MEASUREMENT INVARIANCE OF HEALTH-RELATED QUALITY OF LIFE: A SIMULATION STUDY AND NUMERIC EXAMPLE

BY

JOYKRISHNA SARKAR

A Thesis Submitted to the Faculty of Graduate Studies in Partial Fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department of Community Health Sciences

University of Manitoba

Winnipeg, Manitoba, Canada

ABSTRACT

Measurement invariance (MI) is a prerequisite to conduct valid comparisons of Health-related quality of life (HRQOL) measures across distinct populations. This research investigated the performance of estimation methods for testing MI hypotheses in complex survey data using a simulation study, and demonstrates the application of these methods for a HRQOL measure. Four forms of MI were tested using confirmatory factory analysis. The simulation study showed that the maximum likelihood method for small sample size and low intraclass correlation (ICC) performed best, whereas the pseudomaximum likelihood with weights and clustering effects performed better for large sample sizes with high ICC to test configural invariance. Both methods performed similarly to test other forms of MI. In the numeric example, MI of one HRQOL measure in the Canadian Community Health Survey was investigated and established for Aboriginal and non-Aboriginal populations with chronic conditions, indicating that they had similar conceptualizations of quality of life.

ACKNOWLEDGEMENTS

I am grateful to my supervisor, Dr. Lisa M. Lix, for her diligent guidance, support and encouragement to prepare this thesis. Dr. Lix guided me through many details and devoted countless hours for the production of this thesis. It was really impossible to complete this document without her tireless assistance. Her generosity, kindness and endless patience have been very much appreciated.

I would like to express my sincere gratitude to my thesis committee members for their feedback and kind advice.

I would also like to thank Statistics Canada for providing funding to support this research through a Manitoba Research Data Centre Award.

Finally, I wish to thank the Western Regional Training Centre (WRTC) for Health Services Research for providing financial support during my MSc studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ALOCC	At least one chronic condition
ANOVA	Analysis of variance
BP	Bodily pain
BRR	Balanced repeated replication
CCHS	Canadian Community Health Survey
CES-D	Centre for Epidemiologic Studies Depression Scale
CFA	Confirmatory factor analysis
CFI	Comparative fit index
DF	Degrees of freedom
EORTC QLQ-C30	European Organization for Research and Treatment of Cancer Quality of Life Core Questionnaire
GH	General health perceptions
HRQOL	Health related quality of life
IBDQ	Inflammatory Bowel Disease Questionnaire
ICC	Intraclass correlation
JRR	Jackknife repeated replication
LRT	Likelihood ratio test
MH	Mental health
MI	Measurement invariance
Mini AQLQ	Mini Asthma Quality of Life Questionnaire
Mini OQOL	Mini Osteoporosis Quality of Life
ML	Maximum likelihood

NCC	No chronic condition
NHP	Nottingham Health Profile
NNFI	Non-normed fit index
PF	Physical functioning
PML	Pseudo-maximum likelihood
PML1	Pseudo-maximum likelihood with sample weights
PML2	Pseudo-maximum likelihood with sample weights and clusters
PPS	Probability proportionate to size
RE	Role limitation due to emotional problems
RMSEA	Root mean square error of approximation
RNI	Relative non-centrality index
RP	Role limitations due to physical health problems
SAS	Statistical Analysis Software
SCHIP	California State Children's Health Insurance Program
SD	Standard deviation
SEM	Structural equation modeling
SF	Social functioning
SF-36	Medical Outcomes Study Short Form
SIP	Sickness Impact Profile
SRMR	Standardized root mean square residual
SRS	Simple random sampling
TLI	Tucker-Lewis index

VT	Vitality
WHOQOL	World Health Organization’s quality of life-BREF

CHAPTER ONE: INTRODUCTION

1.1 Background

Health-related quality of life (HRQOL) is a multidimensional construct that encompasses physical, social, and psychological components of health, as well as general or global perceptions of health and well-being (Bergner, 1989; O'Boyle, 1992; Olschewski, Schilgen, & Schumacher, 1994; Ware, 1987; Wood-Dauphinee, 1999). The physical component often refers to the patient's perceived ability to carry out daily activities that require energy expenditure. The social component represents the ability to relate and integrate with members of one's family, neighborhood, workplace and community. The psychological component incorporates perceptions of emotional and mental well-being such as depression, anxiety, fear, anger, happiness, and peacefulness.

