TYPE I ERROR RATES FOR MULTI-GROUP CONFIRMATORY MAXIMUM LIKELIHOOD FACTOR ANALYSIS WITH ORDINAL AND MIXED ITEM FORMAT DATA: A METHODOLOGY FOR CONSTRUCT COMPARABILITY

By

KIM HONG KOH

M.A. (Clinical Psychology) National University of Malaysia (1996)

B.A. (Hons.) (Psychology) National University of Malaysia (1994)

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF EDUCATIONAL AND COUNSELING PSYCHOLOGY AND SPECIAL EDUCATION

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

November 2003

ABSTRACT

Construct comparability studies are of importance in the context of test validation for psychological and educational measures. The most commonly used scale-level methodology for evaluating construct comparability is the Multi-Group Confirmatory Factor Analysis (MGCFA). More specifically, the use of normal-theory Maximum Likelihood (ML) estimation method and Pearson covariance matrix in MGCFA has become increasingly common in day-to-day research given that the estimation methods for ordinal variables require large sample sizes and are limited to 20-25 items. The thesis investigated the statistical properties of the ML estimation method and Pearson covariance matrix in two commonly found contexts, measures with ordinal response formats (binary and Likert-type items) and measures with mixed item formats (wherein some of the items are binary and the remainder are of ordered polytomous items). Two simulation studies were conducted to reflect data typically found in psychological measures and educational achievement tests, respectively. The results of Study 1 show that the number of scale points does not inflate the empirical Type I error rates of the ML chi-square difference test when the ordinal variables approximate a normal distribution. Rather, increasing skewness lead to the inflation of the empirical Type I error rates. In Study 2, the results indicate that mixed item formats and sample size combinations have no effect on the inflation of the empirical Type I error rates when the item response distributions are, again, approximately normal. Implications of the findings and future studies were discussed and recommendations provided for applied researchers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my research supervisor, Professor Bruno D. Zumbo for the kind support, guidance and counsel he provided me, with great commitment and dedication. Without his patience, encouragement, direction, and assistance, I am not sure this dissertation would have been completed. I am in debt to him for his intellectual and emotional support throughout my writing of this dissertation. I am grateful to the committee members, Dr. Kadriye Ercikan, Dr. Anita Hubley, and the University Examiners, Dr. Kimberly Schonert-Reichl and Dr. Ann Anderson, for their inspirations, suggestions, constructive comments, and review.

I would like to thank my Lord who has guided and watched over me in my life. Finally, I thank my mother for her understanding and prayer support during the preparation of this dissertation.

CHAPTER I

BACKGROUND TO THE PROBLEM

Construct comparability or construct equivalence is a prerequisite for valid and meaningful test score comparison between groups. There are three assumptions for construct comparability.

1. The set of manifest measures (observed variables) should evoke the same conceptual or cognitive frame of reference used to make item responses in each group.
2. The pattern of the relationships between observed variables and latent variable(s) are equivalent across groups.
3. The observed variables are influenced to the same degree and perhaps by the same error variances across groups.

When these three assumptions are fulfilled, one can claim that a measurement instrument or a test has captured the same underlying latent variable across groups and the construct is being measured equivalently in two or more groups.

In the literature, construct comparability or construct equivalence is referred to by a variety of terms such as factor structure, structural, or dimensional invariance or equivalence (e.g., Jöreskog, 1971; Sireci, Bastari, & Allalouf, 1998; Sireci, Xing, & Fitzgerald, 1999; Tippets & Michaels, 1997) and measurement invariance, measurement equivalence or factorial structure invariance (e.g., Byrne, 1998; Byrne, Shavelson, & Muthén, 1989; Drasgow & Kanfer, 1985; Horn & McArdle, 1992; Reise, Widaman, & Pugh, 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). As evidence of test fairness, construct comparability is typically examined at the scale level

by looking at the equivalence or invariance of the factor structures and/or measurement models across relevant groups (e.g., gender, ethnic, language).

From the perspective of structural equation modeling (SEM), construct comparability can be examined by multi-group confirmatory factor analysis (MGCFA). The MGCFA model emphasizes testing for both measurement and structural invariances (Byrne, 1998; Jöreskog, 1971). Measurement invariance is tenable when the relations between observed variables and latent construct(s) are identical across relevant groups. In particular, individuals with the same standing on a latent variable but sampled from different subpopulations should have the same expected observed score on a test of that variable. According to Horn and McArdle (1992), measurement invariance refers to "whether or not under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (p. 117). Without measurement invariance, observed means are not directly comparable (Drasgow & Kanfer, 1985). For a strict form of measurement invariance, the observed variables are also expected to be equally reliable across groups. In addition, structural invariance implies that the structural relations among the factors are equivalent across groups (Byrne, 1998). It is worth noting that the testing for structural invariance is not needed when one has a unidimensional or a single-factor structure.

Given that construct comparability studies involving multi-group confirmatory factor analysis (MGCFA) plays an important role in establishing the validity of the inferences and group comparisons one can make from their measure, this thesis will investigate the statistical properties of this methodology in two commonly found contexts, measures with ordinal response formats (binary and Likert-type items) and

measures with mixed item formats (wherein some of the items are binary and the remainder are of an ordered polytomous format). In short, the purpose of this dissertation is to investigate the practice of using maximum likelihood factor analysis of a Pearson covariance matrix to test measurement invariance hypotheses in the framework of MGCFA. With an eye toward contextualizing and articulating the research purpose further, the remainder of this chapter will describe the scaling and statistical methodology in the context of examples of widely used measures of self-esteem and depressive symptomatology, the Rosenberg Self-Esteem Scale and the Center for Epidemiological Studies Depression Scale (CES-D). This chapter will close with the statement of the research purpose. Chapter two is the literature review and statement of the research questions deriving from the purpose, Chapter three is the research methodology and study design, Chapter four presents the results of the studies, and the closing chapter is a discussion of the findings and their implications.

Multi-Group Confirmatory Factor Analysis Methodology

Multi-group confirmatory maximum likelihood factor analysis has become the most commonly recommended scale-level technique to evaluate construct equivalence of a measurement instrument across different groups (e.g., gender, language, ability, age level) in educational, social and behavioral science, and marketing research. Many of the measurement instruments used in the aforementioned fields of research contain the following two types of ordinal-scaled items: binary/dichotomous items, and ordered polytomous items (Byrne, 1998).

The fundamental idea underlying the confirmatory factor analytic models or measurement models in multi-group confirmatory factor analysis is the use of a set of observable variables (i.e., items) to represent the latent variable(s), which in most of the cases are neither observable nor directly measurable. Typically, the latent variable(s) of interest are conceptualized as continuous (i.e., measured on interval scales) and normally distributed. When the ordinal-scaled items are used as proxies for the latent continuous variable(s), the assumptions of interval measurement scale and multivariate normality are likely to be violated.

Two commonly encountered problems associated with ordinal-scaled items are measurement scale coarseness and multivariate nonnormality. Measurement scale coarseness is caused by a crude classification (or measurement) of the latent variables to ordinal scales with small numbers of response categories. Because of the discrete nature of ordinal scales, the distributions of the response data obtained from binary items and/or ordered polytomous items are conducive to multivariate nonnormality.

Binary items are typically associated with multiple-choice items on achievement and aptitude tests that use a dichotomous scoring scheme (Koch, 1983) or statements in psychological and sociological measures that are dichotomously scored according to a scoring key (e.g., true/false, yes/no) (Zumbo, 1999). Although ordered polytomous items can also be referred to as Likert-type items, the use of these two terms is context specific. The term "ordered polytomous items" is more frequently used in the context of large-scale educational achievement assessments wherein the constructed-response items are scored using an ordered polytomous scale (i.e., partial credit scores). In the context of attitudinal and psychological measurement, questionnaire items (e.g., self-report

statements) and rating scales (e.g., bipolar semantic differential scales) with an ordered categorical response format are called "Likert-type items" and "rating scale items", respectively.

Much of the ordinal questionnaire data in the social and behavioral research are derived from a single item response format. For example, all of the items on a scale are five-point Likert scales. Mixed item format data are more often found in educational measurement wherein many achievement tests in use today are "blended" instruments that include a mixture of item formats such as binary and ordered polytomous items. The mixed item response formats are also ordinal in nature.

Conventional wisdom is that the coarseness of a measurement scale can be refined by having more response categories such as the use of five or seven response categories in the Likert-type attitudinal and psychological measures. As a result, the ordinal data may approximate a symmetric distribution. Nonetheless, Likert-type items with four scale points or less are commonly found in the social and behavioral science research. Likewise, many of the achievement tests used in the educational setting are comprised of mixed format items with a relatively small number of response categories (e.g., a test comprised primarily of binary items and a few 3-category ordered polytomous items). This implies that the use of coarse measurement scales is practically unavoidable.

According to Micceri (1989), nonnormality in the form of extreme asymmetry or lumpiness is typical in real data. He found that only 6.8% of the 440 distributions of achievement and psychological data he collected from applied research studies and standardized test databases exhibited both tail weight and symmetry approximating that of the Gaussian distribution. None of the distributions passed the Kolmogorov-Smirnov

test of normality. Micceri's findings indicate that the majority of the data in education and psychology do not follow univariate normal distributions, let alone a multivariate normal distribution. However, one could not know whether the number of scale points might have associated with nonnormality because Micceri did not report the range of the number of scale points used in those data.

Ideally, data derived from an ordinal scale should be analyzed using estimation methods that are designed for use with such data. For example, the Weighted Least Squares (WLS, Jöreskog & Sörbom, 1996) or Asymptotic Distribution Free (ADF, Browne, 1984) estimation of model parameters using the polychoric correlation and asymptotic covariance matrix is theoretically sound for ordinal and mixed item format data, especially when nonnormality of the item distributions is of great concern. In practice, ordinal data are often treated as if they were continuous. The reason for doing so will be discussed in a later section of this chapter. When ordinal data are used with normal theory based Maximum Likelihood (ML) estimation method and Pearson covariance matrix in the single-group confirmatory factor analyses, the chi-square goodness of fit statistic is generally inflated due to departures from multivariate normality in the observed variables, albeit negligible bias is found in the model parameter estimates. Hence, using the normal theory ML chi-square statistic as a measure or formal test statistic of model-data fit under the conditions of multivariate nonnormality will lead to an inflated Type I error rate for rejecting a true model. The impact of analyzing ordinal data with the ML estimation method and Pearson covariance matrix in the framework of MGCFA is yet to be investigated. Given that the use of mixed item formats has become increasingly important in educational measurement, what are the consequences of

ignoring the categorical nature of the mixed item format data when the normal theory ML estimation method and Pearson covariance matrix are applied to such data in MGCFA?

Ordinal Variables, Measurement Scale Coarseness, and Multivariate Nonnormality in the Context of an Example

In order to motivate the discussion of the statistical and psychometric theory, let us consider the following example of a four-point Likert item, which is taken from the Rosenberg Self-Esteem Scale (Rosenberg, 1965, 1979): "I am able to do things as well as most other people." The item responses are scored on a 4-point scale such as (1) Strongly Disagree, (2) Disagree, (3) Agree, and (4) Strongly Agree. This item, along with other items, serve as a set of observed ordinal variables, *x*s, to measure the latent continuous variable x^*, namely self-esteem. For each observed ordinal variable x, there is an underlying continuous variable x^*. If x has m ordered categories, x is connected to x^* through the non-linear step function: $x = i$ if $\tau_{i-1} < x^* \leq \tau_{i,}$ $i = 1,2,3,...,m$, where $\tau_0 = -\infty, \tau_1 < \tau_2 < \tau_3 < ...\tau_{m-1}$, and $\tau_m = +\infty$ are parameters called threshold values. For a variable x with m categories, there are *m-1* unknown thresholds (Jöreskog & Sörbom, 1996). Given that the above item has four response categories, there are three thresholds with the latent continuous variable.

Figure 1 depicts the threshold model for an ordinal variable x that has three response categories and two thresholds with item thresholds for x^*: tau 1(τ_1), tau 2 (τ_2) (values of –1 and 1). The model assumes that underlying the ordinal variable is a continuous variable that determines the categories of x as it crosses different thresholds. In other words, the thresholds serve as decision criteria that elicit examinees' responses to a particular category on an ordinal scale. If x^* is less than τ_1, x is in category one, for

$\tau_1 < x^* \leq \tau_2$, x is in category two, and if x^* exceeds τ_2, x is in category three. It is important to note that the actual intervals or distances between these adjacent categories are generally unknown and need not be equal. Taking the self-esteem item as an example, we cannot say that a person who responds to "Strongly Agree" is *twice* more likely than a person who responds to "Disagree" to have a high self-esteem. Rather, we can only infer that a person who responds to a higher category (i.e., Strongly Agree) has higher self-esteem than a person who responds to a lower category (i.e., Disagree).

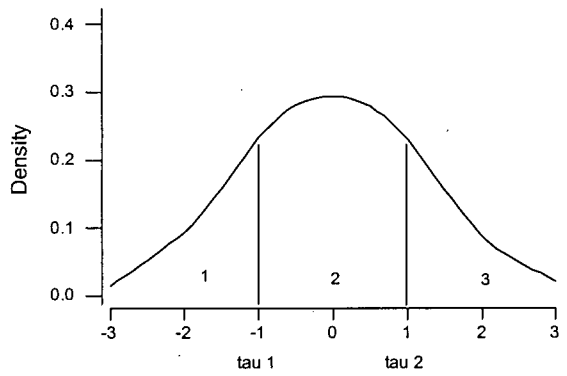


Figure 1. A three category, two threshold x and its corresponding x^*.

When a latent continuous variable is represented by its coarsely categorized version, imperfection of scaling can lead to the violations of normality assumptions in the categorized data. There is a general consensus in the measurement community that dichotomizing a normally distributed continuous variable or collapsing it into three or more categories can lead to the loss of information (i.e., response reduction) and results in the attenuation of Pearson correlations or covariances (Bollen & Barb, 1981; Cohen, 1983). The loss of information is also attributed to a nonreversible transformation of the

original continuous distribution to the ordinal/Likert distribution. As a result, the shape of the distribution of the ordinal variable may differ sharply from that of the latent continuous variable. Compared to the infinite values of a continuous variable, a collapsed variable (i.e., ordinal variable) takes on a relatively small number of possible values on the continuum. Furthermore, unequal threshold discretization may take place. Hence, the scale points of the ordinal variable are discrete rather than continuous. Discrete ordinal scales with unequal thresholds would produce nonnormally distributed ordinal variables.

The Statistical Model

The general structural equation model consists of two components: the measurement model and the structural model. Confirmatory factor analysis makes use of only the measurement model. In an ideal situation, the measurement model is: $x^* = \Lambda\xi + \delta$, where x^* is a $p \times 1$ vector of continuous indicators of latent variables (i.e. multinormally distributed observed variables), Λ is a $p \times n$ regression coefficients matrix (a.k.a. factor loadings matrix) that relates n latent factors to each of the p observed variables designed to measure them, ξ is a $n \times 1$ vector of latent variables or factors, and δ is a $p \times 1$ vector of errors of measurement of x^*. It is assumed that ξ are uncorrelated with δ. Also, the errors of measurement are uncorrelated, that is, the covariance matrix of δ is diagonal. In multi-group confirmatory factor analysis, researchers are interested in finding out whether or not the same measurement model is invariant across subgroups. It is expected that the estimated or implied covariance matrix $\hat{\Sigma}_g$, is as close as possible to the sample covariance matrix S_g: $S_g \approx \hat{\Lambda}_g \hat{\Phi}_g \hat{\Lambda}_g' + \hat{\Psi}_g = \hat{\Sigma}_g$ across groups, where $\hat{\Lambda}_g$ is a $p \times n$ matrix of loadings of the p measured variables on the n latent variables, $\hat{\Phi}_g$ is a n

$\times n$ matrix of covariances among the latent variables, $\hat{\Psi}_g$ is a $p \times p$ matrix of covariances among the measurement errors, and g denotes group membership.

Maximum Likelihood Estimation Method

To date, ML is the default estimation method in almost all of the computer programs designed for MGCFA. The ML fitting function $F[S;\Sigma(\theta)]$ measures how close a given Σ is to the sample covariance matrix, S. In other words, to find the values of $\hat{\Lambda}_g$, $\hat{\Phi}_g$, and $\hat{\Psi}_g$ that generate an estimated covariance matrix, $\hat{\Sigma}_g$ that is as close as possible to the sample covariance matrix S_g. The ML fitting function is defined as $F_{ML}[S;\Sigma(\theta)] = tr(S\Sigma^{-1}) + [\log|\Sigma| - \log|S|] - q$, where Σ is the implied covariance matrix, S is the sample covariance matrix, and q is the number of observed variables. If S and Σ are equal, the fitting function will equal zero (Long, 1983). In MGCFA, the ML estimation produces a chi-square statistic for each hypothesis testing of invariance. The tenability of an invariance hypothesis is determined by the statistical significance of the chi-square difference test (i.e., change in χ^2/change in *df*) between two nested models. For example, a nonsignificant chi-square difference test statistic derived from two nested models (i.e., baseline model versus full measurement invariance model) indicates that the full measurement invariance hypothesis is tenable.

Violations of Measurement Scale and Multivariate Normality Assumptions

As discussed earlier, coarsely categorized observed variables violate the measurement scale assumption of the latent continuous variables. According to Bernstein and Teng (1989), there are two different effects of categorization: one due to

categorization per se (number of scale points) and the other due to differential categorization (differential item distributions). The ML-based χ^2 statistics are highly sensitive to categorization effects when the item distributions differ. In addition, single-group CFA studies found that the number of response categories has little effect on the χ^2 likelihood ratio test of model fit when categorical variables approximate a normal distribution. However, increasing skewness, particularly differential skewness (items skewed in different directions) can lead to inflated χ^2 values (Byrne, 1998).

Given that ordinal-scaled items are discrete rather than continuous indicators of the latent variables, data derived from the ordinal-scaled items do not conform to a multivariate normality distribution. Hence the measurement model for x^* does not hold for the ordinal-scaled items. Moreover, measurement errors induced by a crude categorization of the latent continuous variables may lead to the violations of the covariance structure. Because the Pearson correlation or covariance is attenuated in the ordinal variables, the covariance structure model may hold for the latent variables, but not generally for the observed variables. The normal theory ML estimation method assumes that the observed variables are distributed as multivariate normal. Therefore, ML estimation based on the distorted sample covariance matrix is likely to be biased. Similar to correlations, the differences between the covariance structures become less significant when five or more categories are used and the marginal distributions become similar (Bollen, 1989).

Tests of Measurement Invariance

The methodology for testing the factor structure invariance of a measurement instrument across groups originates from Jöreskog's (1971) work in "simultaneous factor

analysis in several populations" (SIFASP which is equivalent to MGCFA). Within the Jöreskog tradition, tests of factorial invariance begin with a global test of the equality of covariance structures (i.e., matrices) across groups (i.e., $H_0 : \Sigma_1 = \Sigma_2 = ...\Sigma_G$, where G denotes the number of groups). Failure to reject the null hypothesis is interpreted as evidence of factorial invariance across groups; except for mean structure, the groups can be treated as one. Contrariwise, rejection of the null hypothesis leads to testing a series of increasingly restrictive hypotheses in order to identify the source of non-invariance. These hypotheses relate to the invariance of (a) the factor loadings (i.e., $H_0 : \Lambda_1 = \Lambda_2 = ...\Lambda_G$), (b) the error or uniquenesses (i.e., $H_0 : \Theta_1 = \Theta_2 = ...\Theta_G$), and (c) the factor variances and covariances (i.e., $H_0 : \Phi_1 = \Phi_2 = ...\Phi_G$). The tenability of Hypothesis (a) is a prerequisite to the testing of Hypotheses (b) and (c).

In seeking evidence of construct equivalence of a measurement instrument in the MGCFA framework, applied researchers are typically interested in testing for the equivalencies of the following parameters simultaneously across groups: (1) factor loadings (λs), (2) error variances (δs), and if one has more than one factor, (3) factor variances-covariances (Φs). These sets of parameters are tested in an increasingly restrictive way, namely from a weak baseline model with no between group constraints to a strict or full invariance model in which all the above mentioned parameters are constrained to be equal across groups. It is important to note that the evaluation of the construct equivalence of a measurement instrument with a unidimensional construct or a single factor involves only the testing for measurement invariance, implying that the factor loadings (lambda) and error variances (theta-delta) are of primary interest.

Prior to the testing for the invariance of the specific parameters, researchers who follow the Jöreskog tradition will begin with the testing for a global test of the equality of covariance matrices ($\sum^{g} = \sum^{g'}$). According to Byrne (1998) and Muthén (1988), this practice is not necessary because the global test often leads to contradictory findings with respect to equivalencies across groups. For instance, sometimes the global null hypothesis is found tenable, yet subsequent tests of hypotheses related to the invariance of particular measurement or structural parameters must be rejected (Jöreskog, 1971). In contrast, the global null hypothesis may be rejected, yet tests for the invariance of measurement and structural invariance hold (Byrne, 1988).

Implications of Treating Ordinal Data as Interval Data

In factor analysis, the use of interval scale-based correlation measure such as the Pearson product moment correlations with ordinal data results in differential attenuation in the correlations among the observed item responses that affect the factor solution. In MGCFA, bivariate normality between pairs of observed variables or items is an important assumption for the computation of the Pearson covariance matrices. When the distributional properties of the observed variables are neither bivariate nor multivariate normally distributed within groups and/or between groups, the use of Pearson covariance measure is expected to cause distortions to the covariance matrices, which are the key input for the ML estimation. As a result, the ML estimation would yield inflated chi-square values for the hypothesis tests of measurement invariance. Therefore, the chi-square difference test between two nested models that is used to make statistical decisions about the tenability of an invariance hypothesis cannot be trusted.

To provide some context for the discussion of the impact of treating ordinal data as interval data in MGCFA, let us consider a Likert-type measure used in the life and social sciences: the Center for Epidemiological Studies Depression Scale (CES-D; Radloff, 1977). The CES-D consists of 20 Likert-type items that are used as indicators of depression symptoms (i.e., latent variable or construct). The item response format is a 4-point Likert scale. Data were collected by the UNBC Institute for Social Research and Evaluation[1] in a general population health survey from 310 adult males (age: range 17-82, mean = 46.1, standard deviation = 12.1), and 290 adult females (age: range 18-87, mean = 42.2, standard deviation = 13.4 years) who resided in communities in Northern British Columbia. The items are presented as follows:

For each statement, circle the number (see the guide below) to indicate how often you felt or behaved this way **during the past week**.

0 = rarely or none of the time (less than 1 day)

1 = some or a little of the time (1-2 days)

2 = occasionally or a moderate amount of time (3-4 days)

3 = most or all of the time (5-7 days)

	not even 1 day	1-2 days	3-4 days	5-7 days
1. I was bothered by things that usually don't bother me.	0	1	2	3
2. I did not feel like eating; my appetite was poor.	0	1	2	3
3. I felt that I could not shake off the blues even with help from my family or friends.	0	1	2	3

[1] I would like to thank Professor Alex Michalos and Professor Bruno Zumbo for making this data available to me to use in this demonstration.

4. I felt that I was just as good as other people.	0	1	2	3
5. I had trouble keeping my mind on what I was doing.	0	1	2	3
6. I felt depressed.	0	1	2	3
7. I felt that everything I did was an effort.	0	1	2	3
8. I felt hopeful about the future.	0	1	2	3
9. I thought my life had been a failure.	0	1	2	3
10. I felt fearful.	0	1	2	3
11. My sleep was restless.	0	1	2	3
12. I was happy.	0	1	2	3
13. I talked less than usual.	0	1	2	3
14. I felt lonely.	0	1	2	3
15. People were unfriendly.	0	1	2	3
16. I enjoyed life.	0	1	2	3
17. I had crying spells.	0	1	2	3
18. I felt sad.	0	1	2	3
19. I felt that people dislike me.	0	1	2	3
20. I could not get "going".	0	1	2	3

Note: - Items 4, 8, 12, and 16 are reverse coded.

In seeking evidence of construct equivalence or measurement invariance of the CES-D across gender, it is not uncommon for researchers to treat the Likert variables as

if they were continuous for the computation of Pearson correlation or covariance matrices and then proceed with the use of normal theory ML estimation method in MGCFA.

Because all the items in both gender groups yielded similar response distributions, taking into account reverse coding, only the frequency charts for item 1 are presented in Figures 2 and 3. The frequency charts of the remaining items for each gender group can be found in Appendix A.

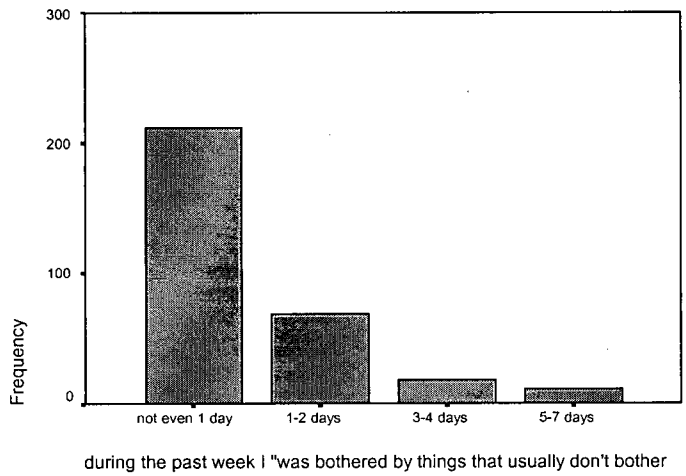


Figure 2. Distribution of responses on CES-D item 1 (I was bothered by things that usually don't bother me) for males.

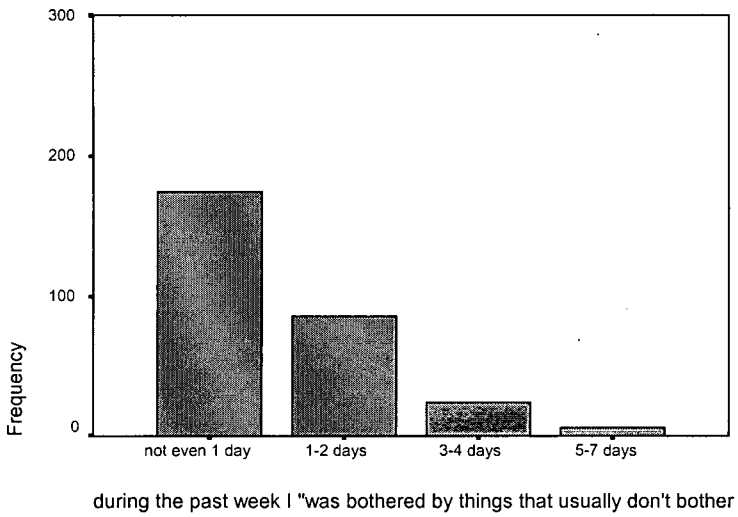


Figure 3. Distribution of responses on CES-D item 1 (I was bothered by things that usually don't bother me) for females.

As in item 1, the distributions of all the Likert-type item responses are positively skewed[2]. Nonnormality is present not only within group but also across groups. Clearly the assumptions of univariate, bivariate and multivariate normality are violated. If the CES-D data are treated as if they were continuous, the tests for multivariate normality revealed the following statistics for skewness (males, $z = 68.58$, $p = .000$; females, $z = 44.25$, $p = .000$), kurtosis (males, $z = 24.31$, $p = .000$; females, $z = 19.80$, $p = .000$), and for third and fourth moments considered jointly (males, $\chi^2[2, N = 310] = 5294.81$, $p = .000$; females, $\chi^2[2, N = 290] = 2349.79$, $p = .000$), indicating that the assumption of multivariate normality is seriously violated.

Although a few options have been suggested for dealing with ordinal variables and nonnormality in SEM/CFA, each option has its own limitations (see Table 1).

[2] Note that, as expected in a general population survey rather than a clinical population, the item responses provided by the survey respondents for items 4, 8, 12, and 16 will be negatively skewed (as seen in Appendix A). However, these data, and the meaning of the survey questions, are reverse coded before the psychometric analysis so that all the items are interpretable in the same scale direction – a larger scale score value meaning more depressive symptomatology.

