type data. *Social Science Research*, 8, 287-301.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304-1312.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Davison, M. L. & Skay, C. L. (1992). Multidimensional scaling and factor models of test and item responses. *Psychological Bulletin, 110*, 551-556.

Dodd, B. G. (1984). *Attitude scaling: A comparison of the graded response and partial credit latent trait models.* Doctoral Dissertation, University of Texas at Austin.

Dodd, B. G., Koch, W. R. & DeAyala, R. J. (1988). Computerized adaptive attitude measurement: A comparison of the graded response and rating scale models. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Gillespie, M. K. (1989). A comparison of factor analysis and multidimensional scaling applied to survey data. Dissertation of the University of Texas at Austin, Austin, TX.

Jîreskog, K G., & Sîrbom, D. (1984). *LISREL VI: User's guide* (3rd ed.). Mooresville, IN: Scientific Software.

Jöreskog, K. G., & Sîrbom, D. (1988). *PRELIS: A preprocessor for LISREL.* Mooresville, IN: Scientific Software, Inc.

Jîreskog, K G., & Sîrbom, D. (1989). *LISREL 7: User's Reference Guide.* Mooresville, IN: Scientific Software, Inc.

Koch, W. R (1983). Likert-type scaling using the graded response latent trait model. *Applied Psychological Measurement, 7*, 15-32.

Koch, W. R. (1984). Degenerate solutions in multidimensional unfolding. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Lawley, D. N. & Maxwell, A. E. (1971). Factor analysis as a statistical method (2nd ed.). London: Butterworths.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

MuthÇn, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometries, 22,* 43-65.

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

Olsson, U. (1979b). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44,* 443-460.

Romney, A. K., Shepard, R. N., & Nerlove, S. B. (1972). Multidimensional scaling: Theory and applications in the behavioral sciences. New York: Seminar Press.

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

SAS Institute, Inc. (1985). SAS user's guide. Cary, NC: Author.

Thissen, D. (1986). *MULTILOG: A user's guide.* Mooresville, IN: Scientific Software, Inc.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104,* 385-395.

Wright, B. D. (1967). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, NJ: Educational Testing Service, 1986.

Young, F. W. (1984). The general Euclidean model. In H. G. Law et al. (Eds.), *Research methods for multimode data analysis.* New York: Praeger. pp. 440-465.

Young, F. W., & Lewyckyj, R. (1979). *ALSCAL-4: User's guide.* Carrboro, NC: Data Analysis and Theory Associates.