HRQOL is investigated in many studies where objective measures of health may not be adequate or sufficient to describe individual or population health. HRQOL is used in clinical trials to investigate the efficacy of new treatments or interventions and in population-based research to compare the health of different populations. HRQOL is also an important outcome measure for individuals with chronic disease (Lam & Lauder, 2000; Sprangers, de Regt, Andries, & van Agt, 2000). For example, it can be used to evaluate the effectiveness of different health care interventions or to monitor changes over time in health status.

The validity of research that uses a HRQOL measure depends, in part, on the psychometric properties of the measure. Psychometric properties such as validity and reliability have been investigated for a number of HRQOL measures in previous research (de Vet, Ader, Berwee, & Pouwer, 2005; Keller et al., 1998; Schlenk et al., 1998). It is

only recently that researchers have begun to focus on the measurement invariance (MI) properties of these measures. An instrument or measure is said to possess MI if it has the same interpretation or meaning across different study groups. Previous research (Gregorich, 2006; Mora et al., 2009) has shown that MI of HRQOL measures may not be tenable for different ethnic or cultural groups because individuals with different backgrounds may not interpret questions about their health in the same way. MI of HRQOL measures may be influenced by other factors, including sex, age, and health status. Testing for differences between groups using statistical techniques such as analysis of variance (ANOVA) or regression analysis without first establishing MI might lead to erroneous conclusions (Bollen, 1989), because true group differences may be confounded with measurement artifact.

The use of confirmatory factor analysis (CFA) to assess the MI of HRQOL in different populations is well documented (Vandenberg & Lance, 2000). CFA is a form of structural equation modeling (SEM). While the use of conventional CFA techniques to assess MI is well-established (Lix, Metge, & Leslie, 2009), there has been only limited consideration of the effect of the study design on the validity of these techniques.

Conventional CFA techniques refer to CFA techniques applied to study designs that adopt simple random sampling (SRS) of observations. Conventional CFA techniques, which are usually based on maximum likelihood (ML) estimation, ignore the dependencies among observations that arise in complex survey data when study participants are not selected using SRS. For example, study participants may be selected using a multistage cluster survey design, which results in non-independence of observations. Inferences about MI made using conventional SEM techniques with ML

estimation may be sensitive to the lack of independence of observations (Stapleton, 2006). In addition, in a multistage sample survey, sample weights are generally assigned at each stage of the sampling process to reflect the unequal sample inclusion probabilities (Pfeffermann, 1993). Sample weights are included in data files and are used in analysis to represent the population from which sample is drawn. The use of sample weights also affects the selection of model estimation methods. A number of methods have been proposed to incorporate both survey design effects and sample weights into SEM analyses. These include pseudo-maximum likelihood (PML) estimation (Asparouhov & Muthén, 2005; Muthén & Satorra, 1995; Stapleton, 2006; Stalpeton, 2008), jackknife repeated replication (JRR), balanced repeated replication (BRR), and bootstrap methods (Stapleton, 2008). However, these techniques have never been investigated or applied in the context of testing MI of HRQOL measures in different populations using CFA techniques.

1.2 Purpose and Objectives

The purpose of this research is to investigate techniques for testing hypotheses about MI in complex survey data and to demonstrate the application of these techniques to real-life HRQOL data. The specific objectives are:

(1) To compare ML and PML estimation methods by a simulation study for testing hypotheses about MI using CFA techniques in complex survey data; and

(2) To demonstrate the application of estimation methods and CFA techniques to test hypotheses about MI by ethnicity for one HRQOL measure.

1.3 Study Rationale

Assessing MI is an important prerequisite to test for differences in HRQOL across different groups defined by ethnic/cultural, demographic, or health variables. If MI is not a tenable assumption, group comparisons may produce misleading results (Hui & Triandis, 1985).

While CFA, a form of SEM, is widely used as a tool to test hypotheses about MI, it is not known what estimation method will produce valid results for complex survey data. Previous work by Stapleton (2008) compared a number of methods for testing hypotheses in SEM using χ^2 tests when the data were obtained from a complex survey design. She included PML method, taking the complex survey design into account and also ignoring the complex survey design in the estimation of model parameters. The PML method was also investigated in previous research for making inferences using the χ^2 test, by including and ignoring the characteristics of a survey design in the analysis (Asparouhov & Muthén, 2005). In particular, the effects of clustering and stratification were examined for estimating the χ^2 statistic. However, these methods have never been compared for testing hypotheses about MI using CFA for complex survey data. Simulation studies have been used in previous research to compare the performance of different methods for testing hypotheses in SEM.