Table 1

Different Options for Conducting MGCFA

Item Level	*Correlation Measure*	*Estimation Method*			
		Normal theory Maximum Likelihood[1]	*Weighted Least Squares/ADF*[2] *with Asymptotic Covariance Matrix or Diagonally Weighted Least Squares with Asymptotic Variance Matrix*	*Categorical Variable Methodology*	*Satorra-Bentler Scaled Chi-Square*[3]
Items	Pearson Covariance Matrix	√ (Byrne, 1998)			Asymptotic Covariance Matrix requires large *N*
	Polychoric Correlation	Not recommended for small *N*	Limited to approximately 20-25 items and Not recommended for small *N*	Limited to 5 response categories; Not suitable for large model sizes and small *N*	
Item Parcels	Pearson Correlation	False factor structure/ Dimension			

Note. [1] Generalized Least Squares method typically has the same function and outcome as ML except that when models are misspecified (Olsson, Foss, Troye, & Howell, 2000), so this method is not included in the table. [2] Asymptotic Distribution Free method is equivalent to the Weighted Least Squares. [3] Satorra-Bentler Scaled Chi-Square is equivalent to the ML-Robust. *N* = Sample sizes.

I will describe the various options depicted in Table 1. Many researchers will face a dilemma while considering what is the most appropriate estimation method for

MGCFA. Given that the CES-D data are extremely skewed, if researchers proceed with their analyses using the interval-level Pearson covariance matrices and the normal theory ML estimation method, serious distortions are expected to occur in the standard errors of parameter estimates and chi-square goodness-of-fit statistics. Consequently, results obtained from the testing for the various hypotheses of measurement invariance may not be valid. Many researchers will choose to use the appropriate correlation measure, that is, the polychoric correlation along with its corresponding asymptotic covariance matrix and proceed with the Weighted Least Squares (WLS) or ADF estimation method. Another alternative is the use of the Diagonally Weighted Least Squares (DWLS) estimation method, which requires the computation of the polychoric correlation and the asymptotic variance matrix. However, all these methods are not feasible due to the small sample sizes in the CES-D data. Compared to the ML estimation method, the WLS/ADF, and DWLS require relatively large sample sizes, i.e., at least 2,000-5,000 observations per group (Browne, 1984). The use of ML with the polychoric correlation is rare in practice and again the requirement of large sample sizes may prohibit researchers from choosing this option. It is worth noting that, as input to the LISREL, the polychoric correlation also has higher rates of nonconvergent and improper solutions due to nonpositive-definite matrices (e.g., Heywood cases).

The use of the ML estimation method and polychoric correlation is more accurate than the ML-Pearson correlation combination in terms of pairwise correlations, parameter estimates and estimated standard errors. However, the polychoric correlation performs worse than the Pearson correlation on the ML chi-square statistics and other practical fit indices such as goodness-of-fit index (GFI) and adjusted goodness-of-fit index (AGFI),

and root mean square residual (RMR), leading to frequent rejections of a correctly specified model (Babakus, Ferguson, & Jöreskog, 1987). Using ADF and Categorical Variable Methodology (CVM, Muthén, 1984), Muthén and Kaplan (1992) and Potthast (1993) found that the chi-square values were still inflated when observed variables were based on nonnormal ordered categorical data, particularly when models were large and sample sizes were small.

The CVM is limited to small model sizes and also needs large sample sizes for the computation of the asymptotic covariance matrix. Furthermore, it makes a very strong assumption that underlying each categorical observed variable is an unobserved latent counterpart that has a continuous scale and these latent variables are assumed to be multivariate normally distributed (Bentler & Wu, 1995). This strict assumption is hard to follow because we can only assume that the latent variables are continuous and multivariate normally distributed, but in reality, the distributions of the latent variables are unknown and may be nonnormally distributed.

The final option is the use of item parcels. By using this technique, researchers can combine the nonnormally distributed categorical items into item parcels, which result in data with more data points and have distributions that are more continuous and normally distributed. This technique can also solve the problem of large numbers of items because its use will allow fewer parameters to be estimated and thus produces more stable results than the item level analyses. However, the usefulness of item parcels is offset by the obfuscation of the true factor structure or dimension (Bandalos, 1999).

Clearly, the requirement of large sample sizes in the estimation methods for ordinal variables poses problems to researchers because CES-D data tend to have

relatively small sample sizes; in this case there are 310 and 290 for males and females, respectively. ML is more suitable for small sample sizes and can handle large numbers of items. Hence, the option of using ML and Pearson covariance matrix seems to be the most viable option in this case. Some researchers have recommended the use of the Satorra-Bentler (SB) scaled chi-square in dealing with data nonnormality, but its utility in MGCFA is still unknown and it is computationally more intensive because it requires the computation of an asymptotic covariance matrix. Also, if the skewness of the distributions varies dramatically across items, the SB scaled chi-square statistic is likely to yield spurious factors (Green, Akey, Fleming, Hershberger, & Marquis, 1997).

Given the above arguments, MGCFA was conducted on the CES-D data by using the ML estimation method and the Pearson covariance matrix. As Cudeck (1989) pointed out, the analysis of correlation matrices may result in the following problems in MGCFA: (1) modify the model being analyzed, (2) produce incorrect χ^2 and other goodness-of-fit measures, and (3) give incorrect standard errors. These reasons justify the use of the Pearson covariance matrix. The MGCFA results of the cross-gender measurement invariance were presented in Table 2.

Table 2

Multi-group Confirmatory Factor Analysis Results for the Measurement Invariance of CES-D across Gender

Hypothesis	χ^2	*df*
Baseline Model[1] (No between-group constraints)	1208.74	340
Strong Invariance[2] (Number of factors and Factor loadings invariant)	1264.50	360
Full Invariance[3] (Number of factors, Factor loadings, and Error variances invariant)	1396.60	380

Note. [1] Configural invariance. [2] Configural invariance and Metric invariance. [3] Configural invariance, Metric invariance, and Item Uniqueness invariance.

From Table 2, the difference in chi-square values between the baseline model and the strong invariance model is statistically significant, $\Delta\chi^2 = 55.76$, $\Delta df = 20$, $p = .000$, indicating that the hypothesis of strong invariance is not tenable. The difference in chi-square values between the baseline model and the full invariance model is also statistically significant, $\Delta\chi^2 = 187.86$, $\Delta df = 40$, $p = .000$, indicating that the hypothesis of full invariance is not tenable. Clearly, the hypotheses of measurement invariance are rejected but, because the Likert variables are extremely nonnormal in both gender groups,

these hypothesis tests of measurement invariance are operating at unknown Type I error rates.

Problem Statement

The question of to what extent the ML estimation method is robust to the violations of measurement scale and multivariate normality assumptions in MGCFA is yet to be answered. Specifically, what are the Type I error rates (likelihood or probability of rejecting true measurement invariance models) of the ML chi-square difference test when ordinal data are analyzed with the normal theory ML estimation method and Pearson covariance matrices across different numbers of scale points and response distributions? It is also of interest to examine whether the use of large numbers of scale points and mixed item formats can compensate for multivariate nonnormality and thus reduce the Type I error rates of the ML estimation method in MGCFA. As Zumbo and Zimmerman (1993) pointed out, the shape of distribution is a better criterion than the measurement level when one decides whether to use parametric or nonparametric statistical methods. Hence, it is important to examine to what extent the shape of the response distribution can justify the use of ML estimation method with ordinal and mixed item format data in MGCFA.

To address the above problems, the effects of the number of scale points, the mixtures of item formats, the distribution of the responses across groups, and the sample size combinations on the Type I error rates of ML-based chi-square difference tests for two hypotheses of measurement invariance (i.e., strong and full invariance) were investigated in the current study. Although Type II error and power issues in multi-group confirmatory maximum likelihood factor analysis are important, the focus of the current study was on Type I error because the Type I error rates of any statistical test need to be

established before one can even turn to the issue of statistical power (and hence Type II error rates).

It is important to note that the research literature on single-group confirmatory factor analysis does not give us precise direction on the use of the chi-square test for MGCFA. That is, the single-group situation is of limited generalizability because the MGCFA involves a test statistic that is the difference between two chi-square test statistics (of nested models), each of which could be biased but may result in an acceptable difference of chi-square of the MGCFA – that is, the single-group biases may cancel out in the MGCFA.

Significance of the Current Study

The robustness properties of various parametric tests such as *t*-test, analysis of variance, bivariate correlation, multiple regression, and single-group CFA have been researched extensively in the published literature. Measurement scales used in education and the social and behavioral sciences are typically ordinal in nature. Researchers tend to use ordinal-scaled variables in statistical procedures that assume that these variables possess an interval scale of measurement. Most of the MGCFA applications to date have been concerned with items or observed variables measured on an ordinal scale. Given that the ML estimation method with Pearson covariance matrix is the most commonly used approach in MGCFA, it is important for the current study to investigate the robustness properties of the ML estimation method when the measurement scales are violated due to the use of small numbers of scale points and unequal threshold discretization of ordinal Likert scales.

Popular SEM texts such as Byrne's (1998) *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming* and

Jöreskog and Sörbom's (1996) *LISREL 8: User's Reference Guide* recommend the use of ML estimation method with the Pearson covariance matrix. Byrne (1998, p. 137) contended that the use of ML is acceptable when one has Likert-type item responses with large numbers of response categories. Jöreskog and Sörbom (1996) stated, "If the sample size is not sufficiently large to produce an accurate estimate of the asymptotic covariance matrix (in WLS), it is probably better to use ML or GLS" (p. 239). Nevertheless, no MGCFA studies to date have investigated the optimal number of response categories and the sample size requirement for use with the ML estimation method in the presence of ordinal and mixed item format data. A thorough investigation of these important issues in the current study is deemed necessary to justify the use of ML estimation method with ordinal and mixed item format data in MGCFA.

A commonly encountered problem among the MGCFA researchers is that the use of large numbers of items in a measurement instrument may prohibit them from using the ordinal-scale estimation methods such as WLS/ADF and CVM. For multi-group conditions, the WLS/ADF and CVM estimation methods require a huge capacity of computer memory because the computation of the asymptotic covariance matrices is heavy. The use of CVM is also restricted to five response categories. In addition, none of the ordinal-scale estimation methods perform better than the ML estimation method in terms of the chi-square statistics when the model sizes are large in single-group CFA. Given the limitations of the ordinal-scale estimation methods, the use of ML in MGCFA seems to be the only convenient solution. David Kaplan (SEMNET, 1996) has postulated that the problems in single-group CFA might also occur in MGCFA. He pointed out that a dissertation examining the effects of multivariate nonnormality in MGCFA might be

worthwhile. To my knowledge, such a research topic has not been studied. Hence, it is important for the current study to examine the effects of multivariate nonnormality in the MGCFA.

The measurement imprecision of a construct is obviously serious when the continuous latent variable is measured using ordinal scales. However, the problem is not due to the use of ordinal measurement per se but the common practice of treating ordinal data as if they were measured at the interval level and analyzing such data with statistical methods which assume interval measures. In many of the single-group CFA studies, the use of normal theory ML estimation method and Pearson covariance matrix with ordinal data was found to result in substantial distortions of standard errors of parameter estimates and model-data fit indices. The impact of treating data derived from the ordinal scales such as binary and Likert-type scales as continuous in the MGCFA is worth investigating.

This study also focuses on the impact of conducting multi-group confirmatory maximum likelihood factor analysis with mixed item format data. A systematic investigation of this issue is deemed to be important given that the assessment of educational outcomes since the 1990s has increasingly taken the form of measurement instruments that combine multiple-choice items with constructed-response items.

In summary, MGCFA are typically carried out on binary and Likert-type variables in measurement invariance research. In the applications of MGCFA for the testing of measurement invariance, it has been common practice for researchers to treat Likert data as if they were continuous for the computation of the Pearson covariance or correlation matrices and for use with the normal theory-based ML estimation method (e.g., Bandalos

& Benson, 1990; Byrne, 1994; Byrne, Shavelson, & Muthén, 1989; Drasgow & Kanfer, 1985; Reise, Widaman, & Pugh, 1993; Sireci, Xing, & Fitzgerald, 1999; Steenkamp & Baumgartner, 1998; Tippets & Michaels, 1997). It is important to note that the ML estimation method makes the implicit assumption that the observed variables have a multivariate normal distribution in the population. Likert-scaled items could hardly follow a multivariate normal distribution. Many of the MGCFA studies have failed to report a test of multivariate normality in justifying of the use of the ML estimation method. Using data simulation, the Type I error rates of rejecting a true measurement invariance model based on the chi-square difference tests produced by the ML estimation method in the current study will inform the applied researchers the operating characteristics of the ML estimation method in the presence of ordinal and mixed item format data.

CHAPTER II

LITERATURE REVIEW

This chapter begins with a review of studies that have examined the consequences of treating ordinal data as if they were continuous and analyzing with the normal theory statistical methods. More specifically, the common psychometric problems of treating ordinal data as interval data in Pearson correlation, multiple regression, and single-group confirmatory factor analysis; and the impact of the number of Likert scale points on reliability and validity are reviewed. The issue of analyzing ordinal and mixed item format data with multi-group confirmatory maximum likelihood factor analysis has not been studied before. Hence, the single-group CFA literature is reviewed with an eye toward the study design variables: number of scale points or response categories, shape of distribution, model size, size of parameters, model specification, sample sizes, and the combination of estimation method and correlation measure. This informs the study design used in the current study. A thorough review of the different estimation methods is deemed to give some methodological contribution to the readers who are interested in pursuing research in CFA. Toward the end of the review, several research questions are raised for the current study.

With the advent of Stevens' (1946) levels of measurement, the applications of the conventional parametric tests to data obtained from ordinal-scaled measures such as binary and Likert-type items have generated vigorous debates. A continuing concern is that ordinal data are discrete and multivariate nonnormal but the parametric statistical methods assume multivariate normality. A substantial body of research has addressed the problems of using observed ordinal variables to measure the underlying continuous

variables, especially the risks involved in treating the ordered categorical or ordinal data (e.g., Likert data) as if they were continuous. Almost all the studies have used computer simulation data for which the true values were known so that the estimated values can be compared to the true values. Discrepancy between the true and estimated values allows the determination of the amount of biases. The impact of using ordinal data with several parametric statistical methods was examined by varying the number of scale points and the form of distributions.

Ordinal Data with Pearson's Correlation Measure

Measurement bias or error caused by using ordinal data obtained from the Likert items in Pearson's correlation has long been debated in psychology and sociology (e.g., Bollen & Barb, 1981; Labovitz 1970; O'Brien, 1979). Labovitz (1970) has strongly advocated for treating ordinal data as interval because he found that the average Pearson correlation between the true scoring system for continuous data and the assigned scoring system for rank-ordered data, especially the equal distance scoring system was quite high. However, Labovitz's findings are valid only for uniform and normal underlying distributions. By looking at various forms of distributions and sample sizes, O'Brien (1979) found that transformation errors resulting from using the equal distance scoring system were very severe when the underlying distribution was neither uniform nor normal. For the condition of nonnormal underlying distribution, the value of average Pearson *r* decreased with increasing sample size. O'Brien also found that the Pearson *r* between an interval variable and its categorized version was not a monotonic function of the number of categories used to rank the data. Instead, *r* was a decreasing function of the number of categories. Similar nonmonotonic relation appeared in the correlation of a collapsed variable with itself.

In contrast to O'Brien (1979), Bollen and Barb's (1981) simulation studies have shown that Pearson's *r* between two different collapsed variables was a monotonic function of the number of categories. The number of categories and the strength of the relations between the continuous variables were varied in their study. Each normally distributed variable was collapsed into a number of categories ranging from two to ten based on equal latent thresholds. Each pair of normally distributed variables was constructed to correlate at one of five magnitudes: 0.2, 0.4, 0.6, 0.8, or 0.9. The difference between the average correlation for the continuous and the collapsed variables was negligible when five or more categories were used. Similarly, the differences between the original continuous correlation and the reproduced continuous correlation converged at five or more categories regardless of the magnitude of the original correlation. Furthermore, the standard deviations of the correlations for the collapsed variables were much larger than the standard deviations of the correlations of the original variables. With five or more categories the standard deviations of the correlations for the collapsed and continuous variables were close.

Number of Likert Scale Points, Reliability and Validity

The optimal number of rating categories for Likert scale items has been a focus in the construction of rating instruments. The use of too many rating categories may limit the rater's power of discrimination (e.g., raters are confused by too many choices). In contrast, a coarse scale with too few rating categories will limit the choices of the raters (e.g., raters find that none of the choices on the scale represent their rating). With regard to reliability (internal consistency and interrater reliability), many researchers contended that the optimal number of scale points or categories to maximize reliability was seven (e.g., Cicchetti, Showalter & Tyrer, 1985; Nunnally, 1967; Ramsay, 1973; Symonds,

1924). When more than seven categories were used, the increase of internal consistency reliability was negligible. Some other researchers have suggested the use of 5-point, 4-point, or 3-point scales in maximizing the reliability (e.g., Bendig, 1954a, 1954b; Jenkins & Taber, 1977). According to Cronbach (1950), there is no merit to increasing reliability of an instrument unless its validity is also increased at least proportionately. Komorita (1963) and Komorita and Graham (1965) found that the utilization of a dichotomous scale would not significantly decrease the reliability of the information obtained when compared to that obtained from a multi-category scale. In addition, they suggested that the use of a two-point response scale could eliminate or minimize an extreme response set. Both Cronbach (1950) and Komorita and Graham (1965) stated that the ultimate criterion in determining the optimal number of scale points is the effect that a change in the number of scale points has on the validity of the scale.

Matell and Jacoby (1971) and Chang (1994) have systematically examined the relationship between the number of scale points and both the reliability and validity. In the Matell and Jacoby study, Likert scale points ranging from 2 to 19 were used to investigate the effects of number of scale points on both the reliability and validity. The test-retest reliability and internal consistency were found to be independent of the number of scale points. As with reliability, validity was not affected by the number of scale points even after correcting the predictive and concurrent validity coefficients for criterion attenuation. The findings were consistent with those reported by Bendig (1954a) and Komorita and Graham (1965).

Chang (1994) used CFA models in relation to a multitrait-multimethod (MTMM) covariance matrix to examine the impact of 4-point and 6-point Likert-type scales on

internal consistency reliability and criterion-related validity, respectively. The following measurement models were nested models, which allow the comparisons of their model-data fit indexes (i.e., goodness-of-fit indexes):

i. A null model (a no-factor model) in which only the error/uniqueness variances were estimated.

ii. A simple CFA model that included the estimation of the factor loadings, trait correlations, and error variances. This model tested the hypothesis that covariation among the observed variables was due only to trait factor and their intercorrelations. Acceptance of this model would support for the equivalence of the 4-point and 6-point Likert-type scales, indicating that items measured by the two scale formats were congeneric indicators of the same traits. A second simple CFA model was a tau-equivalence model wherein the factor loadings corresponding to the same traits were constrained to be equal.

iii. Two MTMM models with the addition of two method factors corresponding to the 4-point and 6-point scales. Acceptance of these two models and rejection of the previous two models would indicate the presence of a method effect due to different number of scale points. As in ii, both congeneric and tau-equivalence constraints were applied to the parameters.

iv. Three models which estimated three traits and one method factor, instead of two method factors as was in the previous two models. The same tau-equivalence constraint used in the second model in iii was used here. In the first model, one common method factor was parameterized. Comparing the second model in iii with this model would determine whether reliability and

validity were affected differently by the 4-point and 6-point scales or if the two scales had the same contamination. In the other two models, one method factor was estimated for items with the 4-point scale and 6-point scale, respectively. Comparing these two models would answer the question of which of the scale formats had less method contamination.

v. The nine items with the 4-point scale loaded onto three trait factors, whereas the nine items with the 6-point scale loaded onto another set of three trait factors. The three traits were correlated in each set of items. The items used with the 4-point and 6-point scale formats measured different traits.

The MTMM covariance matrix was analyzed using ML estimation method. The chi-square/df ratio test was used as one of the goodness-of-fit indexes. Method variance due to number of scale points represents systematic error. If it is left unaccounted for in the separate CFA, the internal consistency reliability can be artificially high. This measurement artifact was found to have affected the 6-point scale more than the 4-point scale. By using the MTMM, systematic error due to method variance can be accounted for. When the method variance was factored out from the trait variance, Chang found that both the reliability and the validity (i.e., heterotrait-monomethod/HTMM correlations) were substantially reduced for the 6-point scale. The 4-point scale had higher reliability than the 6-point scale within the MTMM framework. The number of scale points in a Likert scale affects internal consistency reliability and HTMM validity but not heterotrait-heteromethod/HTHM validity (or criterion-related validity).

By using the CFA MTMM models, Chang's study has given a new insight into the methodology of investigating the number of Likert scale points in relation to internal

consistency reliability and criterion-related validity. However, the study used only one set of real data and the sample size was relatively small (N = 165).

Ordinal Data with Multiple Regression

Ochieng (2001) investigated the implications of using Likert data in multiple regression. He found that the largest bias in the estimation of the model R-squared, the relative Pratt Index, and Pearson correlation coefficients occurred for two or three-point Likert scales. However, the bias did not substantially reduce any further beyond the four-point Likert scale. Interestingly, type of correlation matrix had no effect on the model fit. Skewed response distribution was found to result in large biases in both R-squared and Pearson correlation, but not in the Relative Pratt Index.

Ordinal Data with Single-Group CFA

Several studies have shown that the use of ordinal data such as Likert-type data can introduce biases to the standard errors of estimates, chi-square fit statistics, and practical fit indexes in the single-group CFA (Babakus, Ferguson, & Jöreskog, 1987; Boomsma, 1983; Curran, West, & Finch, 1996; DiStefano, 2002; Dolan, 1994; Finch, West, & MacKinnon, 1997; Green, Akey, Fleming, Hershberger, & Marquis, 1997; Hutchinson & Olmos, 1998; Muthén & Kaplan, 1985, 1992; Olsson, 1979; Potthast, 1993; Rigdon & Ferguson, 1991). In this section, the variables that had been examined in the Monte Carlo computer simulation studies of single-group CFA were reviewed.

Number of Scale Points

The majority of the studies have focused on a 5-point Likert scale (Babakus et al., 1987; DiStefano, 2002; Hutchinson & Olmos, 1998; Muthén & Kaplan, 1985, 1992; Potthast, 1993; Rigdon & Ferguson, 1991). There was a consensus in the CFA simulation studies that the optimal number of scale points to minimize biases in parameter estimates,

standard errors of estimates, chi-square statistics, and practical fit indexes was five. As Pearson's correlation matrix is an integral component of the parameter estimation in LISREL, the attenuation of Pearson correlation coefficients can seriously affect the CFA results. Attenuation was found to occur with observed variables having less than five categories. In those studies that used only 5-point Likert scale, the choice of a single scale point served as a constant variable. Clearly the number of scale points was not a factor of interest. In the Muthén and Kaplan (1985) study, both equal and unequal thresholds were used for the 5-point Likert variables.

A few studies have manipulated the number of scale points and compared their effects on the CFA parameter estimates, standard errors of estimates, chi-square statistics, and practical fit indexes. Olsson (1979), Dolan (1994) and Green et al. (1997) compared the effects of 2-, 3-, 4-, 5-, 7-, and 9-scale points, 2-, 3-, 5-, and 7-scale points, and 2-, 4-, and 6-scale points, respectively. Green et al. used and transformed nonnormally distributed X scores to item data with 2, 4, and 6 categories. The same threshold values were applied to all nonnormally distributed data. For the 2-category, the threshold value was set at .5. A value of 1 was assigned if $X < .50$ and a value of 2 if $X \geq .50$. For a 4-category, a value of 1 was assigned if $X < .25$; 2 if $.25 \leq X < .50$; 3 if $.50 \leq X < .75$; and 4 if $X \geq .75$. For a 6-category, a value of 1 was assigned if $X < .17$; 2 if $.17 \leq X < .33$; 3 if $.33 \leq X < .50$; 4 if $.50 \leq X < .67$; 5 if $.67 \leq X < .83$; and 6 if $X \geq .83$. In Olsson's and Dolan's studies, probabilities under normality and asymmetry were used to create the threshold values.

Shape of Distribution

At this point, it is important to note that, although it is technically not accurate, I will follow the convention in the research literature of referring to symmetric ordinal (Likert) data as "normal" or "normally distributed" data.

In the Muthén and Kaplan (1985) study, five distributions ranging from normal to severely nonnormal were generated for the observed variables: (1) all observed variables were normal with zero skewness and kurtosis, (2) all observed variables with a mild negatively skewness (-0.742) and kurtosis (-0.334), (3) all observed variables with a moderate negatively skewness (-1.217) and kurtosis (0.846), (4) all observed variables were strongly censored (skewness = -2.028 and kurtosis = 2.898), i.e., a 'piling up' of observations at one of the extreme categories, and (5) all observed variables have zero skewness but high kurtosis (2.785). In their 1992 study, these same distributions were used in addition to a negative kurtosis (-1.30) distribution. The symmetric distribution corresponds to a desirable condition because it allows the study of the effects of categorization without the interference of skewness or kurtosis.

Three forms of distributions were varied in Curran et al. (1996): Distribution 1 was a multivariate normal with univariate skewness and kurtosis equal to zero, Distribution 2 was moderately nonnormal with univariate skewness of 2.0 and kurtosis of 7.0, and Distribution 3 was severely nonnormal with univariate skewness of 3.0 and kurtosis of 21.0. Five distributions in the Babakus et al. (1987) study were as follows: (1) all observed variables were normally distributed with zero skewness, (2) all observed variables follow a U-shaped distributions (0.273, 0.183, 0.090, 0.183, and 0.273), (3) all observed variables follow extremely skewed distributions with a skewness vector of

(1.50, 1.50, 1.50, 1.50), (4) two observed variables follow moderately and the other two extremely skewed distributions with a skewness vector of (0.50, 0.50, 1.50, 1.50), and (5) two observed variables follow a moderately skewed, one a normal distribution, and one an extremely skewed distribution with a skewness vector of (0.50, 0.50, 0, 1.50). These same distributions were used by Rigdon and Ferguson (1991). In Olsson's (1979) study, the degrees of skewness were varied as follows: (1) all skewness were the same (0, 0.50, 1, and 2), (2) half the variables have one value of skewness, the remaining variables a second value. The chosen combinations of skewness were 0.50 and -0.50; 0.50 and 0; 1 and -1; 1 and -0.50; 1 and 0; 1 and 0.50; 2 and -2; 2 and -1; 2 and -0.50; 2 and 0; 2 and 0.50; 2 and 1, and (3) the variables were divided into three equal groups, each with one value of skewness. The combinations were -0.50, 0 and 0.50; -1, 0 and 1; -2, 0 and 2.

Four different levels of nonnormality were included in Potthast's (1993) study. Distribution 1 was a normal distribution with zero skewness and kurtosis. Distribution 2 was a distribution with minimal skewness (0.19) but negative kurtosis (-1.12). Distribution 3 was a leptokurtic (very peaked) distribution with zero skewness but positive kurtosis (2.90). Finally, distribution 4 was a distribution with both high skewness (2.52) and kurtosis (5.80), representing a strongly censored distribution.

Only two forms of distributions were used by Dolan (1994): normal and mild asymmetry distributions. Green et al. (1997) have used a different methodology to yield four distributions: (1) uniform distribution, (2) unimodal, symmetric distribution, (3) negatively skewed distribution, and (4) half items negatively skewed, half items positively skewed (opposite skewed). They transformed normally distributed Z scores to yield four types of nonnormally distributed scores by varying the values of the

parameters (p and q) of the beta distribution. The probability density function for the beta distribution is

$$p(X) = \frac{(p+q-1)!}{(p-1)!(q-1)!} X^{(p-1)}(1-X)^{(q-1)},$$

where X is a random variable that can take on values between 0 and 1. The sets of Z scores were transformed to Xs so that the scores followed (1) uniform distribution (p = 1 and q = 1), (2) a symmetric, unimodal distribution (p = 4 and q = 4), (3) a negatively skewed distribution (p = 3 and q = 1.5), and (4) both negatively skewed (p = 3 and q = 1.5) and positively skewed (p = 1.5 and q = 3) distributions.

The distributions in Hutchinson and Olmos (1998) were varied to four forms: (1) a normal distribution with zero skewness and kurtosis; (2) a rectangular distribution with zero skewness and kurtosis of -1.326; (3) a symmetric and leptokurtic distribution with zero skewness and kurtosis of 2.668; and (4) a skewed and leptokurtic distribution with skewness of 2.558 and kurtosis of 5.919.

Finch et al. (1997) used three forms of distributions: (1) normal, (2) moderately skewed (skewness = 2 and kurtosis = 7), and (3) extremely skewed (skewness = 3 and kurtosis = 21). In their second simulation study, three mixed distributions were examined: (1) all variables were mildly and moderately nonnormal, (2) all variables were moderately to severely nonnormal, and (3) three variables were severely nonnormal and the other six were normally distributed.

In the DiStefano (2002) study, only two distributions were investigated. The distributions were approximately normally distributed and nonnormally distributed. To generate the approximately distributed data, for each of the five categories, the area under the curve was approximating 5%, 21%, 48%, 21%, and 5% (skewness = 0, kurtosis =

0.30). For nonnormal ordered categorical data, the percentage of responses in each category was approximately 75, 15, 5, 3, and 2, for categories 1 through 5 (skewness = 2.5, kurtosis = 6.0).

Model Size

Muthén and Kaplan (1985) used a one-factor model with four variables whereas Curran et al. (1996) used an oblique three-factor model with three variables per factor. Olsson (1979) has chosen the use of two one-factor models with 6 and 12 variables, respectively. In Babakus et al. (1987), a single-factor model with four variables was used. The model used in Rigdon and Ferguson's (1991) study consisted of two correlated factors with four variables loading exclusively on each factor. Potthast (1993) employed four increasingly complex oblique models: one-factor model with 4 variables, two-factor model with 8 variables, three-factor model with 12 variables, and four-factor model with 16 variables. Each model had four variables per factor. Dolan (1994) used a one-factor model with 8 variables whereas Green et al. (1997) employed a one-factor model with 20 variables.

Two model sizes were used in Hutchinson and Olmos's (1998) study. Model 1 was a two-factor oblique model with 8 variables and model 2 was a four-factor model with 16 variables. Each factor has four observed variables. Finch et al. (1997) used a three-factor model with 9 variables (3 variables per factor). DiStefano (2002) has examined two sizes of models: two-factor model with 12 variables (4 variables in the first factor and 8 variables in the second factor), and three-factor model with 16 variables (4 variables in each first and second factor and 8 variables in the third factor).

Size of Model Parameters

In Muthén and Kaplan's (1985) study, correlations of the underlying variables were all equal and medium-sized, that is, 0.49. All factor loadings (lambda) were 0.70, the factor variance was 1.0, and error/uniqueness variances were all 0.51. Olsson (1979) used two levels of factor loadings: all were equal (0.80) and all were varied (0.30-0.70). Factor loadings of 0.70, interfactor correlations (phi) of 0.30, factor variances of 1.0, and error variances of 0.51 were used in Curran et al.'s (1996) study. The factor loadings in Babakus et al's (1987) study were at two levels: all were equal to 0.80 and all were varied to 0.40, 0.60, 0.60, and 0.80. The factor loadings used in the Rigdon and Ferguson (1991) study were at two levels: high (0.90, 0.80, 0.80, 0.70) and low loadings (0.70, 0.60, 0.60, 0.50). Meanwhile, the factor correlations were also at two levels: high (0.70) and low (0.40). In the Potthast (1993) study, the factor loadings were all equal to 0.70, the factor correlations were 0.30, factor variances were fixed at 1.0, and error variances were constrained to zero. The factor loadings used in Dolan's (1994) study were 0.80, 0.90, 0.70, 0.80, 0.70, 0.90, 0.80, and 0.90 and the factor variances were fixed at 1.0 and the eight diagonally error variances were 0.36, 0.21, 0.49, 0.36, 0.49, 0.21, 0.36, and 0.21. In the Hutchinson and Olmos (1998) study, the values of the factor loadings were 0.60, 0.70, and 0.80, and the factor correlations were 0.50. The population values for the factor loadings in Finch et al. (1997) were uniform and moderately high, that is 0.70 and the error variances were 0.51. The population values for the structural coefficients were specified as such beta 1: 0.60, beta 2: 0.20, and beta 3: 0.12. The magnitude of the direct effect of latent exogenous variable on latent endogenous variable was specified to be equal to the magnitude of indirect effect of latent variable one on three via two.

Two sizes of factor loadings were used in DiStefano (2002): 0.30 and 0.70. The sizes of factor loadings were varied for four population models. In Population 1A, the loadings on the first four-variable factor (two variables with 0.70 and two with 0.60) were high; and the loadings on the second factor (two with 0.50, four with 0.40, and two with 0.30) were moderate. Population 1B has an opposite condition where the loadings on the first factor was moderate (two with 0.40 and two with 0.30) and the loadings on the second factor were high (two with 0.70 and six with 0.60). The sizes of the loadings in Populations 2A and 2B were manipulated as (1) high loadings for each of the four-variable factors (0.60-0.70) and moderate loadings for the eight-variable factor (0.30-0.50), and (2) moderate loadings for each of the four-variable factors (0.30-0.40) and high loadings for the eight-variable factor (0.60-0.70). The factor correlations were held constant at 0.30.

Model Specification

To date few CFA studies have focused on the model specification. Correct model specification is an important structural assumption because if the sample structure does not accurately reflect the population structure, severe distortions will be introduced into the final CFA solution. In view of this problem, Curran et al. (1996) had included four model specifications. Model 1 was accurately specified such that the sample model reflected the population model. Model 2 contained two factor loadings that were estimated in the sample but did not exist in the population. Although a misspecification of inclusion happened, model 2 was considered as an accurately specified model because the expected value of the parameter estimation was 0. Model 3 excluded two factor loadings from the sample that did exist in the population and was a misspecification of

exclusion. Model 4 was the combination of Models 2 and 3 and involved a misspecification of both inclusion and exclusion.

Sample Sizes

The sample sizes used in Muthén and Kaplan's 1985 and 1992 studies were 500 and 1000, respectively. Only sample size of 1000 was used in Olsson (1979). In the Curran et al. (1996) study, four sample sizes were used: 100, 200, 500, and 1000. Two sample sizes were included in Babakus et al.'s (1987) study: 100 and 500. Rigdon and Ferguson (1991) included three sample sizes in their study: 100, 300, 500. The sample sizes used by Potthast (1993) were 500 and 1000. Dolan (1994) used three sample sizes: 200, 300 and 400. In the Green et al. (1997) study, only one sample was used, that is, N = 1000. The sample sizes used in Hutchinson and Olmos (1998) were 500 and 1000. Sample sizes of 150, 250, 500 and 1000 were studied by Finch et al. (1997). Only two sample sizes were included in the DiStefano (2002) study: 350 and 700. In short, all the sample sizes used in the single-group CFA studies were small and moderate.

Combination of Estimation Method and Correlation Measure

Muthén and Kaplan (1985) examined the performance of the normal theory-based ML and GLS on the estimation of $\Sigma(y)$ parameters as well as on the chi-square and variability measures (true and estimated standard errors). The performance of these estimators in relation to the degree of skewness and kurtosis of the Likert observed variables was investigated in their study. The ADF estimator was examined for its performance on the estimation of $\Sigma(y)$ parameters. Muthén and Kaplan also compared the performance of ML, GLS and ADF on the estimation of $\Sigma(y^*)$ parameters. In addition, the performance of CVM on the estimation of $\Sigma(y^*)$ parameters and sampling variability was studied specifically for dichotomized observed variables. Sample covariance matrix

was analyzed for each of the estimators. For the analyses of nonnormal ordered categorical variables, the ADF (for x models) and the CVM (for x^* models) are two relatively new promising approaches. With strong skewness and/or kurtosis, these two estimators outperform the normal theory-based ML and GLS estimators. The ADF is not based on the assumption of multivariate normality so that nonnormal distributions do not cause any problems in the estimation. As a normal theory estimator, the ML assumes multivariate normality and the fourth-order moments are equal to zero and the multivariate kurtosis is ignored. Because the ADF does not assume multivariate normality, the fourth-order moments are equal to zero, and the multivariate kurtosis is incorporated into the computation of test statistics. The CVM does not require the use of Pearson correlations because it fits the model based on the estimated latent correlations of the underlying variables. Muthén and Kaplan (1985, 1992) found that the normal theory estimators perform quite well even with ordered categorical and moderately skewed/kurtotic variables when the sample size is not small, N=1000. The distortions of ML and GLS chi-squares and standard errors were negligible if most variables have univariate skewness and kurtoses in the range -1.0 to +1.0. They contended that when most skewnesses and/or kurtoses are larger in absolute value than 2.0, and correlations are 0.5 or higher, distortions of the ML and GLS chi-square values and standard errors are very likely to occur, although the parameter estimates are robust.

Olsson (1979) analyzed ordinal data with the ML estimator. He found that the skewness of the variables, rather than the number of scale points, was the major determinant of lack of fit of the factor model. He also noted that ordinal factor analysis of the Pearson correlations for the dichotomous variables (phi coefficients) could lead to

inconsistent and attenuated estimates in addition to incorrect standard errors of estimates and incorrect chi-square test of model fit.

Three estimators were included in Curran et al.'s (1996) study. They were ML, ADF, and SB. The SB provides corrected chi-squares and is an alternative to the normal theory ML estimator when the observed data do not meet the assumption of multivariate normality. The ML chi-square showed no bias across all sample sizes under multivariate normal distributions but was significantly inflated with increasing nonnormality. Under the condition of nonnormality, Type I error rate increased indicating that a correct model was significantly more likely to be falsely rejected based on the ML chi-square statistic. The ADF chi-square was substantially inflated at smaller sample sizes even under multivariate normal distributions. At sample size of 500 and above, the ADF was unbiased regardless of distribution. The SB scaled chi-square performed well across almost all sample sizes and all distributions even under severely nonnormal distribution at sample size of 200 and above.

Using only the ML estimator, Babakus et al.'s (1987) study has shown that the polychoric correlation performs better than Pearson product moment, Spearman's rho, and Kendall's tau-*b* correlations on the estimation of factor loadings and their respective standard errors. The Pearson correlations performed as the second best correlation measure. Although the polychoric correlation produced the best results in terms of the accuracy of parameter estimates and estimated standard errors, it produced the poorest fit statistics, leading to frequent rejections of a true or correctly specified model. Also there was a higher rate of nonconvergent and improper solutions when the polychoric correlations were used as input in LISREL. With respect to the practical fit indexes, the

polychoric correlation performed poorly on GFI, AGFI, and RMR, especially for small sample size. Under normal distribution, Babakus et al. noted that both GFI and AGFI were related directly to sample size but not to the magnitude of true factor loadings. The RMR was affected inversely by both sample size and true factor loadings. All correlation measures performed poorly under severe nonnormal distribution. When sample size was increased from 100 to 500, both categorical and continuous data analyzed with all four correlation measures resulted in much smaller standard errors of factor loadings.

Rigdon and Ferguson (1991) compared the performance of ML, GLS, WLS, DWLS, and ULS. They found that both GLS and ML were susceptible to nonconvergent and improper solutions. Compared to the ML, the GLS produced a higher rate of nonconvergent and improper solutions. The other estimators (WLS, DWLS, and ULS) combined produced only about 1.6% improper solutions. Both nonconvergent and improper solutions were most common under severe nonnormal distribution and when sample size was 100. In order, the WLS, DWLS, and ML estimators were found to have produced estimates with the smallest mean square errors over all experimental conditions. However, these relative positions were dependent on sample size. As sample size was increased from 100 to 300 and 300 to 500, the relative performance of the ML estimator improved to match that of the WLS estimator. No combination of the polychoric correlation coefficient with any of the five estimators produced estimated standard errors that were unbiased under all conditions.

Severe nonnormal distribution of the ordinal data induced large biases in the estimated standard errors. When the strength of the measurement relations was high, high skew level of the distribution produced the fewest large biases. The DWLS and WLS

estimators produced large biases across all the experimental conditions with the DWLS estimator producing the more often. And large sample size did not reduce the likelihood of large biases in the estimated standard errors. For instance, the ML estimator produced larger biases when sample size was increased. At sample size of 100, the ML performed relatively well whereas the GLS, WLS, DWLS, and ULS occasionally produced extremely biased estimated standard errors. When sample size increased to 300, the extreme bias values disappeared for all the estimators except DWLS. The ULS and WLS produced the smallest mean square error. In terms of fit statistics, when the polychoric correlation was used, the model-data fit as indicated by the GFI, AGFI, and RMR improved as sample size increased and as the true values of model parameters become larger. In contrast, when the ML or GLS estimator was used, the model-data fit as indicated by GFI and AGFI improved when measurement relations become weaker whereas the strength of relations between constructs would have no effect. The model-data fit was poorer for ordinal data with a skew distribution.

With respect to the chi-square statistics, no combination of the polychoric correlations with any of the five estimators produced chi-square values that followed the chi-square distribution. The ML, GLS, DWLS, and ULS estimators all produced greatly inflated chi-squares. On the contrary, the WLS produced chi-square values that were much too low. For the ML, GLS, DWLS, and ULS estimators, lower values for the chi-square statistic were also associated with (1) larger sample sizes, (2) weaker measurement relations (factor loadings), and (3) normal distributions. Meanwhile, for the WLS estimator, lower values of the chi-square statistic were associated with higher values of both factor loadings and correlations. The impact of sample size and

distribution was not evident. In short, the polychoric correlation coefficient is best to use with the WLS for ordinal data.

Using only CVM estimator, Potthast (1993) found that biases in the parameter estimates were minimal. However, the estimated factor correlations showed more bias than the factor loadings. In terms of the standard errors of indicator loadings, for small samples a negative bias in excess of 10 per cent was observed in the three largest models. For large samples the negative bias existed in the nine-parameter model, in this case a model with two factors, when positive or highly positive kurtosis was present. Under all conditions of nonnormality, the bias became more severe as the number of parameters increased from 15 to 22. Standard errors of indicator loadings were negatively biased relative to the standard deviation of the estimates in the models of 9 or more parameters, regardless of sample size. For a fixed level of nonnormality, this bias increased with increasing model size. A pattern of increasing negative bias was found as the level of nonnormality in the indicators or observed variables changed from negative to zero to positive and highly positive kurtosis. For the standard errors of factor correlations, increasing nonnormality caused increasing negative bias in the standard errors of the factor correlations. With a fixed sample size, an increase in model size produced larger negative bias in the standard errors of the average factor correlation. For smaller model sizes, less than 15 parameters, the chi-square values were not inflated. Contrariwise, for larger model sizes with 15 and 22 parameters, the chi-square values became excessively inflated. For sample size of 500, chi-square values were inflated. These problems worsened as positive kurtosis increased. At small sample size and when observed variables had strongest nonnormality, the true model with 22 parameters was rejected and

the chi-square value was inflated. In the case of zero kurtosis, large model chi-square values were extremely inflated. The weight matrix was clearly unstable for large models, especially when the observed variables have extreme skewness and kurtosis.

In short, the CVM estimated the parameters with slight positive bias that was not significant for any combination of model size, sample size, and nonnormality in the observed variables. In the model of more than four parameters, the bias worsened as model size and nonnormality of the observed variables increased. The negative bias diminished but did not disappear as sample size increased from 500 to 1000. The chi-square statistics were inflated and the number of rejections of the true model was higher except in the four-parameter model and cases of negative or zero kurtosis in the nine-parameter model. The Potthast findings indicate that the effects of using the CVM methodology with large models and small sample sizes for the CFA of ordered categorical variables are serious. Underestimated standard errors and/or highly overestimated chi-square statistics are inevitable under these conditions, especially when the observed variables are extremely nonnormally distributed.

Dolan (1994) found that the use of Pearson correlation with ML resulted in too large rejection rates of Anderson-Darling (AD) statistics when number of response categories was less than five. Unlike the χ^2, the AD statistic is sensitive to departure from the expected (null) distribution at the tails of the observed distribution. For the five-category, the AD statistics and rejection rates were acceptable when the response distribution was symmetric. Under nonnormality, both AD statistics and rejection rates were too high. However, the AGFI were all over 0.90. Negative biases were found in the mean and standard errors of factor loadings even when the response categories were

seven. The ML with polychoric correlation yielded the similar results except that the actual standard errors of the estimates were much smaller. The ML estimator yielded lower chi-square values when the response categories were up to four categories for all distributions except the opposite skewed distribution. Under the opposite skewed distribution, the chi-square values decreased up to 6-response category and model rejection also decreased. The values of the comparative fit index (CFI) were less than near perfect fit only if skewness varied across items.

The following eight measures of fit were examined for their performance in the Hutchinson and Olmos (1998) study: CFI, critical N, incremental fit index (IFI), measure of centrality (MOC), nonnormed fit index (NNFI), relative fit index (RFI), root mean square error of approximation (RMSEA), and chi-square statistic. Both the sample size and model size were found to have greatest effect on CN, RFI, and MOC. In addition, model size had an extremely large effect on chi-square but not on chi-square/df ratio. Better fit was obtained from a smaller model. Increasing nonnormality had an effect on the decrement of fit for the large model size (i.e., four-factor model). But when the chi-square/df was used, this interaction disappeared. The ML resulted in better fit for normal, rectangular, and symmetric-leptokurtic distributions whereas the WLS yielded better fit for extremely skewed and leptokurtic distribution. All fit indices except CN were adversely affected by increasing level of nonnormality. The most affected fit index was the RFI, especially at N = 500, data were skewed and leptokurtic, and the model was large. CFI, IFI, and NNFI were the least affected by the design variables except level of nonnormality. Only CN was less susceptible to nonnormality. MOC was most affected by the design variables. However, it was never the most affected by any single design

variable. The chi-square statistic was influenced by the same main and interaction effects as the MOC, with the exception of sample size. RFI was found to be sensitive to the main effects of sample size, model size, and level of nonnormality.

The findings reported in Hutchinson and Olmos (1998) were consistent with those of Muthén and Kaplan (1992) and Potthast (1993). That is, chi-square indicated poor fit for nonnormal data, especially when the model was large. Consistent with Babakus et al. (1987), the Hutchinson and Olmos findings showed that greater skewness led to lower values of GFI and AGFI. In addition, increasing nonnormality led to poorer fit for all of the fit indexes except CN. The WLS did not compensate for nonnormality except for chi-square, MOC, and RMSEA. Unlike the χ^2, RMSEA performed generally well in that its values were neither affected by sample size nor model size. While the RMSEA did reflect poorer fit when the data were skewed and leptokurtic, it was one of the few indexes to be amended by WLS. When data are symmetric and only moderately kurtotic, the use of WLS appears to afford little disadvantage over ML. It was also found that the NNFI was sensitive to model specification and independent of sample size.

The use of ADF and WLS is only applicable to large data sets. For applied researchers, analyzing ordered polytomous or Likert data without the benefit of large data sets and when data are nonnormal, the use of ADF and WLS is not feasible. They are forced to resort to the traditional ML estimator, which is the standard default method in almost all of the statistical software packages. Hutchinson and Olmos's (1998) findings suggested that ML could be used albeit minimally biased, provided that the data are not extremely nonnormal. The limitation of Hutchinson and Olmos's study is that all items are based on the same level of normality or nonnormality, which is unrealistic in practice

where distributions of items would vary within a given data set. It is useful to examine to what extent mixed levels of nonnormality can have effects on the various goodness-of-fit indices, especially the commonly reported chi-square test statistic in CFA.

Under multivariate normality, Finch et al. (1997) found that relative bias in the standard errors of indirect effect estimates produced by the ML was negligible. Furthermore, the relative bias in the standard errors of the direct effect structural coefficients was negligible. The standard errors of direct and indirect estimates became increasingly negatively biased as the observed variables became increasingly nonnormal, especially under moderate and severe nonnormal distributions. Under severe nonnormality, ML estimator underestimated the standard errors of the indirect effect by a moderate percentage of 23%. When sample size was large, the relative bias in the standard errors of the structural coefficients decreased.

Normal theory ML standard errors were too small or underestimated when the normality assumption was violated. In contrast, the ADF standard errors were unaffected by the distributional characteristics of the variables, but were substantially negatively biased in small sample size. The practical effect of negatively biased standard errors would be the rejection of the null hypothesis too frequently. Under the nonnormality, the ML robust standard errors performed much better at all sample sizes. The pattern of bias in the standard errors of direct and indirect effects was also not influenced by variation in the population values of the factor loadings.

Under the conditions of mixed nonnormality, no practical significant effects of either sample size or level of nonnormality were observed on the relative bias in the structural coefficients or the indirect effect estimates for either ML or ADF. When all

variables were mildly and moderately nonnormal, no appreciable effect of sample size was found on the estimated standard errors of the indirect effect using normal theory ML-robust. Bias in the standard errors was negligible. However, modest levels of relative bias in the standard errors of structural coefficients were observed for the ML and ML-robust.

The relative bias in the ADF standard errors decreased with increasing sample size. Under the moderately to severely nonnormal distribution, no consistent effects of sample size on ML but estimates of the standard errors of the structural coefficients were negatively biased. Under the mixed normally and severely nonnormal distribution, the ML and ML-robust standard errors of the structural coefficients were severely biased at sample size of 150. These nonnormality effects decreased when sample size became larger.

In the DiStefano (2002) study, under the normal distribution, increasing model size did not greatly affect the pattern of bias observed under both ML and WLS. Similarly, increasing sample size did not greatly reduce the level of bias in the observed parameter estimates. The use of the ML and Pearson correlation introduced little bias in factor correlations. In contrast, the WLS-polychoric correlation combination produced moderate levels of negative bias in factor correlations and the bias levels decreased when sample size was increased to 700. Similar to the bias levels of parameter estimates, the ML-Pearson correlation caused little bias in the standard errors of parameter estimates. The WLS-polychoric correlation introduced a large degree of bias, especially when the sample size was smaller. Both mixed positive and negative biases were produced by the WLS-polychoric correlation for the error variance parameters. Compared to WLS-polychoric correlation, the ML-Pearson correlation resulted in less inflation in chi-square

values at a small sample size and a large model. All fit indexes (χ^2, GFI, SRMR, NNFI, RMSEA) were within acceptable ranges except for WLS-polychoric correlation estimates of chi-square and standardized root mean square residual (SRMR) at smaller sample size.

Under the nonnormal distribution of ordinal data, ML-Pearson correlation introduced little bias in factor loadings and factor intercorrelations. On the contrary, the WLS-polychoric correlation was robust to nonnormality and increasing sample size did not greatly affect bias levels in parameter estimates. In terms of standard errors, the ML-Pearson correlation produced high level of negative bias whereas the WLS-polychoric correlation produced high level of bias only at smaller sample size, in this case $N = 350$. The WLS-polychoric correlation also produced higher chi-square and SRMR at smaller sample size. The SB scaled chi-square was found to be able to reduce the level of bias of standard errors of parameter estimates and the inflation of chi-square.

An important unpublished dissertation is Boomsma's (1983) study on the robustness of ML estimation against nonnormality. The Monte Carlo simulated data used in the study were generated according to four CFA models with number of variables ranging from 6 to 10 and the size of correlations varying across the full range. For each model he generated data according to various combinations of number of scale points (2-7) and skewness (a symmetric condition was also included). 300 replications were run for each combination. By holding the covariance structure true in the population, Boomsma was able to study the effects of skewness, without the confounding effect of categorization of the latent continuous variables into ordinal variables. Under the symmetric and skewed distributions, he found very little bias in parameter estimates. The effects of number of scale points and categorization with no skewness were found to be

very minor in terms of the true model rejection rates of the ML chi-square measure of fit. However, when the skewness value was larger than 1.0, the true model rejection rates were highly inflated. One caveat of the study is that only one sample size, that is, 400 was investigated.

Number of Replications

The number of replications in the computer simulation studies ranged from 100 to 400.

In summary, single-group CFA studies have shown that nonnormally distributed data could affect the performance of normal theory ML and GLS estimation methods in terms of the biases in standard errors of parameter estimates and fit indexes. When the observed variables have excessive skewness and/or kurtosis, the ML and GLS estimates of the standard errors and the associated chi-square statistic are incorrect. The number of scale points per se has relatively little impact on the chi-square goodness-of-fit test when the distribution of the categorized or ordinal variables is approximately normal. As the distributions of the categorized variables become increasingly and particularly differentially skewed, the chi-square values and the Type I error rates become inflated.

Ordinal Data with Measurement Invariance

Cheung and Rensvold (2002) have examined the impact of between-group constraints on the ΔGFI as indicators of measurement invariance based on the various invariance tests in the multi-group CFA. Two model sizes were used in their study: two-factor and three-factor models with 3, 4, and 5 items per factor. The factor loadings for the three items were 1.00, 1.25, and 1.50; for four items were 1.00, 1.25, 1.25, and 1.50; and for five items were 1.00, 1.00, 1.25, 1.50, and 1.50. The factor correlations were either 0.30 or 0.50 and the factor variances were set at either 0.36 or 0.81.

Cheung and Rensvold found that model complexity (i.e. number of items per factor and number of factors) could affect most of the goodness-of-fit indices (except for RMSEA). Only the standard error of the RMSEA was affected by model complexity. The Cheung and Rensvold study is limited to the use of ML as the sole estimation method and only two sample sizes (150 and 300) were used per group. They have reported that the data were generated to two multivariate-normal samples of size *N*. In short, no other details about the data characteristics (e.g., type of measurement scales) were given. However, this study has given a breakthrough in methodology by introducing the idea of looking at the changes in various practical GFIs used in MGCFA. According to Cheung and Rensvold, there is no standard against which a researcher can compare changes in practical GFIs in MGCFA for determining if the changes in the practical GFIs are meaningful when measurement invariance constraints are added to a model.

In their study, nearly all of the practical goodness-of-fit indices (noncentrality parameter, RMSEA, Akaike's information criterion, Browne and Cudeck criterion, Expected value of the cross-validation index, Normed fit index, Relative fit index, Incremental fit index, Tucker-Lewis index, Comparative fit index, Parsimonious normed fit index, Parsimonious comparative fit index, Gamma hat, Rescaled Akaike's information criterion, Cross-validation index, McDonald's noncentrality index, and Critical N) except the GFI were examined for their differences between two nested models. Although Cheung and Rensvold did include the chi-square difference test and the normed chi-square (χ^2/df) in the result section, these statistical fit indices were not used as criteria for evaluating the various hypotheses of invariance. Rather they were used to examine the quality of the simulation. Cheung and Rensvold suggested that ΔCFI,

ΔGamma hat, and ΔMcDonald's NCI were robust statistics for testing the between-group invariance of CFA models. However, "robust" to which type of violation is not mentioned in their study. A general criterion for not rejecting the null hypothesis of all types of invariance was proposed in the study. According to them, a value of ΔCFI smaller than or equal to -0.01 indicated that the null hypothesis of invariance should not be rejected. For ΔGamma hat and ΔMcDonald's NCI, the critical values are -.001 and -.02, respectively. Such criteria are not based on theoretical rationales.

Tests of Latent Mean Invariance

In the construct comparability and measurement invariance literature, fewer studies have included the test of latent mean differences, which are subsumed under the test of structural invariance. The logic is that a test of measurement invariance (associations of observed scores to the latent variable) should precede tests of structural invariance (association of latent variables with each other) and latent mean invariance. In other words, one needs to understand what is being measured before testing associations among what is measured (Anderson & Gerbing, 1988). If the associations between items and the latent variable differ across comparison groups, inferences about the latent variable or construct are not valid because the measures are calibrated to the latent variable differently. When measurement invariance does not hold, it is meaningless for researchers to proceed with the testing of latent mean invariance. Due to the observed variables' or items' psychometric properties that are not generalizable across subgroups of a population, group comparisons based on latent mean differences cannot be made. Bollen (1989), Horn and McArdle (1992), and Vandenberg and Lance (2000) advocate for a strict requirement of measurement invariance (i.e., factor loadings invariance)

evidence before testing restrictions on means and intercepts. Contrariwise, Byrne, Shavelson, and Muthén, (1989) adopt a more liberal perspective by suggesting that further testing for latent mean differences is warranted under the condition of partial metric invariance.