This research will help to address an important gap in the methodological literature about SEM. It will also add to the measurement literature on MI of HRQOL measures across different ethnic groups.

CHAPTER TWO: LITERATURE REVIEW

2.1 Measuring Health-Related Quality of Life

There are over 800 generic and specific instruments that have been developed to measure HRQOL (Guyatt, Feency, & Patrick, 1993; Testa & Simonson, 1996). Generic measures are based on a global conceptualization of HRQOL. These measures allow researchers to investigate various domains of health across populations and disease states (Guyatt et al., 1993; Testa & Simonson, 1996). Generic HRQOL measures include the Sickness Impact Profile (SIP; Bergner, Bobbitt, Carter, & Gilson, 1981), Nottingham Health Profile (NHP; Hunt, McEwen, & McKenna, 1985), World Health Organization WHOQOL–BREF (WHOQOL Group, 1998a), PedsQL 4.0 Generic Core Scale (Limbers, Newman, & Varni, 2008) and 36-item Medical Outcomes Study Short Form (SF-36; Ware & Sherbourne, 1992). These measures have been administered in cross-cultural (Sheila, 2005), cross-national (Keller et al., 1998), and cross-ethnic (Crockett, Shen, Randall, Russell, & Driscoll, 2005; Lix et al., 2009) studies.

Generic HRQOL measures include health profiles and health indices (Camilleri-Brennan & Steele, 1999). Health profiles usually cover a wide range of health related domains and a separate score is computed for each domain (Guyatt et al., 1993). The SF-36 is a well-known health profile (Ware & Sherbourne, 1992). It is useful in surveys of general and specific populations for comparing the relative burden of diseases, and in differentiating the health benefits produced by a wide range of different treatments. The use of the SF-36 has been investigated in chronic disease populations to compare HRQOL across different groups.

The SF-36 is a popular HRQOL measure for several reasons. It is available at no cost to researchers, easy to administer within a short period of time, and is a widely accepted general measure of quality of life (Jordan-Marsh, 2002). The SF-36 is an appropriate measure of HRQOL across different diseases (Yao & Wu, 2005). The SF-36 has been translated into many languages and used in many countries. It has been documented in nearly 4,000 publications (Turner-Bowker, Bartley, & Ware, 2002).

The reliability and construct validity of the SF-36 has been established (Schlenk et al., 1998; de Vet et al., 2005). Previous research has shown that the internal consistency coefficients of the eight scales that comprise this instrument ranged from 0.62 to 0.96 with a median of 0.80. The test-retest reliabilities of the scales ranged from 0.43 to 0.90 with a median of 0.64 after six months, but these values were from 0.60 to 0.81 with a median of 0.76 after two weeks in patients with diabetes (Schlenk et al., 1998). The construct validity of the SF-36 has been proven as well. For example, the eight scales or domains discriminated between groups differing in physical mobility, and seven of the eight scales were sensitive to clinically defined differences in mental health (Schlenk et al., 1998). The mental health scale score was lower for patients with a psychiatric disorder than for patients with minor medical conditions. The correlations, from -0.038 to -0.75, with the Centre for Epidemiologic Studies Depression Scale (CES-D) measure, also supported the construct validity of the SF-36. MI of the SF-36 has been investigated in different ethnic groups (Lix et al., 2009), although the number of papers on this topic is sparse.

Specific HRQOL measures have been developed for particular diseases or conditions, or to allow for in-depth investigation of an individual health domain (Guyatt

et al., 1993). For example, the European Organization for Research and Treatment of Cancer Quality of Life Core Questionnaire (EORTC QLQ-C30) is a disease-specific measure (Aaronson et al., 1993) that has been widely-used in international clinical trials in oncology. Other examples include the Hospital Anxiety and Depression Scale which is used to assess the psychological domain (Zigmond & Snaith, 1983), the mini-Osteoporosis Quality of Life (mini-OQOL) questionnaire which is used to assess HRQOL of individuals with osteoporosis (Lix et al., 2009), and the Inflammatory Bowel Disease Questionnaire (IBDQ; Guyatt et al., 1989).