Hancock, Stapleton, and Berkovits (1999) have addressed the methodological issue of to what extent cross-group equality constraints must hold in order for the integrity of latent structural inference and of latent mean inference to be maintained. They presented an analytical treatment of loading invariance within multisample covariance structure models and of loading and intercept invariance within multisample latent mean structure models. Their study showed that conditions of partial measurement invariance and even configural measurement invariance need not preclude the belief of comparable constructs across population. The inference regarding construct comparability may be considered to rest in large part in the theoretical hands of the researcher and only to some extent by tests of measurement invariance. Improper cross-group measurement constraints may begin with those assigning scale to the factor(s) of interest. Fixing a factor's variance to 1 in both or more groups directly implies latent homogeneity of variance whereas fixing a loading path to 1 in both or more groups implies a one unit change in the factor yields the exact same amount of change in the associated indicator variable. If these implicit invariances do not hold in the population even prior to imposing any other explicit and even proper cross-group constraints, then inaccurate assessment of structural relations will likely result.

For covariance structure models, there are two options. First option, one should minimize loading constraints across groups. Rather, one would choose a loading for each

factor that can be argued on strong theoretical grounds to be the same value in the populations of interest, use this parameter to identify the factors' scale and then test and interpret the key structural parameters as they occur in the unstandardized solutions. Researchers should also conduct sensitivity analyses in order to determine the effects of their choice of scale indicator variables on the stability of the inferences regarding key unstandardized structural parameters. Second option, one may choose a loading for each factor somewhat arbitrarily, and then test and interpret the key structural parameters as they occur in the standardized solution.

For latent means models, the minimum requirement for valid latent mean inference is that the factor scale indicator properly chosen in both or more groups also has equivalent intercepts, thereby making the alternative of a standardized solution approach not feasible. Theoretically, one should identify a variable believed to have invariant relations with the factor across groups and no differential bias such that equal amounts of the factor would be expected to yield equal amounts of this particular variable. Again, sensitivity analyses should be conducted by varying the choice of variable with fixed unit loading and constrained intercept in order to examine the effect of such choice on the resulting latent mean difference.

Summary of Research Concerns

The purpose of this dissertation was to investigate the recommendation made by Byrne (1998), Jöreskog and Sörbom (1996) and others (Chou, Bentler, & Satorra, 1991; Hu, Bentler, & Kano, 1992) to use maximum likelihood estimation and the Pearson covariance matrix when one encounters ordinal data with large numbers of items and insufficient sample sizes in the context of MGCFA. In essence, this dissertation addressed the question that day-to-day researchers face: How does the formal test statistic

such as chi-square goodness-of-fit statistic (likelihood ratio test statistic) for various hypotheses testing of measurement invariance in MGCFA operate following Byrne, Jöreskog and Sörbom, and others' recommendation? This is an important question because this recommendation is widely followed in the research literature of education and psychology. According to Breckler (1990), a review of SEM applications in psychological research over the past 15 years reveals most to be based on Likert-type scaled data with the estimation of parameters using ML estimation method. However, it is important to note that I am not suggesting that the ML is the most appropriate method of analysis but rather that it is a common method and hence widely seen in the educational and psychological research literature.

It is evident from the literature review that treating ordinal data as if they were continuous for use with the normal theory statistical methods such as Pearson correlation, multiple regression, and single-group confirmatory maximum likelihood factor analysis is not a promising practice. Specifically, analyzing data obtained from Likert-type items with small numbers of response categories as well as with severe skewness and kurtosis can cause serious distortions in the Pearson correlations/covariances, ML chi-square statistics, and standard errors of parameter estimates for single-group CFA. The use of statistical methods designed for multivariate normal data for the analysis of ordinal data in the MGCFA may lead to more complex problems because of coarsely categorized scales, nonnormal item response distributions across groups, and unequal sample sizes across groups. As we know that the optimal number of response categories in single-group CFA is five, the optimal number of response categories in MGCFA is yet to be determined. There is no doubt that polychoric correlation and its corresponding

asymptotic covariance matrix and the weighted least square (WLS) estimator should be used for dealing with ordinal data. The polychoric correlation is an estimate of the correlation between two latent variables underlying their respective observed variables, where the latent variables are assumed to have a bivariate normal distribution (Jöreskog & Sörbom, 1996). One caveat is that the use of polychoric correlation and WLS is limited to less than 25 items. In addition, large sample sizes (at least 3,000 to 5,000 in each group) are required in order to obtain correct weight matrix for asymptotic covariance matrices and stable parameter estimates in CFA. Hence, it is important to examine the usefulness of Pearson covariance matrices and ML estimation method for analyzing ordinal data with large numbers of items and small sample sizes. Mixed item formats are widely used in the construction of measurement instruments. Yet little is known about the effects of using mixed item format in MGCFA. To my knowledge, these issues have not been addressed in MGCFA.

Given the above arguments, the research questions of this study are outlined as below:

Research Questions

1. What are the effects of ordinal-scaled data on the Type I error rates of the strong and full measurement invariance hypotheses across a number of scale points (ranging from 2 to 9 categories)?
2. What are the effects of ordinal-scaled data on the Type I error rates of the strong and full measurement invariance hypotheses across item formats (single- and mixed-item formats)?

3. What are the effects of ordinal-scaled data on the Type I error rates of the strong and full measurement invariance hypotheses across response distributions (normal and skewed)?
4. What are the effects of ordinal-scaled data on the Type I error rates of the strong and full measurement invariance hypotheses across sample size combinations?

All the questions are investigated by looking at the MGCFA results of the empirical rejection rates of the two hypotheses of measurement invariance.

CHAPTER III

METHODOLOGY

Two simulation studies were conducted to answer the research questions listed at the end of Chapter II. Study 1 focused on the context wherein one has a measure with all of the items having the same response format. Examples of single response formats abound in the psychological research literature, for example, a psychological measure of self-concept with 35 items to which the response options are on a four-point rating scale of agreement. Study 2 focused on the situation wherein one has a measure or test with a mixture of binary items and polytomously scored items. This mixed response format is most commonly seen in large-scale educational testing, such as the Third International Mathematics and Science Study (TIMSS) wherein an achievement test booklet may have 35 items of which 30 are scored correct/incorrect (i.e., binary) and the remaining 5 are scored on a three-point scale of incorrect (score of 0), partially correct (score of 1), and correct (score of 2). The total test scores for this mixed item format test range from zero to 40.

Study 1 represented the commonly found test and measure format in psychological measurement whereas Study 2 represented the test format found in some large-scale educational achievement tests. The simulation methodology used for these two studies reflected this disciplinary distinction. Of course, one should not interpret the above statements to imply that mixed item formats never occur in psychological measures nor that all educational achievement tests are of mixed item format.

Study 1: Ordinal Data with A Single Item Format

A Monte Carlo approach was used to investigate the research questions. In essence, the examinees' responses to the binary and Likert-type items were simulated to mimic processes in responding to the ordinal scales under controlled conditions. The use of real data would not be able to realize this goal.

To date, the effects of ordinal variables on model-data fits have been extensively examined with commonly used statistical methods such as SEM or single-group CFA. None has been done so far for ordinal variables with multi-group CFA. In view of this, a large population of ordinal responses was simulated in order to adequately examine the effects of the independent variables stated in the research questions and to assess the Type I error rate resulting from the use of ordinal data in the testing of the full and strong invariance hypotheses in the MGCFA framework.

The methodology was adapted from similar studies on ordinal data in correlation, multiple regression, and single-group CFA described in the literature review. A single-factor first-order measurement model with 30 indicators (items) was used for the MGCFA. The factor structure was assumed to be unidimensional both within and across groups. In this context, the model specifications were correct in both groups. Hence, the invariant single-factor CFA model across groups serves as the true model. There were two rationales for the use of a single-factor model. First, the majority of the measurement instruments in the educational and psychological research assess a unidimensional latent construct or unidimensional sub-scale scores. For example, the Coopersmith Self-Esteem Inventory for Children, Form B (Coopersmith, 1975) test items were designed to measure a unidimensional construct, that is, self-esteem. Second, large numbers of items (i.e.,

more than 25 items) and small sample sizes (i.e., less than 1000 observations per group) were purposely selected in this study to reflect many common testing situations. The reader should recall from the literature review that with large numbers of items and small sample sizes, the recommended polychoric or tetrachoric correlations and WLS/ADF, DWLS, and CVM estimation methods for dealing with ordinal data could not be used.

Study Design

The simulation study was set up as an 8 × 6 ×2 factorial design. The design variables were eight scale points (ranging from two to nine), two item distribution shapes (normal and positively skewed), and six sample size combinations ranging from 200 per group to 800 per group. This resulted in 96 cells in the simulation design with 100 replications per cell. MGCFA was conducted on each combination of the design variables. Each cell consisted of the dependent variable (i.e., Type I error rates) derived from the MGCFA results of testing for the two hypotheses of measurement invariance: Full and Strong Measurement Invariance.

At this point I will provide more details on each of the factors in the simulation experiment.

Number of Scale Points

For this study all of the items comprising a measurement instrument have the same number of scale points. For instance, all items have a 5-point scale. The number of scale points varies from 2 to 9, which reflect common practice of binary and Likert scales used in psychological research.

Distribution of the Item Responses

The shapes of the distributions for the observed ordinal variables were of two conditions: normal (symmetric) and nonnormal (positively skewed). For a normal distribution, all the observed ordinal variables are of equal intervals, resulting in symmetric responses in the middle of the scale range. There are two possible conditions of nonnormality in the observed ordinal variables, namely positively and negatively skewed. For a positively skewed distribution, the ordinal variables are of unequal intervals with responses bunching to the left whereas for the negatively skewed distribution, the ordinal variables are of unequal intervals with responses bunching to the right. In the current study, only positively skewed distribution was investigated because the majority of the ordinal data in the applied social and psychological research tended to be positively skewed.

Study 1A: Equal Latent Thresholds

The ordinal variables with equal interval scale points and normal (symmetric) distribution was similar to that used in many of the single-group CFA studies (e.g., DiStefano, 2002; Dolan, 1994; Finch, West, & MacKinnon, 1997). The responses for ordinal scales with equal interval were assumed to be normally distributed and the range of the scale points included a standardized scale of $z = -3$ to $z = +3$ (Bollen & Barb, 1981). The scale points were divided equally for each ordinal item response process in which the ordinal scale points were simulated. The cutting points are determined by considering the area under the normal curve between $\pm$ 3 standard deviations. On the average this range includes nearly all (i.e., over 99.7%) of the cases. Therefore, the generated symmetric ordinal data are, in essence, interval data. In order to determine the number of unit intervals per category, the six-unit interval between $\pm$ 3 is divided by the

number of categories. Let us consider a four-category case. The six intervals are divided by four, which results in 1.5 intervals for each category. If the value of the normally distributed variable is less than or equal to -1.5, the categorized or collapsed variable is coded as 1. If the value of the normal variable is greater than -1.5 and less than 0, the collapsed variable is coded as 2. If the value of the normal variable is greater than 0 and less than 1.5, the collapsed variable is coded as 3. Finally, if the value of the normal variable is greater than 1.5, the collapsed variable is coded as 4. The thresholds used for scale points ranging from 2 to 9 were appended in Appendix B.

Study 1B: Unequal Latent Thresholds

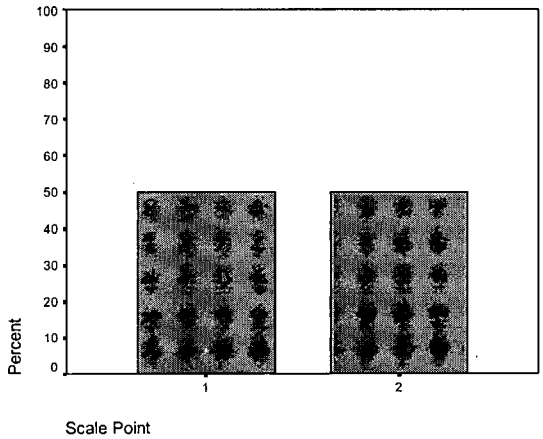
For ordinal variables with unequal intervals, the thresholds were set to generate item responses that were positively skewed. Consider a four-category case. Three thresholds were needed in order to collapse the latent continuous variable into a four-category observed variable starting at the latent z-score of zero. This leads to one interval for each category. If the value of the normal variable is less than or equal to 0, the collapsed variable is coded as 1. If the value of a normally distributed variable is greater than 0 and less than 1, the collapsed variable is coded as 2. If the value of the normal variable is greater than 1 and less than 2, the collapsed variable is coded as 3. For the value of the normal variable greater than 2, the collapsed variable is coded as 4. The thresholds used for collapsing the latent continuous variables into each number of scale points were attached in Appendix B.

In studies 1A and 1B, six combinations of equal and unequal sample sizes were considered for the two groups: 200 vs. 200; 500 vs. 500; 800 vs. 800; 200 vs. 500; 200 vs. 800; and 500 vs. 800. These were the typical sample sizes across two groups used with the ML estimation method and Pearson covariance matrix in MGCFA applied

research. It is expected that a large number of parameters are to be estimated but so far the minimum sample sizes needed for the two groups in multi-group confirmatory maximum likelihood factor analysis are still unknown. Hence, it is of this study's interest to examine the sample sizes required to maintain the Type I error rate for the ML estimation method in MGCFA.

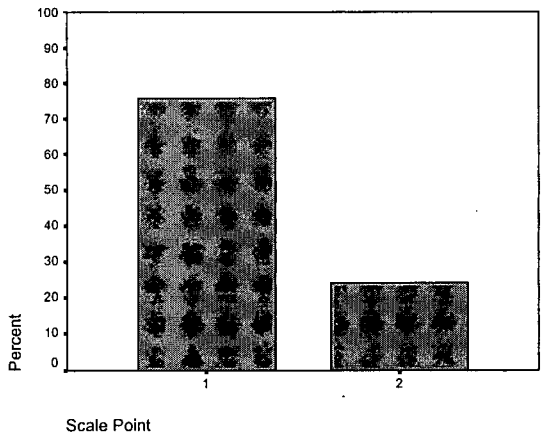
Study 1C: Controlling the Skewness of the Observed Variables

Study 1C was designed to control the skewness of the observed variables so that the effect of the number of scale points could be determined without the confounding variable of skewness – this confounding was present, by design, in Study 1B. Based on the results of Study 1B, the degree of skewness was varied into skewness values of 1.22, 1.34, and 2.03. A uniform or normal distribution (zero skewness) was included as a baseline condition. Because the confounding effect of skewness in Study 1B was more profound at 2, 3, and 5 scale points, only these scale points were examined for their separate effects on the empirical Type I error rates of the ML chi-square difference test based on the maximum likelihood estimation in the hypotheses testing of full and strong measurement invariance. The levels of skewness of the observed variables are presented graphically in Figures 4-15. Studies 1A and 1B varied sample size combinations but Study 1C only used 200 respondents per group (this was based on the findings in Studies 1A and 1B that sample size combination did not have an impact on the empirical Type I error rate).



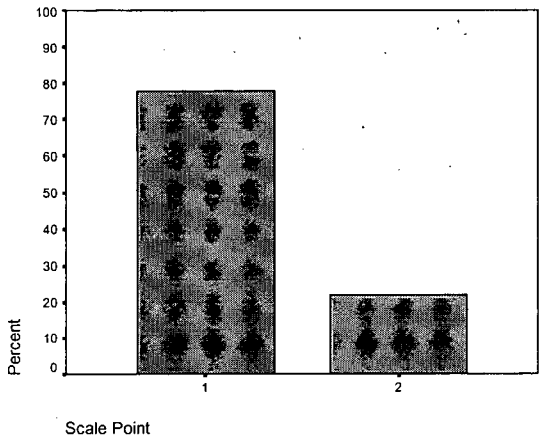
Percent in Categories: 50.0, 50.0

Figure 4. Histogram for two-point variable with a skewness value of 0.



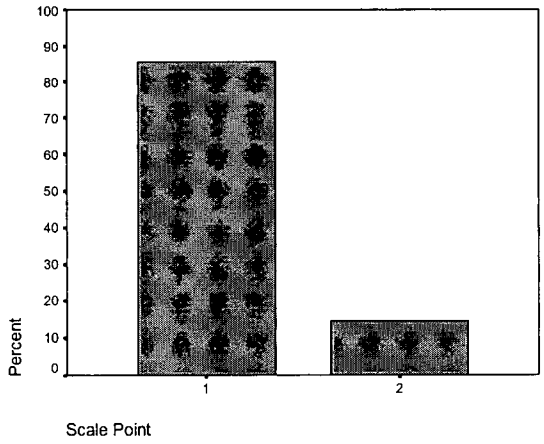
Percent in Categories: 75.9, 24.1

Figure 5. Histogram for two-point variable with a skewness value of 1.22.



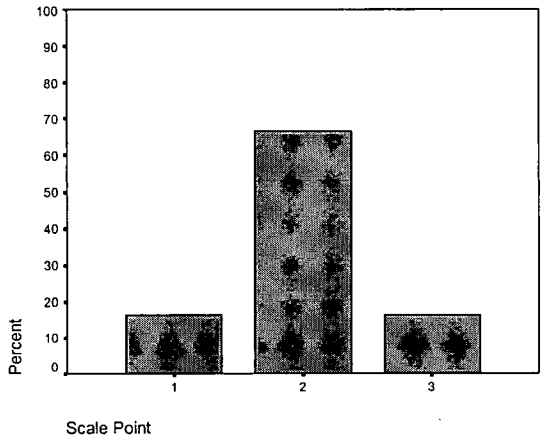
Percent in Categories: 77.9, 22.1

Figure 6. Histogram for two-point variable with a skewness value of 1.34.



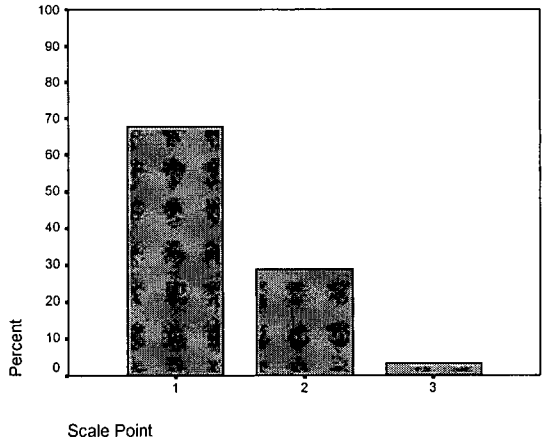
Percent in Categories: 85.6, 14.4

Figure 7. Histogram for two-point variable with a skewness value of 2.03.



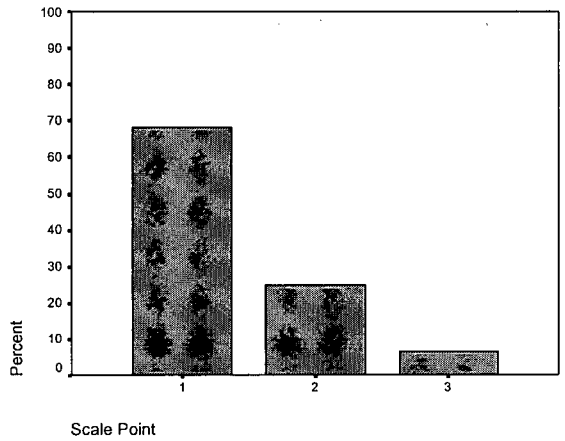
Percent in Categories: 16.7, 66.6, 16.7

Figure 8. Histogram for three-point variable with a skewness value of 0.



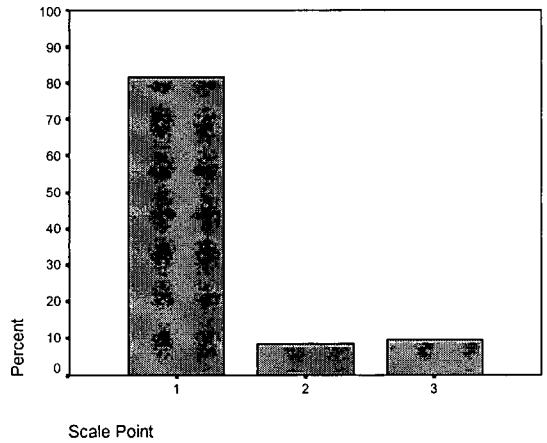
Percent in Categories: 67.7, 28.8, 3.5

Figure 9. Histogram for three-point variable with a skewness value of 1.22.



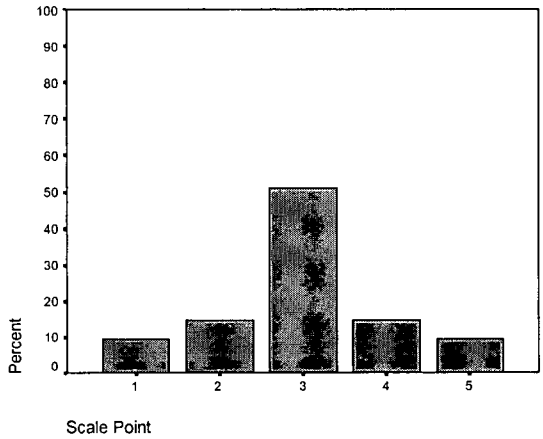
Percent in Categories: 68.3, 25.0, 6.6

Figure 10. Histogram for three-point variable with a skewness value of 1.34.



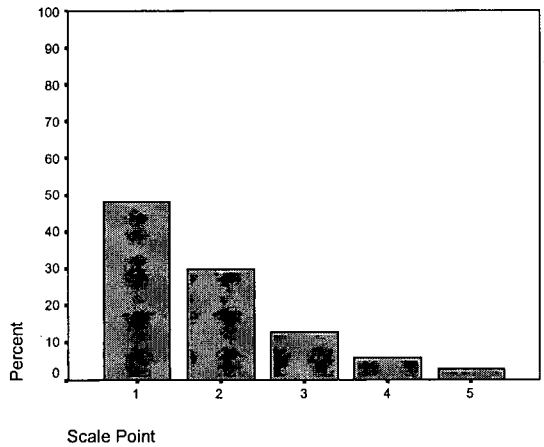
Percent in Categories: 81.6, 8.7, 9.7

Figure 11. Histogram for three-point variable with a skewness value of 2.03.



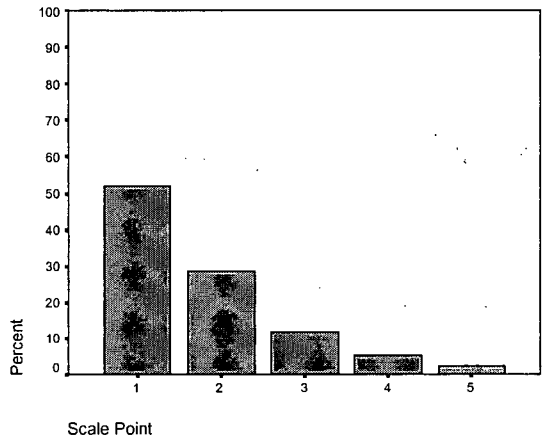
Percent in Categories: 9.7, 14.8, 50.9, 14.9, 9.7

Figure 12. Histogram for five-point variable with a skewness value of 0.



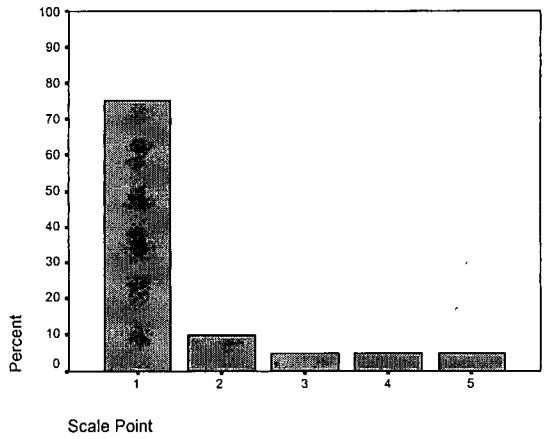
Percent in Categories: 48.1, 29.9, 13.1, 6.0, 2.9

Figure 13. Histogram for five-point variable with a skewness value of 1.22.



Percent in Categories: 51.9, 28.8, 11.8, 5.2, 2.3

Figure 14. Histogram for five-point variable with a skewness value of 1.34.



Percent in Categories: 75.2, 9.8, 5.0, 5.1, 4.9

Figure 15. Histogram for five-point variable with a skewness value of 2.03.

Simulation Procedure

Computer code was written to generate the latent responses set for 100,000 examinees. The resulting data set serves as a population from which observed response conditions were generated. A simulation iteration begins by specifying a common

population covariance matrix, Σ, as $\Sigma_g = \Lambda_g \Phi_g \Lambda_g' + \Theta_g$ for the two subgroups. The population model was a single-factor CFA model, with 30 items as indicators of the latent variable. The CFA model was invariant across the two groups. The population model parameters are summarized in Table 3.

Table 3

Model Parameters for Simulation

Parameter	*Value*
Factor variance	1.00
Item loadings	0.30-0.90
Error variances	0.19-0.91

The model parameters were composed of the following: (1) item/factor loadings with lambdas ranging from 0.30 to 0.90, (2) factor variance was set to 1.0 for identification and scaling purposes, and (3) the covariance matrix of the errors was defined to be diagonal with elements $\theta_i = 1 - \lambda_i^2$, assuming that the errors of measurement were uncorrelated. The values of lambda were selected to reflect the range of true item loadings commonly encountered in practice. To provide a realistic set of values, a set of educational research data (TIMSS data) was analyzed by using Principal Components Analysis. The resulting item loadings were used to specify the population covariance matrices. Other data could have been used but the TIMSS data are widely available hence enhancing the ease of replicability. In addition, values of the item

loadings in the ranges of 0.30 to 0.90 were also used in the majority of the reviewed single-group CFA simulation and applied psychological studies.

Taking one cell as an example, I outline the steps of my simulation, for Studies 1A, 1B, and 1C as follows:

1. A population covariance matrix Σ was created and normal continuous data was generated based on the item loadings typical of real data.
2. To approximate data from binary scales (2-scale point), the generated continuous data were cut into two ordered categories. An algorithm was written to classify thresholds in the generated continuous data to obtain approximately normally distributed (equal latent thresholds) and positively skewed (unequal latent thresholds), ordered categorical data with two categories. The one cut point or threshold used to categorize the continuous data into two ordered categories was chosen in accordance with area under the normal curve. For Likert-scale data, say a five-scale point, the generated continuous data were converted to five ordered categories using four cut points or thresholds. Again, an algorithm was used to classify item thresholds in the generated continuous data to obtain both normally and nonnormally distributed, ordinal data with five categories.
3. Both equal and unequal intervals in the Likert response categories or scale points were considered in this study. It has been common practice to design Likert scales with equal intervals in the applied research. In reality, not all the Likert scales have equal intervals. Therefore, it is of this study's interest to examine the effects of using Likert scales with unequal intervals in MGCFA. The shape of the

distributions of the Likert variables were created by adjusting item threshold values corresponding to both normal and positively skewed conditions.

4. The simulated data were imported to the PRELIS. Boostrapping was used to generate 100 Pearson covariance matrices for each group and LISREL 8.53 was used to run the MGCFA with Maximum Likelihood Estimation Method.

5. The LISREL output was transferred into text files. The files were exported to a statistical software package (SPSS) to include only the indexes of interest (i.e., 100 chi-square values and degrees of freedom for each hypothesis). The $\Delta\chi^2$, Δdf, p-values, and decision categories for full and strong invariance hypotheses were computed. These data were then read by a syntax file for the computation of mean rejection rates (empirical Type I error rates) for full and strong measurement invariance hypotheses. All the decision categories and the independent variables were entered into a single data file for later analysis: the Binary Logistic Regression analyses for examining the main effect of each independent variable and their interaction in predicting the decisions for full and strong invariance, respectively.

6. In short, the ordinal data used in this study represent characteristics encountered in the empirical research. The use of a single-factor CFA model across two groups approximate a realistic measurement model which is most frequently tested in the applied research for studying construct equivalence or factor structure comparability across two different groups. Number of scale points, sample size combinations, and shape of distributions were created after reviewing empirical

MGCFA studies in the applied literature between 1980 and 2002 in psychology and education.