2.2 Testing Measurement Invariance Using Conventional Structural Equation Modeling Techniques

Some constructs that are commonly investigated in health research, such as HRQOL, intellectual ability, depression, anxiety, and attitude, cannot be measured directly. Measurement of these constructs is very important in decision making in a variety of environments (French & Finch, 2006). Researchers must ensure that measurements of the same attribute are equivalent under different conditions (e. g., stability of measurements over time, across different populations defined by characteristics such as age or sex, across rater groups or over different modes of instrument administration).

In SEM, a measurement model defines an association among a set of observed variables and latent variables (i.e., factors). The measurement model can be represented using the regression equation,

$$y_{nj} = \tau_j + \lambda_{j1}\eta_{1n} + \lambda_{j2}\eta_{2n} + \ldots\ldots\ldots\ldots + \lambda_{jp}\eta_{pn} + e_{nj}, \quad (1)$$

where y_{nj} denotes the nth (n = 1, 2,……, N) person's score on the jth (j = 1, 2,…., J) observed variable, τ_j is the intercept at which the latent variable score is zero, λ_{jp} are regression coefficients for observed variable j on the kth latent variable (k = 1,…, p), η_{kn} is the kth latent variable, and e_{nj} is the error term for the nth individual. The direct effects of latent variables on observed variables are also called factor loadings.

There are a number of forms of MI that can be tested using CFA techniques in SEM. The tests are conducted by applying constraints to the regression coefficients of equation 1 across independent groups of study participants. Factor loadings may be free, fixed, or constrained depending on the researcher's specifications. A free parameter is estimated, whereas a fixed parameter is specified to be equal to a constant. A constrained factor loading is a regression coefficient that is estimated under some model restriction. The recommended forms of MI tests are: (i) configural or pattern invariance: a test of equality of the pattern of factor loadings across groups, (ii) metric or weak invariance: a test of equality of factor loadings across groups, (iii) scalar or strong invariance: a test that factor loadings and intercepts of like items (i.e., same indicators) are invariant across groups, (iv) complete or strict invariance: a test that like item factor loadings, means, and error variances are equivalent across groups (Muthén & Muthén, 2007; Vandenberg & Lance, 2000). If configural invariance is satisfied then it can be concluded that the same construct or pattern of fixed and free parameters is being measured in each group. The model of configural invariance serves as a baseline model to which more restrictive models are compared. When metric or weak invariance is established, this implies that the same latent variables (i.e., factors) are being measured across groups. Like in metric invariance, scalar invariance also implies that the measurement of the latent variables is

the same across groups. Moreover, the invariance of intercepts allows evaluating mean differences in latent variables across groups. Any differences in means of the indicators are attributable to differences in means on the latent variables. If scalar invariance is not satisfied then a comparison across groups for indicator means will not be valid (Meredith & Teresi, 2006). The key difference between complete and scalar invariance involves how the variances of the indicators are accounted for. In complete invariance, group differences in variances of indicators are attributable only to group differences in variances of latent variables since error variances are invariant across groups. Complete invariance is a highly constrained model and may often not hold in practice, even if scalar invariance does hold. If complete invariance is established then comparisons across groups on the global or domain scores should be unbiased (Lix et al., 2009).

To test different forms of MI, two identified measurement models must be specified for each group. In one model (i.e., the constrained model), one or more constraints on the model parameters are specified, while in the second model (i.e., the unconstrained model), these constraints are removed. The models are nested, that is, the unconstrained model is a special case of the constrained model. In order to be identified, a CFA model must have the following characteristics: (1) the number of parameters to be estimated is less than or equal to the number of observations (i.e., $j(j + 1)/2$, where j is the number of indicators); (2) every latent variable must have a scale; and (3) a model with a single latent variable must have at least three observed variables, and a model with two or more latent variables must have at least two observed variables per latent variable (Kline, 2005). In a standard CFA model, each observed variable is represented by a

single underlying latent variable and an error term, the error terms are independent, and the latent variables are assumed to be correlated (Kline, 2005).