Study 2: Mixed Item Format Data

Study 1 was based on data with a single item format in which the numbers of ordinal scale points were identical across all items whereas Study 2 focused on multiple item formats in which data with a mixture of binary and polytomous items were examined. Study 2 was a 6 × 3 factorial design, resulting in 18 cells. The design variables were mixed item formats and sample size combinations.

The mixed item formats were varied according to the proportion of ordered polytomous items as follows:

1. 67% (20) binary items and 33% (10) polytomous items (3 scale points),
2. 50% (15) binary items and 50% (15) polytomous items (3 scale points), and
3. 33% (10) binary items and 67% (20) polytomous items (3 scale points).

These item format proportions reflect the real achievement assessment data found in large-scale educational testing contexts such as TIMSS or the National Assessment of Educational Progress (NAEP). Given that most of the achievement data, when partial scores are allotted, use 3-category polytomous items, the polytomous items in the simulation were limited to item responses with 3 scale points. The sample size combinations were the same as those stated in Study 1.

Simulation Procedure

Given that the mixed response format is most commonly found in large-scale educational testing, the simulation model will reflect the item response models commonly

found in educational measurement – item response theory. Note that item response theory was only used as a model for generating item responses and not for test analysis.

For unidimensional binary items, the item responses were generated from the three-parameter logistic (3PL) item response theory model (Birnbaum, 1968),

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+\exp[-1.7a_i(\theta - b_i)]},$$

where a_i, b_i and c_i are the item i discrimination, difficulty, and guessing parameters, respectively. The $P_i(\theta)$ denotes the probability of answering correctly to item i by a randomly selected examinee with ability θ. The 3PL item parameters a, b, and c of each binary item were real item parameter estimates taken from the TIMSS Mathematics Achievement Test (1995, 1999). The sample distributions of these item parameter estimates are summarized in Table 4. The full information of all the 20 item parameter estimates is appended in Appendix C.

Table 4

Means for the Binary Item Parameters Used in This Study

$\bar{a}$	σ_a	$\bar{b}$	σ_b	$\bar{c}$	σ_c
0.984	0.073	0.078	0.078	0.233	0.025

Using a random number generator to produce numbers uniformly distributed on the interval [0,1], the probabilities were converted to either 0s or 1s to reflect examinee item scores. When the random number selected was less than or equal to $P_i(\theta)$, a "1" was assigned to an examinee for item i, and a "0" otherwise (Hambleton & Rovinelli, 1986).

It is useful to describe the Hambleton and Rovinelli methodology at this point as a series of steps.

Step #1: Examinees were simulated by randomly drawing an ability value from the normal distribution with a mean of zero and unit variance.

Step #2: For each examinee one now has the item parameter values (as described above) and the ability θ value. With the item and person statistics at hand one can compute each examinee's $P_i(\theta)$, i.e., their probability of a correct response.

Step #3: For each item and examinee, the examinee's probability of a correct response is compared to a uniformly distributed random number between 0 and 1, u. When the probability, $P_i(\theta)$, was greater than the random draw, u, the examinee's item response was coded as responding correctly to the item, 1. When the probability was less, the examinee was coded as responding incorrectly, 0.

Thus, as with real testing, individual simulated examinees sometimes respond incorrectly to items they should have been able to answer correctly. Step three, in essence, makes the item response generation a stochastic process, which is what is needed for simulation studies. That is, at the end of step two there is nothing stochastic, per se, in the simulation because at that point in the simulation $P_i(\theta)$ is fixed. Comparing $P_i(\theta)$ to a random number, between 0 and 1, makes the process stochastic and hence sometimes examinees respond incorrectly to items they should have been able to answer correctly -- hence making the simulated data like real data, if $P_i(\theta) > u$ then the item response is one.

For the polytomously scored items, the generalized partial credit model (GPCM)(Muraki, 1992) was used to generate unidimensional polytomous item responses, which were categorized into $r_i + 1$ ordered score categories (0, 1, …, r_i) for i-th item.

The model states that the probability of getting item score $U_j = q$ for a randomly sampled examinee with ability θ to the *i*-th item is given by

$$P_{i,q}(\theta) = \Pr ob(U_i = q|\theta) = \frac{\exp[\Sigma_{v=0}^{q} 1.7a_i(\theta - b_i + d_{iv})]}{\Sigma_{j=0}^{ri} \exp[\Sigma_{v=0}^{j} 1.7a_i(\theta - b_i + d_{iv})]}, \quad q = 0,1,\ldots, r_i,$$

where a_i is the slope parameter of item *i*; b_i is the location parameter of item *i*; and d_{iv} are a set of threshold parameters of item *i* with associated constrains $d_{i0} = 0$ and $\Sigma_{v=1}^{r_i} d_{iv} = 0$ (Muraki, 1992).

Because nearly all of the polytomous items in the TIMSS Mathematics Test consist of three-category items, polytomous items simulated in this study were all three-category polytomous items. A total of 20 polytomous item parameters (*a*s, *b*s, *d*s) were obtained from the TIMSS data. A summary of the sample distributions of the item parameter estimates is shown in Table 5. The full information of each item parameter estimates can be found in Appendix D.

Table 5

Means for the Polytomous Item Parameters Used in This Study

$\overline{a}$	σ_a	$\overline{b}$	σ_b	$\overline{d_1}$	σ_{d_1}	$\overline{d_2}$	σ_{d_2}
0.757	0.017	0.658	0.018	-1.046	0.048	1.046	0.051

The approach described by González-Romá, Hernández & Gómez-Benito (2002) was used to generate ordered polytomous items. For each examinee, a latent trait estimate θ was generated from a normal standard distribution, *N*(0,1). The GPCM probabilities were summed across categories to create a cumulative probability for each score level, and then the probability of responding above category k [$P_k^*(\theta)$] was computed. For each

simulated item and examinee a single random number (u) was randomly sampled from a uniform distribution over the interval [0,1], and the item scores were assigned as follows:

k = 3 if $P_2^*(\theta) \geq u$

k = 2 if $P_2^*(\theta) < u \leq P_1^*(\theta)$

k = 1 if $P_1^*(\theta) < u$.

It is important to note that the logic of the scoring rules for the ordered polytomous items are opposite of the binary items because in the ordered polytomous items $P_k^*(\theta)$ is the probability of responding *above* category k.

In total, two population data were simulated with equivalent parameters (i.e., measurement invariance across the two populations). The population data consist of 20 binary and 20 polytomous items. Three population data with different proportions of polytomous items were created by a random selection of the items. The skewness values of the population data according to the three conditions of mixed item formats are presented in Table 6. The response distributions for each of the mixed item format conditions were approximately normal.

Table 6

Mean Skewness of the Mixed Item Format Population Data

Mixtures of Item Formats	Mean Skewness
67% Binary and 33% Polytomous Items	-0.39
50% Binary and 50% Polytomous Items	-0.44
33% Binary and 67% Polytomous Items	-0.40

Testing for Measurement Invariance Hypotheses

Before the testing for the invariance of particular parameters across groups, a baseline model was first determined for each group. In this sense, all sets of parameters were estimated separately in the two groups (no between-group constraints). According to Byrne (1998), there are four specific hypotheses that need to be considered in the testing for the invariance of a measurement instrument: (1) the number of underlying factors is equivalent across groups, (2) the pattern of factor loadings is equivalent across groups, (3) structural relations among the factors are equivalent across groups, and (4) the reliabilities of item pairs from each subscale of the instrument are equivalent across groups. However, note that the testing for hypothesis three is not relevant to one-factor CFA models.

Figure 16 summarizes the two measurement invariance hypotheses that were tested in this study.

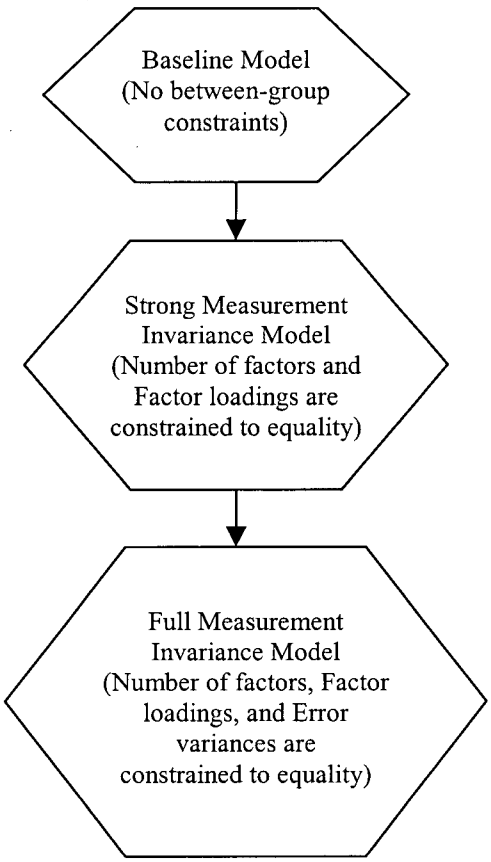


Figure 16. MGCFA nested models for the testing for two hypotheses of measurement invariance.

Estimation Method

The MGCFA was conducted by using the Pearson product moment covariance matrices along with the normal theory ML estimation method in the LISREL 8.53. There were 100 replications for each cell.

Dependent Variables

For each combination of the conditions, MGCFA was conducted for the tests of two hypotheses of measurement invariance. Effects of ordinal data and mixed item formats on the tests of hypotheses of measurement invariance were analyzed through the

mean rejection rates of the true models (Type I error rates). The continuous data with equal interval and normal distribution served as a baseline for comparisons and a quality check on the simulated data.

Analysis for the Simulation Results

Empirical Type I Error Rates

The empirical Type I error rates for each invariance hypothesis were computed as follows: *Empirical Type I error rates = number of rejections divided by 100 replications.*

In order to determine whether an empirical Type I error rate is inflated or not, the two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%) was computed using the normal approximation. Appendix E shows a table of the lower and upper confidence limits for each empirical alpha. The empirical alpha of .13 or below is within the two-tailed confidence interval. Hence, if an empirical alpha fell in the range of 0 and .13, the empirical Type I error rate is not inflated.

Binary Logistic Regression Analysis

Binary logistic regression was used to determine the main and interaction effects of the independent variables on the decisions of full and strong measurement invariance.

CHAPTER IV

RESULTS

In this chapter the simulation study outcomes will be presented and analyzed. Each study is presented in the order they are described in the previous chapter, the Methodology Chapter.

Study 1: Ordinal Data with A Single Item Format

It should be recalled that study one consists of three sub-studies: Studies 1A, 1B, and 1C. Together these sub-studies were designed to determine the effects of analyzing ordinal data with the normal theory Maximum Likelihood estimation method and Pearson covariance matrices in the MGCFA framework on the Type I error rates of the chi-square difference test (difference in chi-squares between two nested models) for testing full and strong measurement invariance hypotheses. The study was designed to have a large number of items (with all items loading on one factor) and small sample sizes to depict the research situation wherein the use of (a) polychoric correlation along with its corresponding asymptotic covariance matrix as well as LISREL's Weighted Least Squares method or Browne's (1984) Asymptotic Distribution Free method or (b) Muthen's (1984) Categorical Variable Modeling are not feasible. The ordinal data with a single item format were simulated to represent typical binary, Likert-type or rating scale data used in psychology. Many of the measurement instruments such as questionnaires, attitude and opinion surveys, rating scales, and, interest inventories consist of items with a single ordered categorical response format. One hundred replications were run in each of the conditions.

Check on the Simulation Methodology

As a first step, a quality check was made of the data generation process. MGCFA was run with 100 replications for the continuous data under varying sample size combinations before introducing any categorization. Table 7 shows that the mean rejection rates for the full and strong measurement invariance hypotheses are, as expected for multivariate normal item responses, within their two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%). Taking sampling variability into account, none of the chi-square test statistics are inflated. This indicates that the true models, namely full measurement invariance and strong measurement invariance models all hold for the simulated multivariate normal data. The empirical Type I error rates of the normally distributed continuous data also serve as a baseline for the comparison with the empirical rates of the symmetric and of the positively skewed ordinal data.

Table 7

Empirical Type I Error Rates of ML Chi-square Difference Test for the Full and Strong Measurement Invariance Hypotheses under Continuous Condition

Sample Sizes (*n1: n2*)	*Hypothesis*	
	FI	SI
200 : 200	.04	.05
500 : 500	.01	.02
800 : 800	.05	.06
200 : 500	.05	.08
200 : 800	.03	.04
500 : 800	.03	.05

Note. Those empirical Type I error rates that have the nominal alpha (.05) outside of their two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%) would be in bold font. FI and SI denote Full and Strong Invariance Hypotheses, respectively.

Study 1A: Equal Latent Thresholds

As discussed in Chapter Three, each of the latent continuous variables in the simulation were categorized into a number of scale points ranging from two to nine following the equal latent thresholds used by Bollen and Barb (1981). Using equal thresholds, the observed ordinal data were symmetric.

Symmetric Ordinal Variables.

The empirical Type I error rates of the ML chi-square difference test for the full and strong measurement invariance hypotheses across different number of scale points are presented in Table 8. The empirical rejection rates are zero for the two scale points.

At three and higher scale points, all of the Type I error rates had the nominal alpha (.05) within their two-tailed Bonferroni-corrected confidence interval. Likewise this protection of the nominal Type I error rate was upheld for equal and unequal sample sizes.

Table 8

Empirical Type I Error Rates of ML Chi-square Difference Test for the Full and Strong Measurement Invariance Hypotheses Across Number of Scale Points (Symmetric Distributional Condition) and Sample Size Combinations

Sample Sizes (n1: n2)	*Hypothesis*	*Number of Scale Points*							
		2	3	4	5	6	7	8	9
200 : 200	FI	.00	.09	.03	.04	.01	.03	.04	.00
	SI	.01	.08	.02	.05	.03	.03	.04	.00
500 : 500	FI	.00	.08	.04	.03	.04	.05	.03	.02
	SI	.00	.08	.04	.08	.05	.05	.03	.08
800 : 800	FI	.00	.05	.03	.02	.04	.03	.02	.05
	SI	.00	.07	.04	.04	.02	.02	.02	.04
200 : 500	FI	.00	.13	.02	.02	.04	.04	.07	.02
	SI	.00	.08	.04	.01	.02	.04	.09	.01
200 : 800	FI	.00	.08	.01	.00	.04	.07	.03	.03
	SI	.00	.06	.05	.02	.04	.04	.03	.02
500 : 800	FI	.00	.09	.01	.05	.04	.02	.01	.01
	SI	.00	.07	.03	.06	.05	.00	.00	.04

Note. Those empirical Type I error rates that have the nominal alpha (.05) outside of their two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%) would be in **bold font**. FI and SI denote Full and Strong Invariance Hypotheses, respectively.

Logistic Regression Analysis

Table 9 shows the results of logistic regression analysis for the simulation results in Table 8. The logistic regression was used to predict the two categories of decision for the full measurement invariance (1 - reject, and 0 - accept) from the number of scale points, sample size combination, and the interaction. The Wald Test values indicate that

neither the number of scale points nor the sample size combination main effects were statistically significant predictors for the decision for the full measurement invariance hypothesis. Likewise their interactions are not statistically significant at the alpha level of .05. The Hosmer and Lemeshow Test indicates a good model-data fit, $\chi^2 = 12.05$, $df = 8$, $p = .149$.

In Table 10, the logistic regression analysis shows that the number of scale points, sample size combination, and their interaction do not predict the decision for strong measurement invariance. The Wald Test values for each of the factors and their interaction are not statistically significant. The Hosmer and Lemeshow Test indicates a good fit of the LR model to the data fit, $\chi^2 = 11.16$, $df = 8$, $p = .193$.

Table 9

Logistic Regression Analysis of Decision for Full Measurement Invariance as A Function of Number of Scale Points and Sample Size Combination (Symmetric Distribution)

Variable	*Wald Test*	*df*	*Sig.*
Number of Scale Points	0.79	1	.372
Sample Size Combination	6.69	5	.245
Number of Scale Points* Sample Size Combination	6.06	5	.301

Table 10

Logistic Regression Analysis of Decision for Strong Measurement Invariance as A Function of Number of Scale Points and Sample Size Combination (Symmetric Distribution)

Variable	*Wald Test*	*df*	*Sig.*
Number of Scale Points	0.00	1	1.000
Sample Size Combination	0.94	5	.967
Number of Scale Points* Sample Size Combination	3.16	5	.676

Study 1B: Unequal Latent Thresholds

Given that the underlying thresholds are typically unknown and need not necessarily be equal, which is the case in most of the data collected from the social and behavioral sciences, unequal latent thresholds were thus used for the categorization of the latent continuous variables in the simulation into observed variables with a number of scale points ranging from two to nine.

Positively Skewed Ordinal Variables.

Table 11 shows the results of the empirical Type I error rates of the ML chi-square difference test for the full and strong measurement invariance hypotheses across different number of scale points under the condition of a positively skewed distribution. From Table 11, it seems evident that the empirical Type I error rates are inflated due to the increasing skewness rather than the increased number of scale points. The highest skewness is at 2 scale points. Then there is an increasing of skewness between 3 and 9 scale points. When the univariate skewness values are around 0.61, the empirical Type I

error rates for the two hypotheses of measurement invariance are all less than or closer to the nominal alpha level of .05. With a skewness value of 0.91, all the empirical Type I error rates for the full measurement invariance hypothesis are inflated whereas most of the empirical Type I error rates for the strong measurement invariance have the nominal alpha (.05) within their two-tailed Bonferroni-corrected confidence interval of 99%. The empirical rejection rates are increasingly inflated when the skewness values are increased from 1.07 to 3.48. The majority of the empirical Type I error rates have a nominal alpha (.05) outside of their two-tailed Bonferroni corrected confidence interval of 99%.

Table 11

Empirical Type I Error Rates of ML Chi-square Difference Test for the Full and Strong Measurement Invariance Hypotheses across Number of Scale Points (Positively Skewed Distributional Condition) and Sample Size Combinations

Sample Sizes (n1: n2)	*Hypothesis*	*Number of Scale Points*							
		2 s = 3.48 k =10.09	3 s = 0.61 k = -0.55	4 s = 0.91 k = 0.02	5 s = 1.07 k = 0.41	6 s = 1.17 k = 0.68	7 s = 1.24 k = 0.91	8 s = 1.29 k = 1.04	9 s = 1.63 k = 2.21
200 : 200	FI	**1.00**	.03	**.21**	**.32**	**.46**	**.64**	**.61**	**.99**
	SI	**.99**	.03	.12	**.23**	**.25**	**.29**	**.29**	**.34**
500 : 500	FI	**1.00**	.02	**.15**	**.43**	**.49**	**.63**	**.60**	**.94**
	SI	**.99**	.05	.07	**.21**	**.23**	**.33**	**.35**	**.26**
800 : 800	FI	**1.00**	.01	**.19**	**.36**	**.48**	**.55**	**.65**	**.97**
	SI	**1.00**	.03	.06	.12	**.25**	**.24**	**.31**	**.28**
200 : 500	FI	**1.00**	.02	**.20**	**.35**	**.44**	**.65**	**.65**	**.97**
	SI	**1.00**	.03	**.15**	**.19**	**.16**	**.22**	**.28**	**.31**
200 : 800	FI	**1.00**	.04	**.16**	**.33**	**.50**	**.60**	**.54**	**.96**
	SI	**1.00**	.02	**.14**	.13	**.22**	**.29**	**.33**	**.23**
500 : 800	FI	**1.00**	.02	**.14**	**.44**	**.37**	**.62**	**.68**	**.98**
	SI	**1.00**	.03	.11	**.20**	**.23**	**.28**	**.35**	**.28**

Note. Those empirical Type I error rates that have the nominal alpha (.05) outside of their two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%) would be in **bold font**.
FI and SI denote Full and Strong Invariance Hypotheses, respectively. s = Skewness; k = Kurtosis.

Logistic Regression Analysis

The Wald Test results from Tables 12 and 13 indicate that number of scale points is a statistically significant predictor for the decisions for full and strong measurement invariance. Sample size combination is not a statistically significant predictor. For each of the invariance decisions, the interaction between the number of scale points and sample size combination is not statistically significant. Although the main effects of the number of scale points were statistically significant for both hypotheses, the Hosmer and Lemeshow Test indicates a bad model-data fit for the Logistic Regression Analysis of the decision for full measurement invariance, $\chi^2 = 1513.11$, $df = 8$, $p = .000$. Likewise, the model-data fit does not hold for the logistic regression analysis of the decision for strong measurement invariance, $\chi^2 = 949.36$, $df = 8$, $p = .000$.

Table 12

Logistic Regression Analysis of Decision for Full Measurement Invariance as A Function of Number of Scale Points and Sample Size Combination (Asymmetric Distribution)

Variable	*Wald Test*	*df*	*Sig.*
Number of Scale Points	40.58	1	.000
Sample Size Combination	0.59	5	.988
Number of Scale Points* Sample Size Combination	1.16	5	.949

Table 13

Logistic Regression Analysis of Decision for Strong Measurement Invariance as A Function of Number of Scale Points and Sample Size Combination (Asymmetric Distribution)

Variable	*Wald Test*	*df*	*Sig.*
Number of Scale Points	25.05	1	.000
Sample Size Combination	0.53	5	.991
Number of Scale Points* Sample Size Combination	1.63	5	.897

Going back to Table 11, the inflated empirical Type I error rates do not really reflect the effects of the number of scale points because the Type I error rates at 2 scale points are the highest, then they level off at 3 scale points and increase throughout the remaining scale points. The Type I error rate inflation may reflect the increasing of skewness. As discussed in the previous paragraph, the highest Type I error rates are found at 2 scale points in which the skewness is also the highest. From 3 to 9 scale points, there is an increasing skewness. Hence, the logistic regression was conducted with skewness. Tables 14 and 15 indicate that skewness is the only variable that is significant. The Hosmer and Lemeshow Test statistics indicate a lack of fit for the logistic regression models for the decision for full measurement invariance ($\chi^2 = 19.24$, $df = 8$, $p = .014$) and for the strong measurement invariance ($\chi^2 = 77.85$, $df = 8$, $p = .000$), respectively.

Table 14

Logistic Regression Analysis of Decision for Full Measurement Invariance as A Function of Skewness and Sample Size Combination (Asymmetric Distribution)

Variable	*Wald Test*	*df*	*Sig.*
Skewness	137.92	1	.000
Sample Size Combination	1.82	5	.874
Skewness* Sample Size Combination	2.10	5	.835

Table 15

Logistic Regression Analysis of Decision for Strong Measurement Invariance as A Function of Skewness and Sample Size Combination (Asymmetric Distribution)

Variable	*Wald Test*	*df*	*Sig.*
Skewness	83.98	1	.000
Sample Size Combination	2.78	5	.734
Skewness* Sample Size Combination	1.46	5	.918

In order to have a better control of the skewness, the logistic regression analyses were re-conducted with the data that consist of 3 to 9 scale points. Only the number of scale points was examined here. The logistic regression results are presented in Tables 16 and 17.

Table 16

Logistic Regression Analysis of Decision for Full Measurement Invariance as A Function of Number of Scale Points (3-9) and Sample Size Combination (Asymmetric Distribution)

Variable	*Wald Test*	*df*	*Sig.*
Number of Scale Points	175.51	1	.000
Sample Size Combination	1.65	5	.895
Number of Scale Points* Sample Size Combination	2.21	5	.819

Table 17

Logistic Regression Analysis of Decision for Strong Measurement Invariance as A Function of Number of Scale Points (3-9) and Sample Size Combination (Asymmetric Distribution)

Variable	*Wald Test*	*df*	*Sig.*
Number of Scale Points	41.18	1	.000
Sample Size Combination	2.05	5	.843
Number of Scale Points* Sample Size Combination	1.33	5	.932

As seen in Tables 16 and 17, the main effects of the number of scale points remain statistically significant even though the skewness was controlled by taking out the data for 2 scale points. For both the decisions for full and strong measurement invariance, the Hosmer and Lemeshow Test statistics indicate poor model fits (full: $\chi^2 = 156.88$, $df = 8$, $p = .000$; strong: $\chi^2 = 49.81$, $df = 8$, $p = .000$).

Due to the nature of the simulated data, it is not clear whether the inflation of the Type I error rates is due to number of scale points or skewness.

Study 1C: Controlling the Skewness of the Observed Variables

Because the effects of the number of scale points were confounded by the skewness of the observed variables in Study 1B, Study 1C was designed to disentangle the confounding effects of skewness by controlling the skewness of the observed variables.

Disentangling the Effect of Skewness from the Number of Scale Points.

Table 18 shows the empirical Type I error rates of the ML chi-square difference test for the full and strong measurement invariance hypotheses when the confounding effects of skewness are disentangled from the number of scale points. Because the sample size combinations and their interaction with skewness were found to have no impacts on the empirical Type I error rates, only one condition of the sample size combination, that is, 200 : 200 was investigated in Study 1C.

Table 18

Empirical Type I Error Rates of ML Chi-square Difference Test for the Full and Strong Measurement Invariance Hypotheses When the Effects of Skewness and Number of Scale Points Are Disentangled (Sample Size Combination of 200 : 200)

Distribution	*Hypothesis*	*Number of Scale Points* 2	3	5
Not Skewed (s = 0)	FI	.00	.04	.05
	SI	.01	.04	.04
Positively Skewed (s = 1.22)	FI	**.21**	**.44**	**.62**
	SI	*.13*	**.29**	**.17**
Positively Skewed (s = 1.34)	FI	**.32**	**.63**	**.74**
	SI	**.19**	**.31**	**.34**
Positively Skewed (s = 2.03)	FI	**.94**	**.98**	**.98**
	SI	**.72**	**.78**	**.72**

Note. Those empirical Type I error rates that have the nominal alpha (.05) outside of their two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%) would be in **bold font**. Empirical Type I error rate that is marginally within the two-tailed confidence interval is *italized.* FI and SI denote Full and Strong Invariance Hypotheses, respectively. s = Skewness.

Using eyeballing, it appears that increasingly skewness and numbers of scale points have led to the inflation of empirical Type I error rates. When the skewness values are above 1.0, the empirical Type I error rates are increasingly higher and have the

nominal alpha outside of their two-tailed Bonferroni-corrected confidence interval of 99%.

Logistic Regression Analysis

The inferential statistics based on the logistic regression analyses show that skewness is indeed a statistically significant predictor for the decision for full measurement invariance (see Table 19). The interaction between number of scale points and skewness is not statistically significant at the .05 alpha level. The chi-square and *p*-value of the Hosmer and Lemeshow Test indicate a good model fit, $\chi^2 = 11.23$, $df = 7$, $p = .129$.

Table 19

Logistic Regression Analysis of Decision for Full Measurement Invariance as A Function of Number of Scale Points and Skewness

Variable	*Wald Test*	*df*	*Sig.*
Number of Scale Points	1.89	1	.170
Skewness	17.75	1	.000
Number of Scale Points* Skewness	0.13	1	.722

As indicated by the Wald Test statistic and its respective *p*-value in Table 20, skewness is the only statistically significant predictor for the decision for strong measurement invariance. Neither number of scale points nor its interaction with skewness is statistically significant at the .05 alpha level. The Hosmer and Lemeshow Test indicates a good model fit, $\chi^2 = 14.19$, $df = 7$, $p = .048$.

Table 20

Logistic Regression Analysis of Decision for Strong Measurement Invariance as A Function of Number of Scale Points and Skewness

Variable	*Wald Test*	*df*	*Sig.*
Number of Scale Points	1.55	1	.213
Skewness	32.29	1	.000
Number of Scale Points* Skewness	0.97	1	.325

Study 2: Mixed Item Format Data

In this study, the data were simulated to reflect achievement data that have a mixture of binary items and ordered polytomous items. A quality check on the simulated data was conducted by testing the full and strong measurement invariance hypotheses at the population level for each mixed item format combination. As can be seen in Table 21, the differences in chi-squares between models, that is, baseline vs. full invariance, and baseline vs. strong invariance are not statistically significant at the alpha level of .05. The results indicate that the factor structure of the artificial achievement test is invariant across groups. Thus, any sample data drawn from the population data are expected to yield equivalent factor structures for the two groups in the MGCFA framework.

Table 21

Maximum Likelihood Chi-square Goodness-of-Fit Statistics between Models

Mixed Item Format	*Model*	*Chi-square Difference Statistic*	*p*
67% Binary Items 33% Polytomous Items (20:10)	Baseline vs. Full Invariance	$\Delta\chi^2 = 32$, $\Delta df = 60$	1.00
	Baseline vs. Strong Invariance	$\Delta\chi^2 = 21$, $\Delta df = 30$	.89
50% Binary Items 50% Polytomous Items (15:15)	Baseline vs. Full Invariance	$\Delta\chi^2 = 38$, $\Delta df = 60$	.99
	Baseline vs. Strong Invariance	$\Delta\chi^2 = 23$, $\Delta df = 30$	.82
33% Binary Items 67% Polytomous Items (10:20)	Baseline vs. Full Invariance	$\Delta\chi^2 = 39$, $\Delta df = 60$	.98
	Baseline vs. Strong Invariance	$\Delta\chi^2 = 23$, $\Delta df = 30$	.82

Note. Numbers of binary and polytomous items are in parentheses.

Table 22 reports the results of empirical Type I error rates of the ML chi-square difference test for the full and strong measurement invariance hypotheses across mixed item formats and sample size combinations. For both hypotheses, the empirical rejection rates of the ML chi-square difference test have the nominal alpha (.05) that fall within their two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%).

Table 22

Empirical Type I Error Rates of ML Chi-square Difference Test for the Full and Strong Measurement Invariance Hypotheses Across Mixed Item Formats and Sample Size Combinations

Sample Sizes (n1: n2)	*Hypothesis*	*Mixed Item Formats*		
		67% Binary 33% Polytomous	*50% Binary 50% Polytomous*	*33% Binary 67% Polytomous*
200 : 200	FI	.01	.02	.01
	SI	.00	.00	.00
500 : 500	FI	.00	.01	.00
	SI	.02	.01	.02
800 : 800	FI	.00	.01	.00
	SI	.01	.01	.00
200 : 500	FI	.00	.03	.00
	SI	.02	.00	.01
200 : 800	FI	.00	.03	.00
	SI	.00	.02	.00
500 : 800	FI	.00	.02	.02
	SI	.01	.01	.01

Note. Those empirical Type I error rates that have the nominal alpha (.05) outside of their two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%) would be in **bold font**. FI and SI denote Full and Strong Invariance Hypotheses, respectively.

Logistic Regression Analysis

For both hypotheses, none of the predictors in the logistic regression analyses are statistically significant at the alpha level of .05. The inferential statistics in Tables 23 and 24 support the results of the empirical Type I error rates as shown in Table 22. It is evident that mixed item formats, sample size combination, and their interactions do not have an impact on the decisions for full and strong measurement invariance.

Table 23

Logistic Regression Analysis of Decision for Full Measurement Invariance as A Function of Mixed Item Formats and Sample Size Combination

Variable	*Wald Test*	*df*	*Sig.*
Mixed Item Formats	0.00	1	1.000
Sample Size Combination	2.07	5	.840
Mixed Item Formats* Sample Size Combination	1.10	5	.954

Table 24

Logistic Regression Analysis of Decision for Strong Measurement Invariance as A Function of Mixed Item Formats and Sample Size Combination

Variable	*Wald Test*	*df*	*Sig.*
Mixed Item Formats	0.68	1	.411
Sample Size Combination	1.31	5	.934
Mixed Item Format* Sample Size Combination	0.84	5	.974

CHAPTER V

DISCUSSION

In this chapter, I will first restate the purposes of the study and then discuss the Monte Carlo simulation study findings and their implications in the order of the previous chapters. The methodological and educational contributions of the study, the limitations of the study, and the suggestions of future studies will also be discussed.

This study examined the impact of analyzing ordinal data and mixed item format data with the normal-theory Maximum Likelihood estimation method and Pearson covariance matrix in the framework of multi-group confirmatory factor analysis. The empirical Type I error rate of the ML chi-square difference test for full and strong measurement invariance hypotheses was systematically examined under varying conditions of numbers of scale points, response distributions, mixed item formats, and sample size combinations. Given that much of the data in the social and behavioral science (psychology) research are of binary and Likert-type data, Study 1 focused on the effects of analyzing ordinal data with the normal theory Maximum Likelihood estimation and Pearson Covariance Matrix in the MGCFA framework. The increasing use of the mixed item formats in achievement tests and of the MGCFA for evaluating the comparability of the factor structures of such tests across different groups set the stage for Study 2, which was designed to examine the effects of analyzing mixed format data with the normal-theory Maximum Likelihood estimation method and Pearson covariance matrix in MGCFA.

The Weighted Least Squares (Jöreskog & Sörbom, 1996) or Asymptotic Distribution Free (Browne, 1984) estimation of the model parameters using polychoric or

tetrachoric correlation and asymptotic covariance matrix and Muthén's (1984) Categorical Variable Methodology are theoretically sound estimation methods for dealing with data derived from an ordinal scale in the multi-group confirmatory factor analysis methodology. Unfortunately, the positive effects of these methods are offset by the following practical limitations: (a) due to computer memory constraints they are limited to approximately 20-25 items for computing the asymptotic covariance matrices, and (b) they are not recommended for small sample sizes (less than 1,000 per group) because of the instability of the weight matrix. In addition, ADF does not work well as a means of compensating for the effects of nonnormality unless the model is small (i.e., 9 observed variables or 24 degrees of freedom). Muthén and Kaplan (1992) found that ADF chi-square was quite sensitive to model size even for the multivariate normal continuous variables. Likewise, the standard errors of parameter estimates produced by ADF estimator were seriously downward biased as the model size increased.

Given that data with mixed binary and ordered polytomous items are of ordinal nature, the above-mentioned problems related to ordinal variables also apply to such data. Hence, the performance of the ML estimation method and Pearson covariance matrix in analyzing ordinal data and mixed item format data in the multi-group confirmatory factor analysis framework is worth investigating.

Study 1: Ordinal Data with A Single Item Format

As a data simulation check, as expected, the use of the Maximum Likelihood estimation and Pearson covariance matrix with continuous and multivariate normal data did not result in the inflation of the empirical Type I error rates for the full and strong measurement invariance hypotheses. The results are not surprising given that the ML

estimation method and Pearson covariance matrix are built upon the multivariate normality and interval measurement assumptions of the observed variables.

Study 1A: Equal Latent Thresholds

Under the conditions of a symmetric distribution, the number of scale points has little or no effect on the empirical Type I error rates of the ML chi-square difference test. The use of a coarse measurement scale such as a 2- or 3-point scale under the symmetric distribution did not lead to an inflation of the empirical Type I error rate for the full and strong measurement invariance hypotheses. This indicates that when the assumption of multivariate normality of the ordinal variables is not violated the ordinal data, regardless of number of scale points, could be analyzed with the normal theory ML estimation method and Pearson covariance matrix in MGCFA. These findings are consistent with the single-group CFA (Babakus et al., 1987; Boomsma, 1983; Muthén & Kaplan, 1985, 1992; Olsson, 1979) that found that when categorical variables approximate a normal distribution, the number of categories has little effect on the chi-square likelihood ratio test of model fit.

Study 1B: Unequal Latent Thresholds

Surprisingly, the number of scale points has little or no effect on the empirical Type I error rates when the observed variables are skewed due to unequal thresholds. The MGCFA of the simulation data indicate that the ML chi-square difference test statistics are robust to a small degree of skewness (< 1.0). When the skewness value is larger than 1.0, the empirical Type I error rates of the ML chi-square difference tests become inflated. The inflation increases as the distributions become increasingly skewed. This is in line with Boomsma (1983) and Muthén and Kaplan's (1985, 1992) findings on normal theory estimators. Based on single-group CFA with Likert-type data, Muthén and

Kaplan's (1985, 1992) results showed that for univariate skewnesses and kurtoses in the range of -1.0 to 1.0, the chi-square goodness-of-fit tests obtained from the normal theory estimation methods such as ML and GLS were quite robust to a moderate degree of nonnormality. The ML and GLS chi-square tests were found to be highly sensitive to univariate skewness greater than two in absolute value, which reflects a severe degree of nonnormality. Boomsma's findings showed that when the mean absolute value of the skewnesses of the observed variables was larger than 1.0, the use of the ML estimation with ordinal variables would affect model fit in structural equation modeling.

Under the worse case scenario in the present study, where the number of scale points is two and the response distribution is severely skewed, the empirical Type I error rates are profoundly inflated. The full and strong measurement invariance hypotheses were rejected more often than would be expected (at .05 level) even with a two-tailed Bonferroni corrected confidence interval of 99% on the empirical Type I error rates. The findings are consistent with Olsson (1979), who noted that ordinal factor analysis of the Pearson correlations for the dichotomous variables (phi coefficients) resulted in inconsistent and attenuated estimates in addition to incorrect standard errors of estimates and incorrect chi-square test of model fit. On a related note, when dichotomizing continuous variables to produce binary variables, the choice of cutting points can affect the values of the expected phi coefficients. According to Mislevy (1986), factor analyses of phi coefficients (Pearson correlation coefficients) of binary variables produced by the same underlying correlational structure but dichotomized at different points can lead to factor models with different structures and possibly different numbers of factors. Hence,

linear MGCFA should not be conducted with binary item responses that are severely skewed.

The use of equal and unequal sample sizes between the two groups does not result in an inflated empirical Type I error rate of the ML chi-square difference test for the full and strong measurement invariance hypotheses. Sampling variability has no impact on the statistical test of measurement invariance across the two groups. Increasing sample sizes across groups does not reduce the empirical Type I error. This is parallel to the single-group CFA findings. Under multivariate normality, Curran, West, and Finch (1996) found that the ML chi-square test statistic rejected the expected number of models across all sample sizes.

In short, the findings of the present study suggest that the use of a response scale with a greater number of response categories, when the response distributions are severely skewed, does not attenuate the inflation of the empirical Type I error rates of the true measurement invariance models.

The effects of the number of scale points and categorization with equal thresholds (which results in a symmetric distribution) seem to be negligible on the false rejection rates of the invariance hypotheses when the ordinal data are used with the ML estimation method and Pearson covariance matrix in MGCFA. Unfortunately, normally distributed response distributions are rare in psychology and education (Micceri, 1989). The skewness of the variables, rather than the number of scale points, is the major determinant of lack of fit of the measurement invariance models. As in Bollen and Barb (1981)'s investigation of the impact of using coarsely categorized measures on Pearson's *r*, the difference between the correlation of the continuous measures and that of the

collapsed variables are relatively small when equal latent thresholds were used for collapsing the continuous variables. Collapsing in this manner leads to variables that are symmetric and, as the number of response categories increases, to variables that approximate a normal distribution. The use of equal latent thresholds did not attenuate the Pearson correlations/covariances.

One can use Bollen and Barb's findings, as well as the findings by Muthen and Kaplan to postulate an explanation for my findings. That is, severe skewness (taken together with small number of categories) will distort ordinary Pearson product moment correlations/covariances. When the distorted Pearson covariance matrix is used as the input to the LISREL multi-group confirmatory maximum likelihood factor analysis, the standard errors of parameter estimates are underestimated and the ML chi-square difference test statistics are declared to be statistically significant more often than expected. This leads to the inflation of the empirical Type I error rates. As a result, an applied researcher may conclude erroneously that the factor structure of a measurement instrument is not invariant across groups – i.e., the research would conclude that the measure functions differently across groups.

Study 1C: Controlling the Skewness of the Ordinal Variables

In study 1B the effect of number of scale points and skewness are naturally confounded. When the confounding effects of number of scale points and skewness were disentangled, the increasingly inflated empirical Type I error rates of the ML chi-square difference test were found to be due to increasing skewness. This confirms that the inflation of the empirical Type I error rates is mainly attributed to the multivariate nonnormality of the observed ordinal variables. When these variables approximate a normal distribution, the number of scale points does not have an impact on the empirical

Type I error rates. This is consistent with the earlier findings, discussed above, by Muthen and his colleagues.

Implications of Findings

The combination of the ML estimation method and Pearson covariance matrix in MGCFA of ordinal data was robust to small numbers of scale points when the observed ordinal variables were approximately normally distributed. In other words, the empirical Type I error rates for the ML estimation method tended to be conservative for the symmetric distribution. However, when the assumption of multivariate nonnormality was severely violated (i.e., skewness value is larger than 1.0), the normal theory ML chi-square difference test statistic as a test of measurement invariance could lead to an inflated Type I error rate for model rejection. The ML estimation method committed an inflated empirical Type I error rate for testing both full and strong measurement invariance hypotheses in the MGCFA framework.

Unlike other computer simulation studies on reliability, validity, multiple regression, and single-group CFA, the implications of the current study into the question of the optimal number of scale points or response categories in MGCFA when ML estimation and Pearson covariance matrix are employed with ordinal data are not straightforward. When the multivariate normality of the observed ordinal variables is not violated, it seems evident that there is no noticeable difference of the empirical Type I error rates of the ML chi-square difference test between a dichotomous or binary scale and a 3-, 4-, 5-, 6-, 7-, 8-, or 9-category ordinal scale. It is important to ensure that equal latent thresholds are applied to the categorization of a continuous scale. However, this is not the case in most of the applied research. Given that the number of rating scale categories, behavioral anchors, wording, and response category descriptions can affect

the thresholds, the use of unequal latent thresholds is more realistic. Furthermore, many observed variables in the social and behavioral sciences are positively skewed. For example, depressive symptoms, child abuse, and psychopathology in the general populations are positively skewed (Curran et al., 1996).

The findings of dichotomous-scored data are consistent with the caveats expressed by Cohen (1983), who quantified the substantial losses in information (accuracy) that can occur when a continuous scale of measurement is dichotomized. Apparently, a violation of the multivariate normality assumption can seriously invalidate statistical hypothesis testing of measurement invariance.

This study has several important implications for practice. Regarding the ML estimation method and Pearson covariance matrix, the findings provide support for and extend cautions raised, by the single-group CFA research. ML estimation method and Pearson covariance matrix are robust to the violation of the interval measurement scale under the condition of multivariate normality. When the response distributions of the observed ordinal variables are severely skewed (> 1.0) across the two groups, the chi-squared difference test between the nested models cannot be used for making statistical decisions about measurement invariance across groups because of the inflation of empirical Type I error. Apparently, number of scale points does not compensate for multivariate nonnormality. The distributional assumption of the observed variables is the most important assumptions in multi-group confirmatory maximum likelihood factor analysis. One can also conclude that the shape of the distribution is more important than the levels of measurement as the criterion for deciding whether to use normal theory estimation method in structural equation modeling. This is akin to Zumbo and

Zimmerman's (1993) findings on the two-sample Student *t*-tests in which the authors concluded that "when deciding whether to use parametric or nonparametric statistical methods for a two-sample location problem, the shape of the probability distribution is a better criterion than levels of measurement for making such a decision" (p. 398).

Study 2: Mixed Item Format Data

Mixed item formats do not affect the empirical Type I error rates of the ML chi-square difference tests in the hypotheses testing of full and strong measurement invariance. The proportion of the polytomous items in the mixed item format data has no impact on the empirical Type I error when such data are treated as continuous and analyzed with the ML estimation and Pearson covariance matrix. Keep in mind that the distributions of the mixed item format data were approximately symmetric across groups or subpopulations. This is the first study of this kind so there is no literature to compare it to.

Implications of Findings

The use of ML estimation and Pearson covariance matrix is appropriate for analyzing data with mixed item formats, especially for data with large numbers of items and small sample sizes. Practically, the findings indicate that multiple-choice items and constructed-response items are psychometrically equivalent. Both item formats measure the same unidimensional construct provided that the item distributions are normally distributed. This provides support for substantive research on using mixed format test and factor structure invariance. Studies conducted by Bennet, Rock, and Wang (1991); Bridgeman (1992); Lukhele, Thissen, and Wainer (1994); Perkhounkova, Hoover, and Ankemann (1997); and Thissen, Wainer, and Wang (1994) showed that multiple-choice

and constructed-response items measured the same basic trait or proficiency. The incorporation of mixed item formats in large-scale achievement assessments is a good example of the use of multiple measures for enhancing the validity of inferences made from the test scores.

Contribution of the Study to the Measurement Literature

Methodological Contribution

Although a variety of techniques have been used to assess measurement invariance, there is a general agreement that the multi-group confirmatory factor analysis model represents the most powerful and versatile approach to testing for cross-group measurement invariance. Construct comparability is typically assessed by MGCFA. The current study findings provide the applied researchers with some statistical properties of MGCFA.

One way to ameliorate the distorting effects of employing Likert-type measures of underlying continuous variables is to construct measures in a way that increases the number of response categories into which a respondent's answers can be placed. However, the current study findings postulate that even a fine grained of measure can result in responses that are not normally distributed due to the examinees' characteristics. The response distribution matters the most in multi-group confirmatory maximum likelihood factor analysis research.

Although there has been no earlier study of mixed response formats involving binary and ordered-polytomous variables, some related methodological research has examined the analysis of structural equation models with mixed type of ordered polytomous and continuous variables. Muthén's CVM and Lee, Poon, and Bentler's (1992) two-stage estimation procedures are devoted to analyze mixed polytomous and

continuous data. Such multistage procedures have at least three pitfalls. First, the complexity of the procedures makes it unattractive to the applied researchers. Second, the analyses are computationally unwieldy and this makes it impossible to work with large and moderately large (more than 25 items) sets of items. Finally, large sample sizes (at least 3,000 per group) are needed to obtain a stable weight matrix. This study provides some insight into the use of ML estimation method with mixed item format data (i.e., a mixture of binary and ordered polytomous items) in MGCFA when large numbers of items and small sample sizes are the issues.

The findings of the current study echo Hutchinson and Olmos's (1998) recommendations that applied researchers analyzing polytomous data without benefit of the large sample sizes required to estimate the weight matrix in WLS should obtain fairly accurate, albeit minimally biased, measures of fit with ML, provided that their data are not extremely nonnormal. The study findings also contribute to the SEM literature on multivariate nonnormality. Kaplan (SEMNET, 1996) contended that the single-group CFA findings on multivariate nonnormality might be generalizable to MGCFA. He called for research on the issue of multivariate nonnormality in MGCFA. To my knowledge, this is the first study responding to Kaplan's important call.

Educational Contribution

While applying MGCFA to the factor structure invariance or measurement invariance and construct comparability studies, researchers must take into account the nature of the measurement scale and the characteristics of the data. It is important to advocate a multi-method approach to investigating construct comparability – item-level IRT and scale-level MGCFA methods. Empirical evidence of scale-level measurement

invariance via MGCFA is essential because without measurement invariance, it is difficult to interpret observed mean score differences meaningfully.

Both the International Test Commission Guidelines and the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) have emphasized the importance of reporting statistical evidence for test comparability in dual- and multi-language tests. To ensure that a test or measurement instrument is comparable across groups, one needs to verify that the test or measurement instrument has the psychometric properties of measurement invariance. Construct comparability or factor structure invariance is one of the two requirements (the other requirement is lack of item bias or DIF items) for measurement invariance (Hambleton & van der Vijver, 1996). In test construction, the evaluation of the dimensionality of a test across groups provides another piece of evidence regarding test fairness. Construct comparability is also an important assumption in test equating and linking (Kolen & Brennan, 1995). Hence, this study gives some methodological insights into the utility of multi-group confirmatory maximum likelihood factor analysis in construct comparability research.

MGCFA allows the testing of specific invariance hypotheses about whether certain features of a common factor model can be taken as invariant across populations. Measurement precision of a measurement instrument, for example, a questionnaire can be determined by looking at whether the factor loadings of an item can be construed as invariant. These results can then be compared to the IRT DIF analyses, keeping in mind that the item loadings in the measurement models of MGCFA are akin to the item slopes in the IRT framework. When the factor distributions and unique variances differ from one

group to the next, one can suspect the existence of varying population distributions and measurement precision.

From a psychometric point of view, it is important to study the various measurement invariance models because they imply different measurement properties. For example, the baseline model in MGCFA is equivalent to the testing for configural invariance. Following Thurstone (1947), the most basic and fundamental conceptualization of a construct is the pattern of zero and nonzero loadings, not the particular magnitude of the nonzero loadings. Therefore, in order to conclude that a construct can be conceptualized in the same way across groups, a set of items has to be at least cross-group congeneric, not necessarily tau equivalent. However, the researcher should refrain from making quantitative comparisons until more stringent forms of invariance such as strong (tau-equivalence) and full invariance (parallel) have been established. The current study has provided an illustration on how to conduct these various invariance hypotheses testing.

Finally, an MGCFA of Likert-type item is essential to psychological research wherein these scale types are widely used but that is not the only domain using Likert scales. The Programme for International Student Assessment (PISA 2000), a large-scale achievement assessment, for example, has incorporated Likert-type questionnaires to measure students' attitudes toward learning along with the achievement tests. The findings of the current simulation study will remind educational researchers, especially those who are involved in high-stakes international assessments, of the importance of checking for multivariate normality before proceeding with their MGCFA.

Limitations of the Study and Future Research

The greatest limitation to this research is that only one model size was examined. Future research should systematically explore varying model sizes and model complexity (e.g., models with cross-loadings). In addition, in the present study all the items have the same response distributions. Although a good place to start in conducting simulations of this nature, empirical research characterizing the kinds of patterns of response distributions found in real data is needed and these findings can then inform simulation studies. For example, no research to date has investigated how having response distributions vary across the items of a test might affect the ML chi-square difference test in MGCFA. In addition, the item response distributions may vary across groups.

There are still many situations and issues that need to be investigated. It is not certain that a normal distribution accurately reflects the true distribution of many underlying variables. For example, some underlying variables are inherently not normal. Future simulation research should examine the operating characteristic of the ML chi-square difference test when the latent variable is not normally distributed. According to Coenders, Satorra, and Saris (1997), the WLS estimation method and polychoric correlation coefficient/asymptotic covariance matrix should not be used with ordinal data that have a nonnormally distributed latent variable.

Collapsing the categories may alleviate problems due to skewness and kurtosis of the ordinal variables. This method was used by Muthen (1984) in a simultaneous study of causal model characteristics across gender in research about attitudes toward blood donation. Certain variables were collapsed to three or four categories. Although some information is lost when doing this, the compensatory effects may outweigh the loss.

Future research should address the systematic effects on standard error bias and chi-square overestimation, when ordered categories are collapsed in MGCFA.

Finally, research on the power of the ML chi-square difference test and model misspecification in MGCFA deserves considerable attention as does the matter of statistical power. Like much of the previous published research, the present study focused on the Type I error rate because that needs to be maintained before one can even discuss statistical power.

Recommendations

The objective of the current study was to investigate how the ML estimation method functions when the observed ordinal variables are treated and analyzed as if they were continuous, as per Byrne's (1998) recommendation. The ML chi-square difference test becomes liberal (i.e., inflated) when the response distributions significantly depart from normality, resulting in rejection of the true models more often than expected.

The outcomes of the MGCFA of ordinal and mixed item format data echo Boomsma 's findings on single-group CFA. Thus, I would like to make the same conclusion and recommendation that we shall not dissuade researchers to apply maximum likelihood estimation and Pearson covariance matrix in MGCFA, when the observed variables are discrete but symmetric. However, I do not recommend the use of such method when the mean absolute value of the skewness of the observed variables is larger than 1.0 because it would artificially inflate the chi-square values and empirical Type I error rates. As a consequence, crucial decisions of measurement invariance of an instrument are not valid.

It is fair to remind readers that the sample sizes used in the simulation studies are small to moderate. The use of equal and unequal sample sizes across groups, typical of the cross-group sample sizes in real psychological and educational data, does not have any impact on the statistical analysis of construct comparability. For large sample sizes, the chi-square difference test statistics should not be used because any hypotheses would be rejected by large sample sizes.

Is there a minimum number of scale points that a researcher should employ in the construction of items to be used in correlation/SEM analysis? Based on the current study findings, the answer is 'yes' or 'no', depending on the response distributions. If the response distributions approximate a normal distribution, then the number of scale points is not an important factor. In contrast, if the response distributions are skewed, then researchers should avoid the use of 2 scale points. The use of dichotomous variables in correlation or covariance analysis can be problematical in cases where the underlying continuous distributions for such variables are even moderately skewed. Correlations computed for such dichotomous variables provide very poor estimates of the true correlations between the continuous forms of these variables. A logical recommendation to researchers engaged in MGCFA using ML estimation method and Pearson covariance matrix might be to avoid entirely the use of dichotomous variables; however, such a recommendation is not realistic.

It should be noted, however, that deciding on the number of scale points one should use is not just a matter of statistical significance testing. An important consideration in deciding on the number of scale points to use is whether the population for whom the scale was developed can comprehend the task and respond to the

statements using the response categories. For example, if the target population for the test (or measure) is primary school children (typically ages 5 – 7 years old) then a two- or three-point response format may be more appropriate than five-point response format because children at that developmental stage may not be able to handle the subtlety of the discrimination needed to use a five-point response format. Likewise, the response points may be best described with a picture (e.g., a cat with various degrees of a smile indicating greater agreement) rather than the words "none", "a little", and "a lot". Of course, pilot testing is needed to determine if a response scale is appropriate.

As my results show, when the response distribution is severely skewed, the ML chi-square difference test as a formal statistical test of measurement invariance will lead to an inflated Type I error rate for hypothesis rejection. Consequently, in practice a researcher may mistakenly reject or modify a model by releasing the constraint of certain parameters because the distribution of the observed variables is not multivariate normal rather than because the model itself is not invariant across groups. According to Zumbo, Sireci, and Hambleton (2003), a viable alternative would be to avoid the distribution theory and statistical hypothesis testing of confirmatory factor analysis for evaluating construct comparability. In short, what these authors are suggesting is to abandon the reliance on statistical distribution theory when the assumptions are clearly violated. Instead, they suggest the use of (multi-group) exploratory factor analysis, MGEFA, which is more descriptive in nature. In addition, because the sampling distribution of the factor model statistics is not of concern, one can use robust and scale-appropriate correlation matrices involving full-information estimation, polychoric or outlier-resistant estimators with MGEFA. This recommendation reflects the inherent trade-offs one needs

to make in day-to-day psychometric and data analysis – sometimes one needs to trade-off the strength of the sampling theory and its focus on a population model, for the deleterious effect of violating those same assumptions.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anderson, J.C., & Gerbing, D.W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*, 411-423.

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

Bandalos, D. (1999, April). *The effects of item parceling in structural equation modeling: A Monte Carlo study*. Paper presented at the Structural Equation Modeling Special Interest Group of the American Educational Research Association Annual Meeting, Montréal, Quebec.

Bandalos, D., & Benson, J. (1990). Testing the factor structure invariance of a computer attitude scale over two grouping conditions. *Educational and Psychological Measurement, 50*, 49-60.

Bendig, A.W. (1954a). Reliability and the number of rating scale categories. *Journal of Applied Psychology, 38*, 38-40.

Bendig, A.W. (1954b). Reliability of short rating scales and the heterogeneity of the rated stimuli. *Journal of Applied Psychology, 38*, 167-170.

Bennet, R.E., Rock, D.A., & Wang, M. (1991). Equivalence of free response and multiple-choice items. *Journal of Educational Measurement, 28*, 77-92.

Bentler, P.M., & Wu, E.J.C. (1995). *EQS for Windows: User's guide*. Encino, CA: Multivariate Software Inc.

Bernstein, I.H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*, 467-477.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Bollen, K.A., & Barb, K.H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review, 46*, 232-239.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality*. Unpublished doctoral dissertation, University of Groningen, Groningen, The Netherlands.

Breckler, S.J. (1990). Applications of covariance structural modeling in psychology: Cause for concern? *Psychological Bulletin, 107*, 260-271.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice format. *Journal of Educational Measurement, 29*, 253-271.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*, 62-83.

Byrne, B.M. (1988). The Self Description Questionnaire III: Testing for equivalent factorial validity across ability. *Educational and Psychological Measurement, 48*, 397-406.

Byrne, B.M. (1994). Testing for the factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research, 29*, 289-311.

Byrne, B.M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Byrne, B.M., Shavelson, R.J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Chang, L. (1994). A psychometric evaluation on 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18,* 205-215.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Chou, C.P., Bentler, P.M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*, 347-357.

Cicchetti, D.V., Showalter, D., & Tyrer. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement, 9*, 31-36.

Coenders, G., Satorra, A., & Saris, W.E. (1997). Alternative approaches to structural equation modeling of ordinal data: A Monte Carlo study. *Structural Equation Modeling, 4*, 261-282.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249-253.

Coopersmith, S. (1975). *Coopersmith Self-Esteem Inventory*, Technical Manual. Palo Alto, CA: Consulting Psychologist Press, Inc.

Cronbach, L.J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*, 3-31.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105*, 317-327.

Curran, P.J., West, S.G., & Finch, J.F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16-29.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9*, 327-346.

Dolan, C.V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47*, 309-326.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*, 662-680.

Finch, J.F., West, S.G., & MacKinnon, D.P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling, 4*, 87-107.

González-Romá, V., Hernández, A., & Gómez-Benito, J. (2002). *An evaluation of the multiple-group mean and covariance structure analysis model for detecting differential item functioning in graded response items*. Paper presented at the International Test Commission (ITC) Conference on Computer-Based Testing and the Internet. Winchester, UK.

Green, S.B., Akey, T.M., Fleming, K.K., Hershberger, S.L., & Marquis, J.G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling, 4,* 108-120.

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

Hambleton, R.K., & Van der Vijver. (1996). Translating tests: Some practical guidelines. *European Psychologists, 1*, 89-99.

Hancock, G.R., Stapleton, L.M., & Berkovits, I. (1999, April). *Loading and intercept invariance within multisample covariance and mean structure models*. Paper presented at the Annual Meeting of the American Educational Research Association, Montréal, Quebec.

Horn, J.L., & McArdle, J.J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

Hu, L.T., Bentler, P.M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*, 351-362.

Hutchinson, S.R., & Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analyses using ordered categorical data. *Structural Equation Modeling, 5,* 344-364.

Jenkins, G.D., Jr., & Taber, T.D. (1977). A monte carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62*, 392-398.

Jöreskog, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Jöreskog, K. & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago, IL: Scientific Software International.

Kaplan, D. (1996). Multivariate Normality. *SEMNET.*

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

Kolen, M.J., & Brennan, R.L. (1995). *Test equating methods and practices*. New York: Springer.

Komorita, S.S. (1963). Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology, 61*, 327-334.

Komorita, S.S., & Graham, W.K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement, 4*, 987-995.

Labovitz, S. (1970). The assignment of numbers of rank order categories. *American Sociological Review, 36*, 515-524.

Lee, S-Y., Poon, W-Y., & Bentler, P.M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika, 57*, 89-105.

Long, J.S. (1983). *Confirmatory factor analysis: A preface to LISREL*. Beverly Hills, CA: SAGE Publications, Inc.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response and examinee selected item on two achievement tests. *Journal*

of Educational Measurement, 31, 234-250.

Matell, M.S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: reliability and validity. *Educational and Psychological Measurement, 31*, 657-674.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156-166.

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

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

Muthén, B.O. (1988). *LISCOMP: Analysis of linear structural equations using a comprehensive measurement model: User's guide*. Chicago, IL: Scientific Software International.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of nonnormal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38*, 171-189.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis for non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19-30.

Nunnally, J.C. (1967). *Psychometric Theory*. New York: McGraw-Hill.

O'Brien, R.M. (1979). The use of Pearson's r with ordinal data. *American Sociological*

Review, 44, 851-857.

O'Chieng, O.O. (2001). *Implications of using Likert data in multiple regression analysis.* Unpublished doctoral dissertation, University of British Columbia, Vancouver.

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

Olsson, U.H., Foss, T., Troye, S.V., & Howell, R.D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7,* 557-595.

Perkhounkova, Y., Hoover, H.D., & Ankemann, R.D. (1997, March). *An examination of construct validity of multiple-choice versus constructed response tests.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Potthast, M.J. (1993). Confirmatory factor analysis of ordered categorical variables with large models. *British Journal of Mathematical and Statistical Psychology, 46*, 273-286.

Radloff, L.S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401.

Ramsay, J.O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika, 38*, 513-533.

Reise, S.P., Widaman, K.F., & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Rigdon, E.E., & Ferguson, Jr. (1991). The performance of the polychoric correlation

coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research, Vol. XXVIII*, 491-497.

Rosenberg MI. (1965, 1979). Rosenberg self esteem scale [RSE, RSES]. In K. Corcoran & J. Fischer (Eds.), *Measures for clinical practice: A sourcebook.* New York: Free Press.

Satorra, A., & Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *American Statistical Association 1988 Proceedings of the Business and Economic Section* (pp. 308-313). Alexandria, VA: American Statistical Association.

Sireci, S.G., Bastari, B., & Allalouf, A. (1998, April). *Evaluating construct equivalence across adapted tests*. Paper presented at the Annual Meeting of the American Psychological Association, San Francisco, CA.

Sireci, S.G., Xing, D., & Fitzgerald, C. (1999, April). *Evaluating adapted tests across multiple groups: Lessons learned from the IT industry.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montréal, Quebec.

Steenkamp, J-B E.M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *The Journal of Consumer Research, 25*, 78-90.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.

Symonds, P.M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology, 7*, 456-461.

Thissen, D., Wainer, H., & Wang, X.B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice

tests: An analysis of two tests. *Journal of Educational Measurement, 31*, 113-123.

Thurstone, L.L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Tippets, E., & Michaels, H. (1997, March). *Factor structure invariance of accommodated and non-accommodated performance assessments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance hypothesis literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B.D., Sireci, S.G., & Hambleton, R.K. (2003, April). *Re-visiting exploratory methods for construct comparability: Is there something to be gained from the ways of old?* Paper presented in the symposium Construct Comparability Research: Methodological Issues and Results, at the National Council on Measurement in Education Annual Meeting, Chicago, IL.

Zumbo, B.D., & Zimmerman, D.W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology, 34*, 390-400.

APPENDIXES

Appendix A

Note that items 4, 8, 12, and 16 are reversed coded before analysis with the MGCFA, therefore their skew will be the opposite of what is displayed in this Appendix.

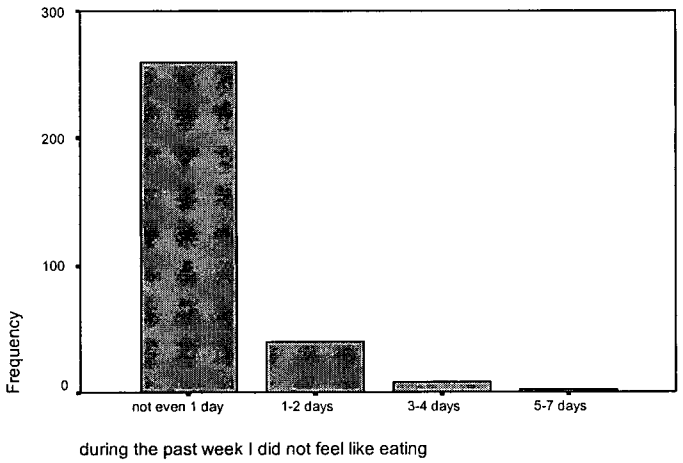


Figure A1. Distribution of responses on CES-D item 2 (I did not feel like eating) for males.

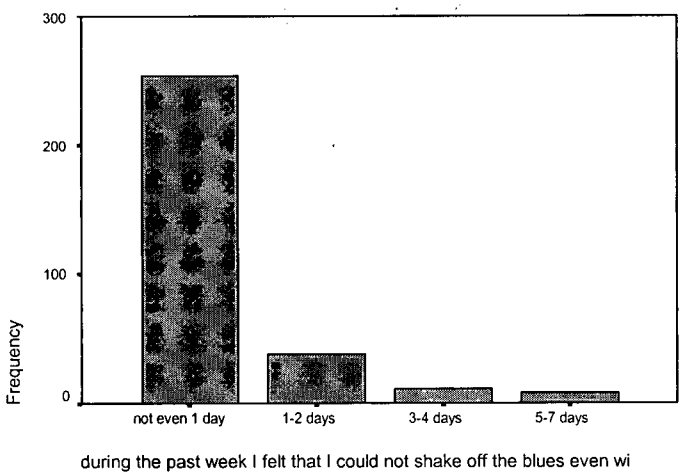


Figure A2. Distribution of responses on CES-D item 3 (I felt that I could not shake off the blues even with help) for males.

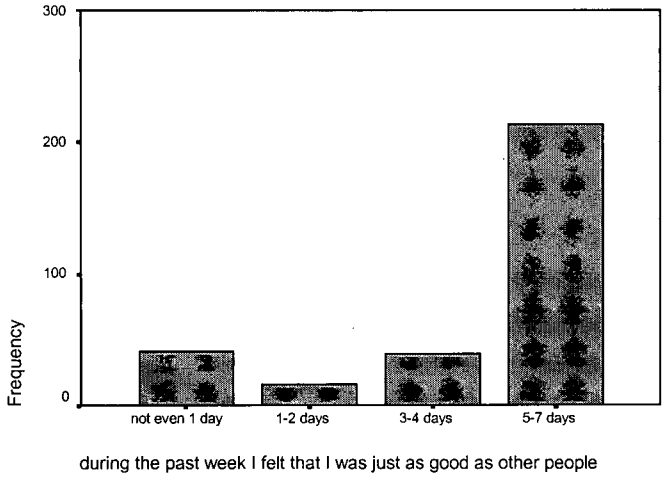


Figure A3. Distribution of responses on CES-D item 4 (I felt that I was just as good as other people) for males.

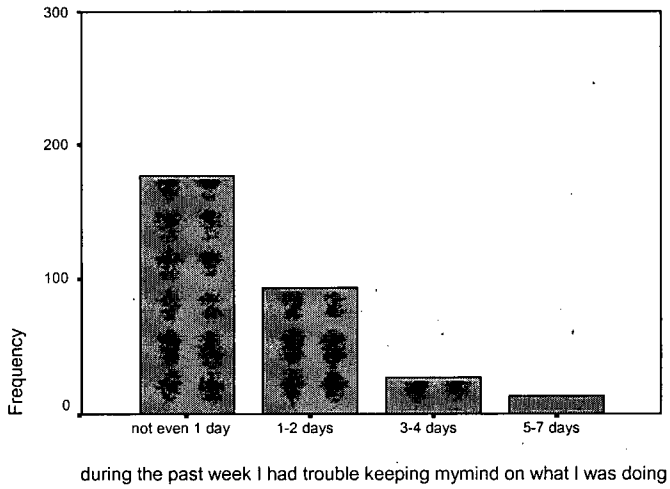


Figure A4. Distribution of responses on CES-D item 5 (I had trouble keeping my mind on what I was doing) for males.

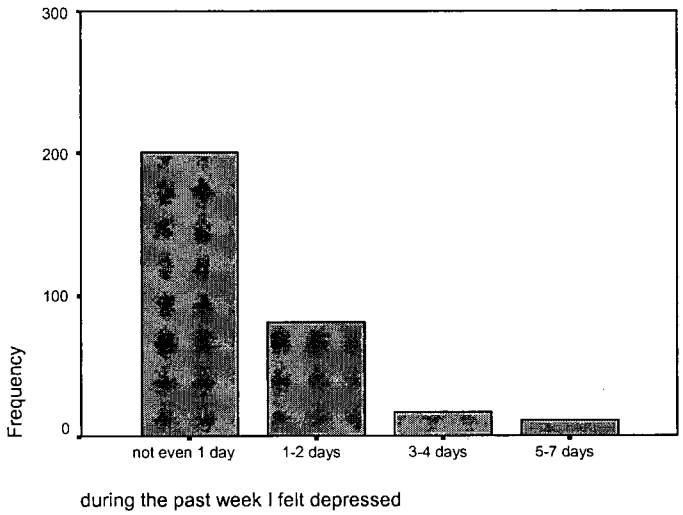


Figure A5. Distribution of responses on CES-D item 6 (I felt depressed) for males.

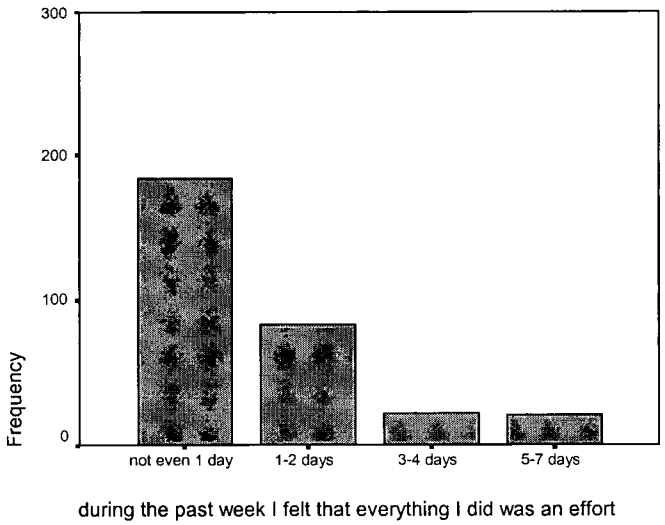


Figure A6. Distribution of responses on CES-D item 7 (I felt that everything I did was an effort) for males.

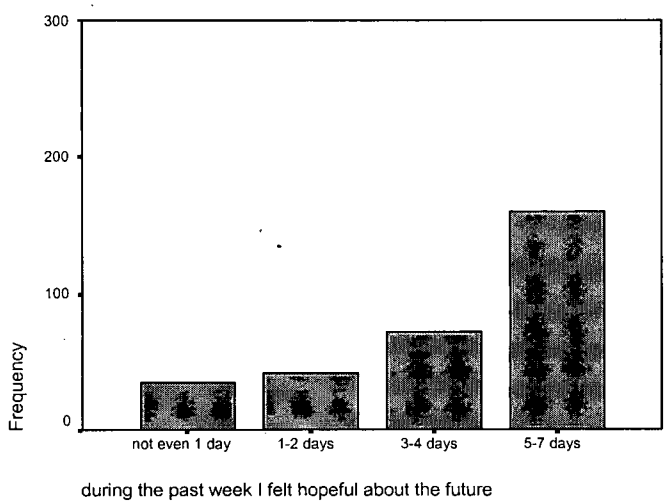


Figure A7. Distribution of responses on CES-D item 8 (I felt hopeful about the future) for males.

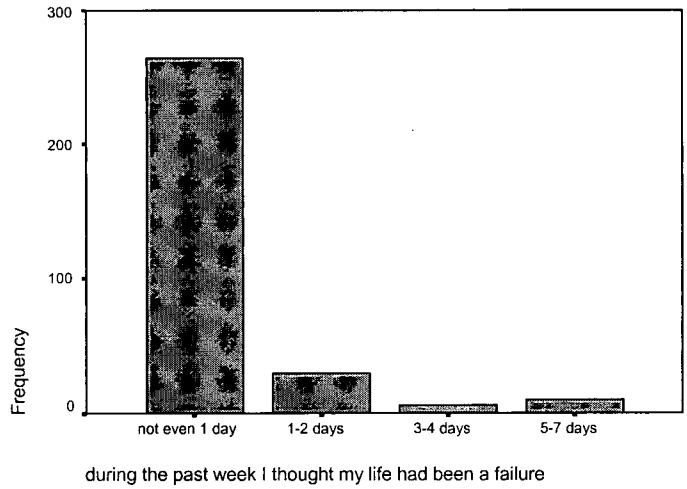


Figure A8. Distribution of responses on CES-D item 9 (I thought my life had been a failure) for males.

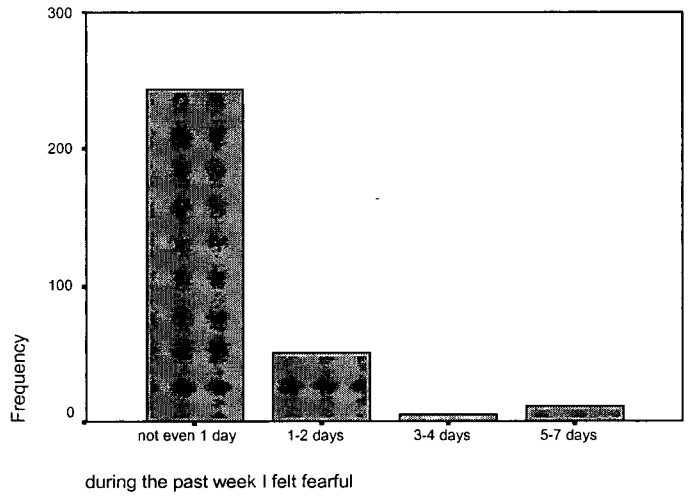


Figure A9. Distribution of responses on CES-D item 10 (I felt fearful) for males.

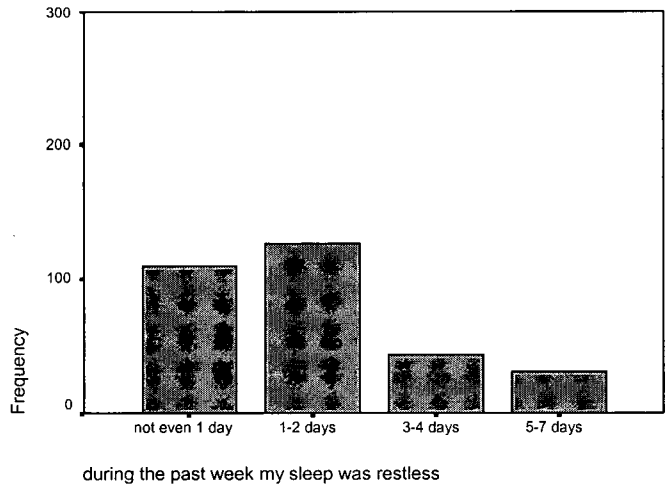


Figure A10. Distribution of responses on CES-D item 11 (My sleep was restless) for males.

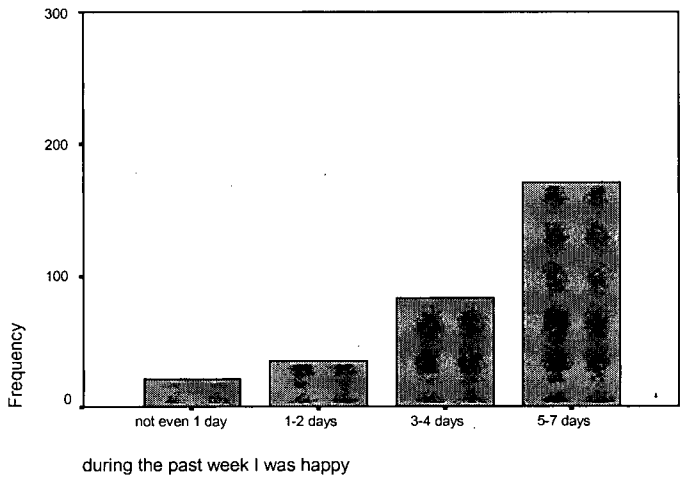


Figure A11. Distribution of responses on CES-D item 12 (I was happy) for males.

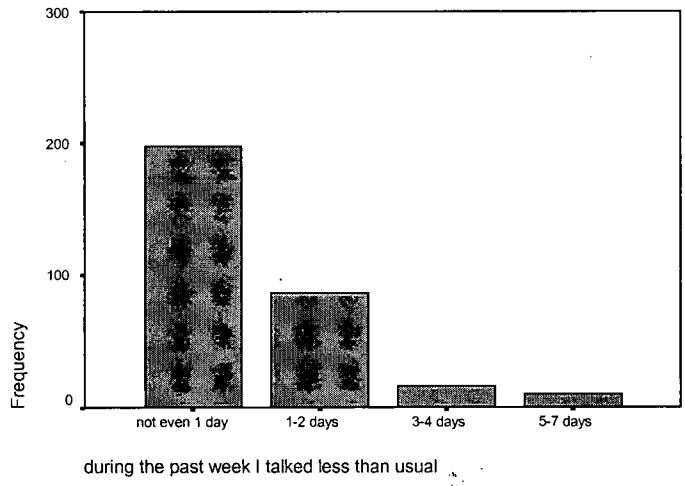


Figure A12. Distribution of responses on CES-D item 13 (I talked less than usual) for males.

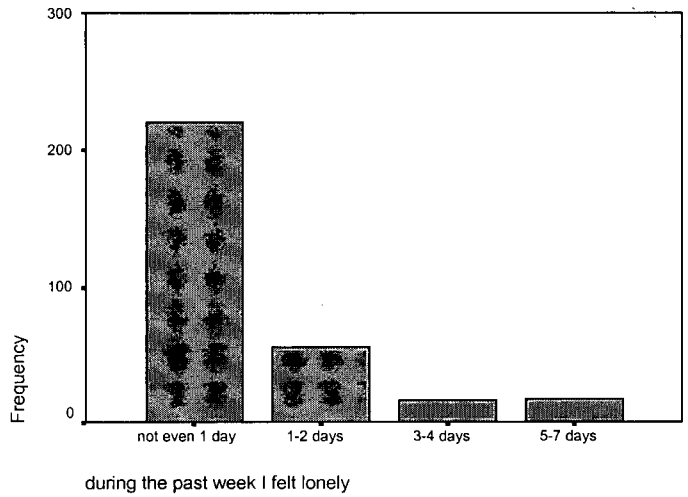


Figure A13. Distribution of responses on CES-D item 14 (I felt lonely) for males.

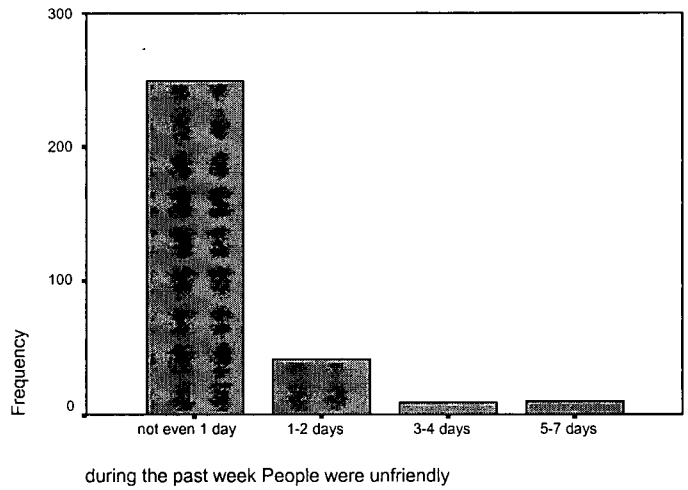


Figure A14. Distribution of responses on CES-D item 15 (People were unfriendly) for males.

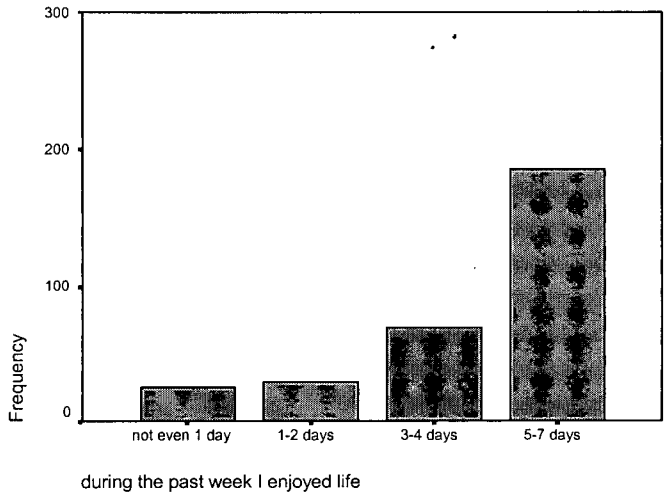


Figure A15. Distribution of responses on CES-D item 16 (I enjoyed life) for males.

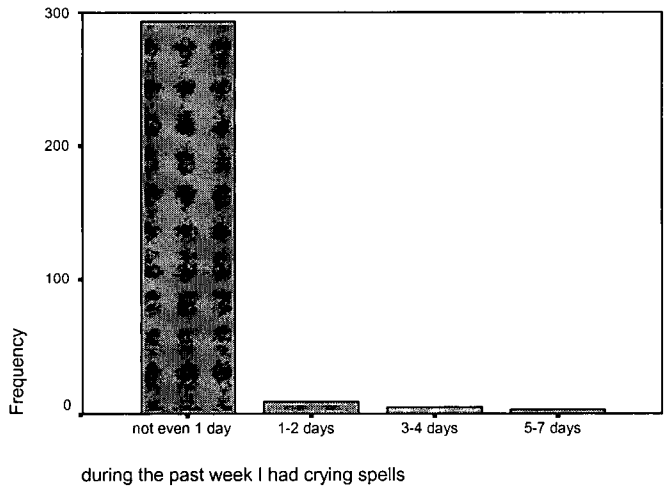


Figure A16. Distribution of responses on CES-D item 17 (I had crying spells) for males.

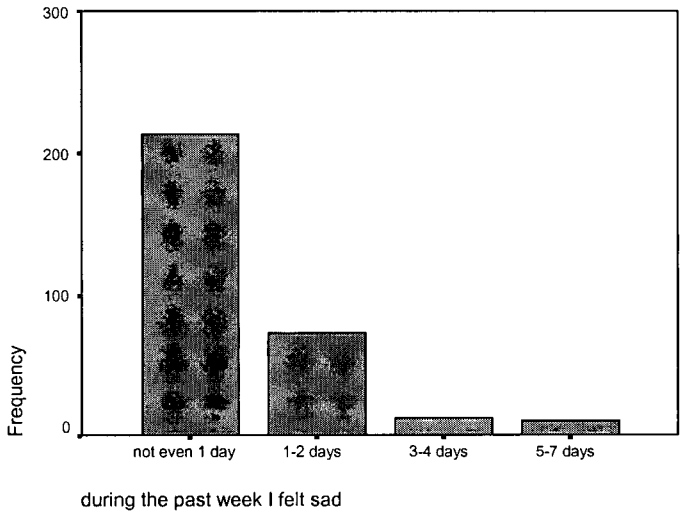


Figure A17. Distribution of responses on CES-D item 18 (I felt sad) for males.

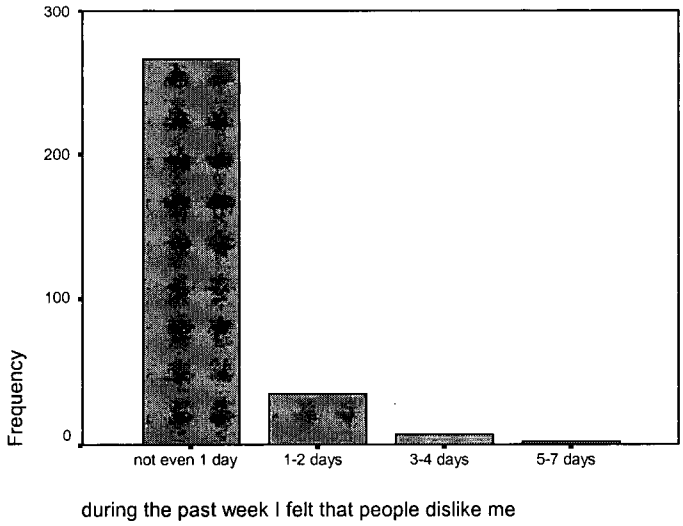


Figure A18. Distribution of responses on CES-D item 19 (I felt that people dislike me) for males.

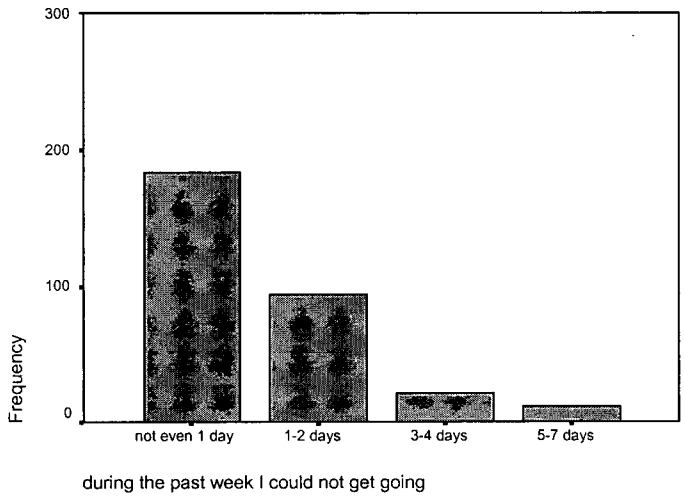


Figure A19. Distribution of responses on CES-D item 20 (I could not get going) for males.

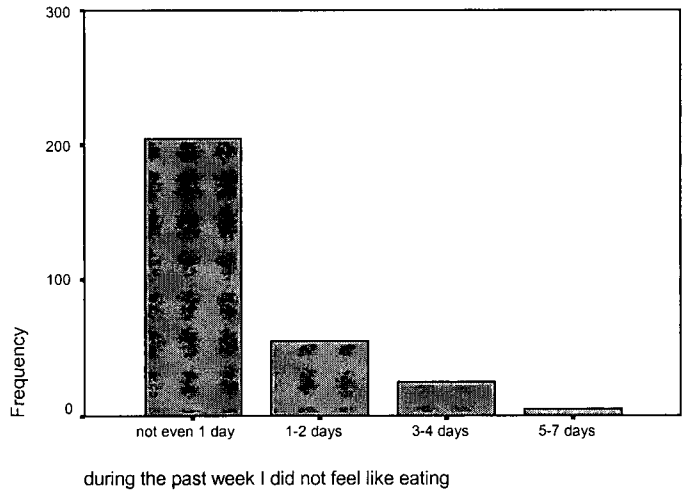


Figure A20. Distribution of responses on CES-D item 2 (I did not feel like eating) for females.

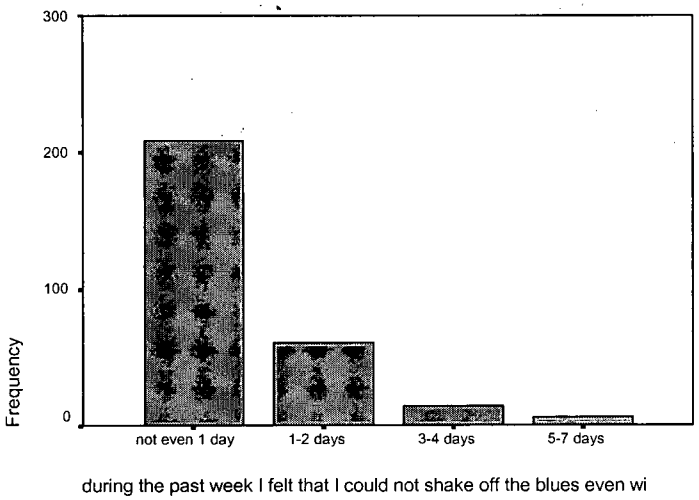


Figure A21. Distribution of responses on CES-D item 3 (I felt that I could not shake off the blues even with help) for females.

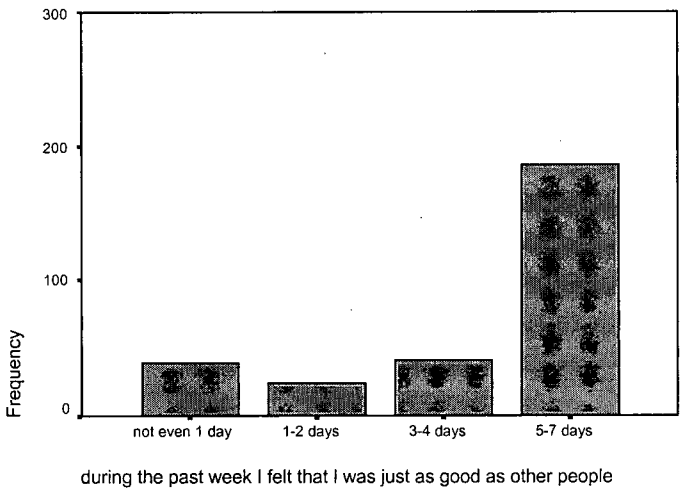


Figure A22. Distribution of responses on CES-D item 4 (I felt that I was just as good as other people) for females.

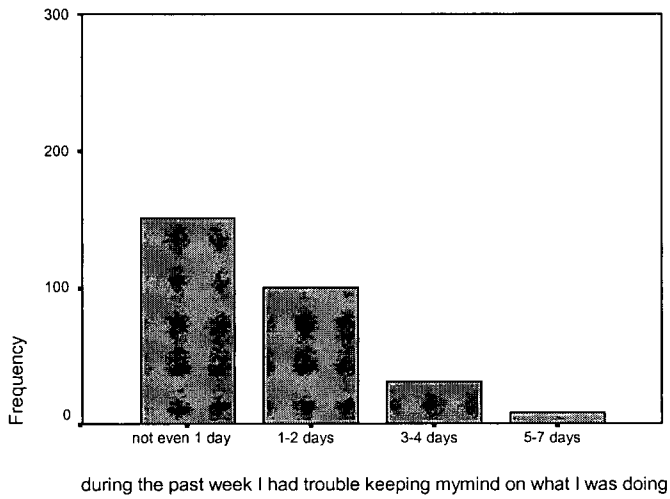


Figure A23. Distribution of responses on CES-D item 5 (I had trouble keeping my mind on what I was doing) for females.

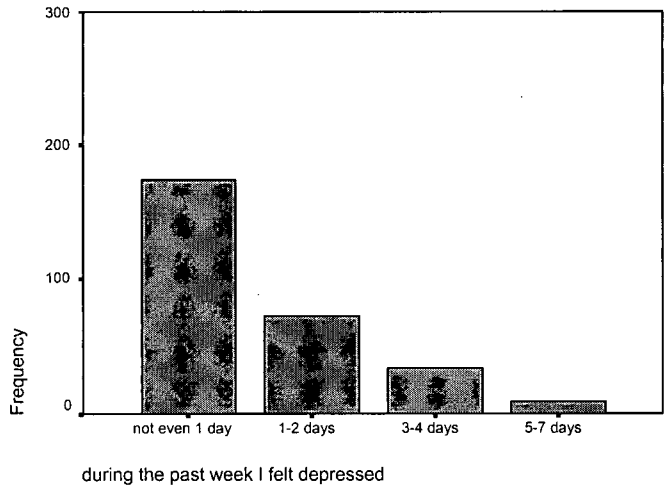


Figure A24. Distribution of responses on CES-D item 6 (I felt depressed) for females.

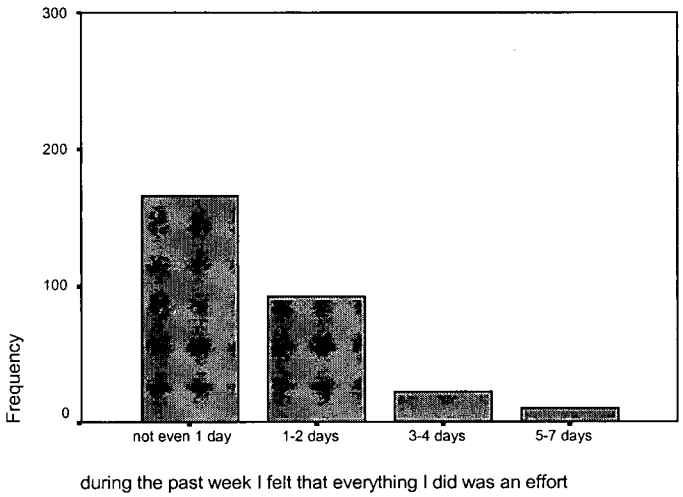


Figure A25. Distribution of responses on CES-D item 7 (I felt that everything I did was an effort) for females.

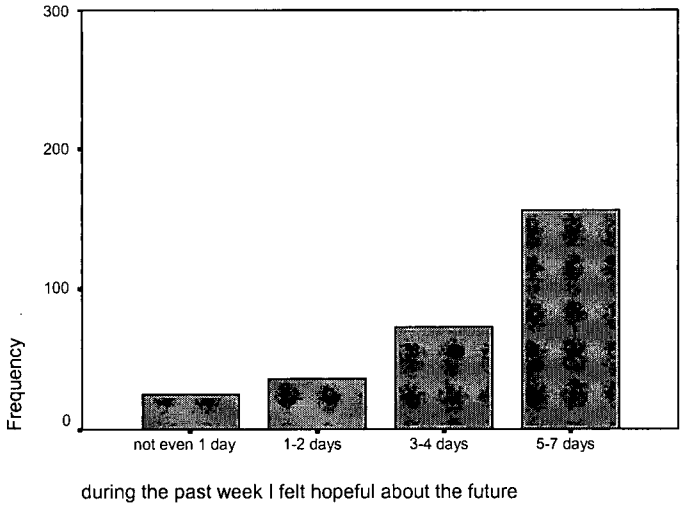


Figure A26. Distribution of responses on CES-D item 8 (I felt hopeful about the future) for females.

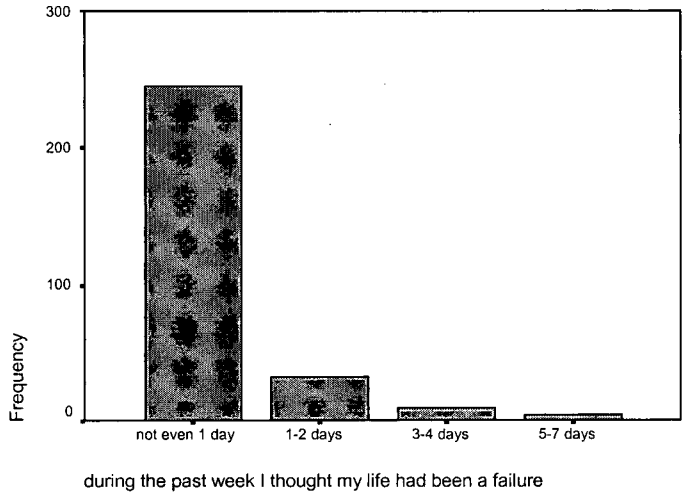


Figure A27. Distribution of responses on CES-D item 9 (I thought my life had been a failure) for females.

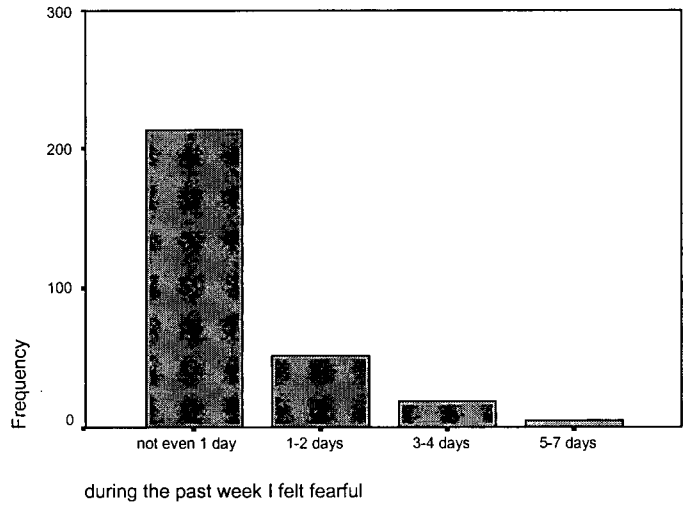


Figure A28. Distribution of responses on CES-D item 10 (I felt fearful) for females.

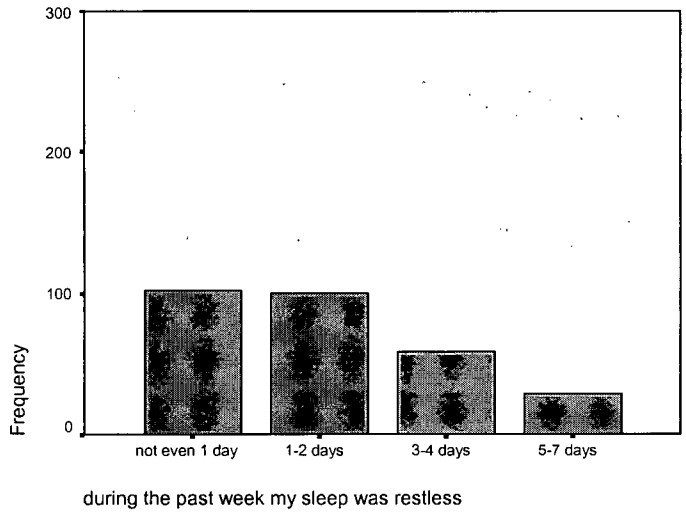


Figure A29. Distribution of responses on CES-D item 11 (My sleep was restless) for females.

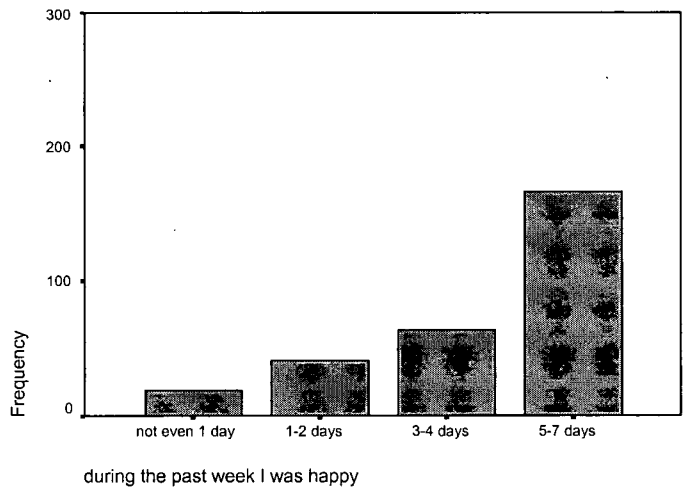


Figure A30. Distribution of responses on CES-D item 12 (I was happy) for females.

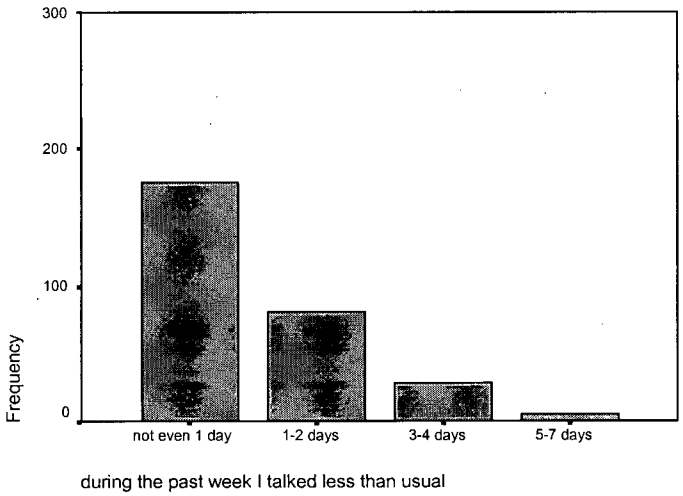


Figure A31. Distribution of responses on CES-D item 13 (I talked less than usual) for females.

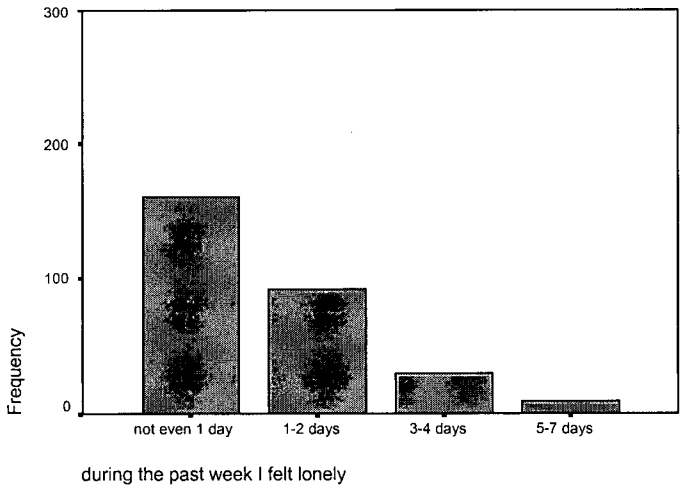


Figure A32. Distribution of responses on CES-D item 14 (I felt lonely) for females.

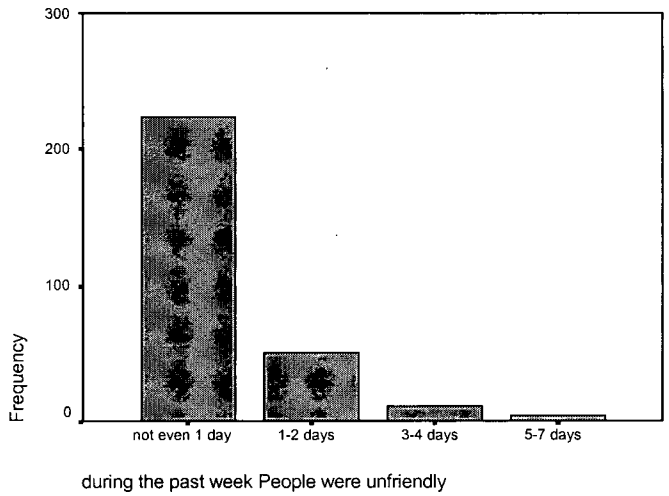


Figure A33. Distribution of responses on CES-D item 15 (People were unfriendly) for females.

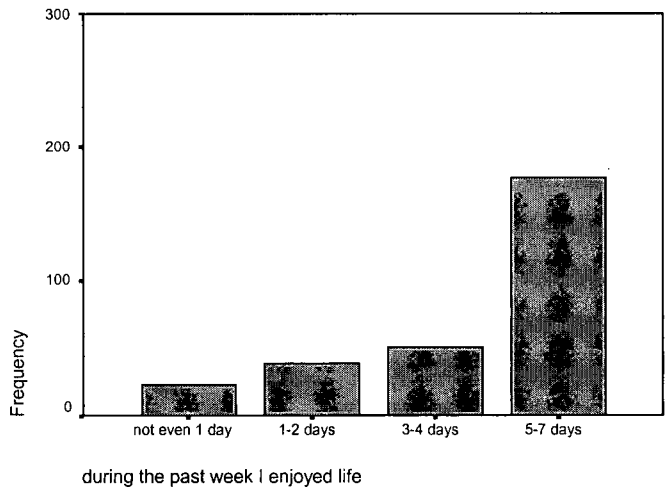


Figure A34. Distribution of responses on CES-D item 16 (I enjoyed life) for females.

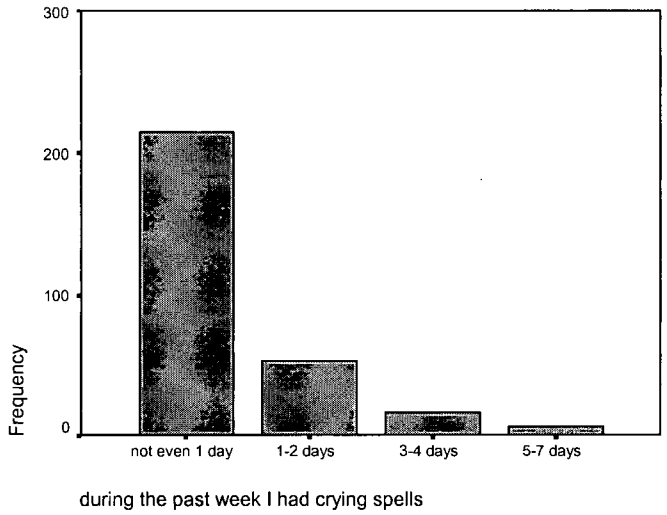


Figure A35. Distribution of responses on CES-D item 17 (I had crying spells) for females.

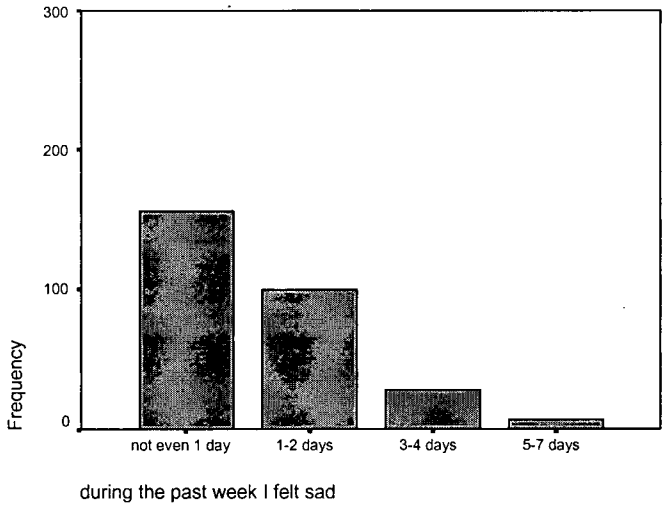


Figure A36. Distribution of responses on CES-D item 18 (I felt sad) for females.

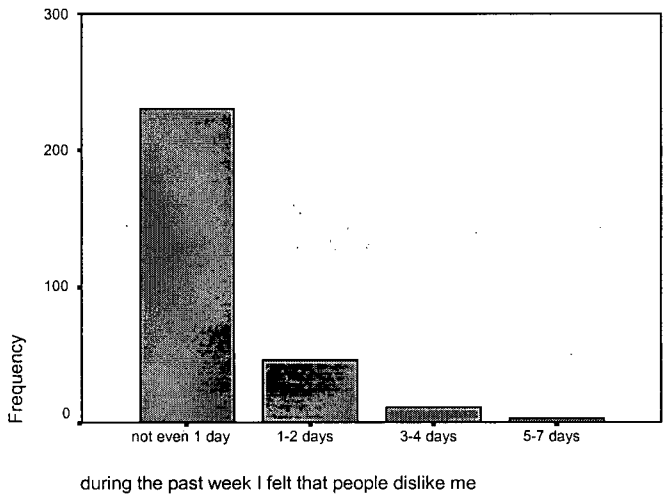


Figure A37. Distribution of responses on CES-D item 19 (I felt that people dislike me) for females.

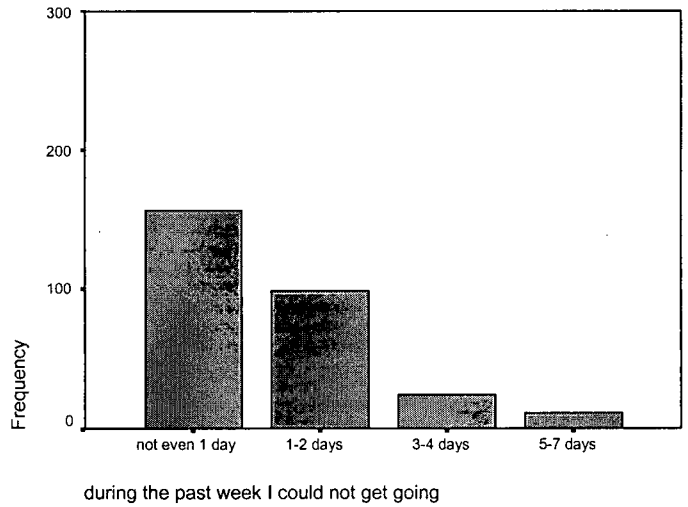


Figure A38. Distribution of responses on CES-D item 20 (I could not get going) for females.

Appendix B

Scale Interval Thresholds

Table B1

Threshold values for the scale intervals in symmetric response distribution with equal intervals

1	2	3	4	5	6	7	8
0.0000							
-1.0000	1.0000						
-1.5000	0.0000	1.5000					
-1.8000	-0.6000	0.6000	1.8000				
-2.0000	-1.0000	0.0000	1.0000	2.0000			
-2.1429	-1.2857	-0.4286	0.4286	1.2857	2.1429		
-2.2500	-1.5000	-0.7500	0.0000	0.7500	1.5000	2.2500	
-2.3333	-1.6667	-1.0000	-0.3333	0.3333	1.0000	1.6667	2.3333

Table B2

Threshold values for the scale intervals in positively skewed response distribution (left bunching) with unequal intervals

1	2	3	4	5	6	7	8
1.5000							
0.0000	1.5000						
0.0000	1.0000	2.0000					
0.0000	0.7500	1.5000	2.2500				
0.0000	0.6000	1.2000	1.8000	2.4000			
0.0000	0.5000	1.0000	1.5000	2.0000	2.5000		
0.0000	0.4286	0.8571	1.2857	1.7143	2.1429	2.5714	
0.0000	0.3750	0.7500	1.1250	1.5000	1.8750	2.2500	2.6250

Appendix C

IRT Parameters for the Binary Items in TIMSS Mathematics Achievement Test

Item #	a	SE for a	b	SE for b	c	SE for c
1	0.807	0.033	-0.176	0.051	0.196	0.020
2	0.939	0.039	-0.137	0.043	0.172	0.018
3	1.390	0.093	0.940	0.033	0.428	0.010
4	0.881	0.045	0.088	0.053	0.273	0.020
5	1.096	0.059	0.141	0.046	0.396	0.017
6	0.650	0.031	0.329	0.039	0	0
7	0.827	0.055	-0.684	0.093	0.194	0.037
8	1.258	0.109	0.777	0.045	0.236	0.018
9	0.844	0.062	0.263	0.065	0.153	0.025
10	1.014	0.129	1.353	0.073	0.301	0.018
11	0.886	0.086	0.942	0.061	0.188	0.021
12	0.702	0.065	0.507	0.085	0.176	0.029
13	1.858	0.110	0.402	0.026	0.120	0.012
14	0.916	0.076	-0.167	0.094	0.308	0.035
15	1.497	0.136	1.259	0.049	0.281	0.013
16	0.980	0.083	-0.533	0.107	0.366	0.040
17	0.391	0.037	-2.464	0.284	0.219	0.056
18	0.556	0.041	-1.827	0.175	0.205	0.052
19	0.841	0.072	-0.088	0.098	0.281	0.036
20	1.353	0.097	0.625	0.037	0.160	0.016

Note. a = slope parameter; b = location parameter; c = guessing parameter; SE = standard error.

Appendix D

IRT Parameters for the 3-point Polytomous Items in TIMSS Mathematics Achievement Test

Item #	a	SE for a	b	SE for b	d1	SE for d1	d2	SE for d2
1	0.466	0.011	0.662	0.023	-0.887	0.047	0.887	0.052
2	0.459	0.011	0.752	0.026	-0.910	0.051	0.910	0.057
3	0.569	0.013	0.525	0.020	-1.277	0.053	1.277	0.056
4	0.869	0.028	1.238	0.026	0.095	0.029	-0.095	0.043
5	0.497	0.010	0.425	0.020	-1.491	0.056	1.491	0.058
6	0.828	0.021	0.917	0.018	-0.120	0.028	0.120	0.035
7	0.652	0.012	0.299	0.016	-1.647	0.053	1.647	0.054
8	0.557	0.012	1.250	0.025	-2.077	0.070	2.077	0.076
9	0.815	0.016	0.445	0.014	-0.652	0.032	0.652	0.034
10	0.863	0.017	0.791	0.014	-1.342	0.048	1.342	0.050
11	0.755	0.018	0.381	0.015	0.093	0.027	-0.093	0.029
12	0.780	0.022	0.837	0.023	-1.893	0.094	1.893	0.097
13	0.648	0.012	0.454	0.016	-1.520	0.052	1.520	0.053
14	0.585	0.011	0.284	0.017	-1.887	0.059	1.887	0.060
15	1.719	0.053	0.687	0.013	-0.701	0.044	0.701	0.045
16	0.861	0.016	0.372	0.013	-1.073	0.039	1.073	0.040
17	0.739	0.017	0.332	0.016	0.114	0.027	-0.114	0.030
18	0.525	0.011	1.319	0.027	-2.195	0.075	2.195	0.081
19	0.892	0.017	0.420	0.013	-0.537	0.030	0.537	0.031
20	1.057	0.021	0.760	0.013	-1.011	0.040	1.011	0.041

Note. a = slope parameter; b = location parameter; d = step parameter; SE = standard error.

Appendix E

Criteria for Identifying Inflated Empirical Type I Error Rates of ML Chi-square Statistics

Two-tailed confidence interval (at a Bonferroni corrected confidence interval of 99%)

empirical alpha	*Lower*	*Upper*
.15	.058	.242
.14	.051	.230
.13	.043	.217
.12	.036	.204
.11	.029	.191
.10	.023	.177
.09	.016	.164
.08	.010	.150
.07	.004	.136

Note: The confidence intervals were computed based on the normal approximation.