An important distribution that is widely used for model fitting is the χ^2 distribution. This distribution is defined as

$$f(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}, \quad x > 0, \quad \nu = 1, 2, \ldots, \tag{2}$$

where ν is the degrees of freedom, e is the base of the natural logarithm and Γ is gamma function. A variety of indices have been proposed to evaluate the overall fit of the initial measurement model and also to test hypotheses about MI by comparing two measurement models. The use of the large-sample χ^2 test is recommended in conjunction with other quantitative fit indices (Marsh & Yeung, 1996) to test the goodness of fit of the initial measurement model. A non-significant χ^2 test indicates adequate fit of the initial measurement model; however this test is sensitive to sample size and tends to reject the null hypothesis too often when sample size is large. Other fit indices, such as Tucker-Lewis index (TLI; Tucker & Lewis, 1973) or non-normed fit index (NNFI; Bentler & Bonett, 1980), relative non-centrality index (RNI; Bentler, 1990), root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR) are therefore recommended to evaluate the fit of the initial measurement model (Vandenberg & Lance, 2000). The fit indices TLI and RNI reflect the systematic variation in model misspecification and are not systematically related to sample size. TLI and RNI values of 0.90 or above are indicative of well-fitted models (Hu & Bentler, 1999). The RMSEA is also sensitive to model misspecification, especially to misspecified factor loadings (Hu & Bentler, 1999). The value of RMSEA is provided with its 90% confidence intervals. A common rule of thumb is that the value of RMSEA

less than equal to 0.05 implies close approximate fit; the values of RMSEA between 0.06 and 0.10 indicate acceptable fit; and the values of RMSEA greater than 0.10 indicate poor approximate fit (Browne & Cudeck, 1993). The SRMR is sensitive to model misspecification among the covariances; a value of 0.08 is recommended as an indication of a well-fitted model, with the value of 0.10 as an upper limit.

To test different forms of MI, the χ^2 test or likelihood ratio test (LRT) is commonly employed. This test statistic is computed as twice the difference between the log-likelihood of two nested models, i.e.,

$$\chi^2_{ML} = 2\,(L_C - L_{UC}), \tag{3}$$

where L_C and L_{UC} are log-likelihood values of the constrained and unconstrained models, respectively and are computed using conventional ML estimation. The conventional LRT statistic follows a χ^2 distribution with the degrees of freedom equal to the difference in the number of parameters between the unconstrained and constrained models. Moreover, difference in LRT statistics for two nested models (i.e.,$\Delta\ \chi^2$) is also a χ^2 distribution and is used for testing different forms of MI. No other fit indices used to test MI have known distributions (Vandenberg & Lance, 2000). However, one fit index that has been recommended for testing MI along with LRT is the comparative fit index (CFI; Yuan, 2005),

$$CFI_{ML} = 1 - \frac{\max[(\chi^2_C - df_C), 0]}{\max[(\chi^2_{UC} - df_{UC}), (\chi^2_C - df_C)]}, \tag{4}$$

where χ^2_C and χ^2_{UC} are χ^2 statistics evaluated at the constrained and unconstrained models using conventional ML, respectively; df_C and df_{UC} are the corresponding degrees

of freedom. The CFI assesses the relative improvement in model fit compared to the baseline or null model. It is assumed that a value of CFI greater or equal to 0.90 indicates an acceptable fit of the proposed model to the data (Hu & Bentler, 1999). This fit index is recommended to use for MI tests because it performs very well at all sample sizes (Bentler, 1990). It is also recommended that the difference in CFI of 0.01 or less between constrained and less constrained models (i.e.,ΔCFI) indicates that null hypothesis of MI should not be rejected (Bentler, 1990). Both χ^2_{ML} and CFI_{ML} were developed for conducting conventional SEM analyses.

2.3 Measurement Invariance of Health-Related Quality of Life Measures

MI of several different HRQOL measures has been investigated in previous research; the majority of these studies have used CFA techniques to test MI. In cross-sectional studies, researchers have investigated the MI of HRQOL instruments across independent population groups (Limbers et al., 2008; Lix et al., 2009; Yao & Wu, 2005), while in longitudinal studies, researchers have investigated MI over time within a single population group (Ahmed et al., 2005; Feldt et al., 2007; Krause, Kaltman, Goodman, & Dutton, 2007; Makikangas et al., 2006; Varni, Limbers, Newman, & Seid, 2008).

A recent study (Lix et al., 2009) investigated the MI of the SF-36 and mini OQOL measures for Canadian Aboriginal and non-Aboriginal women using CFA techniques. Four forms of MI (configural, metric, scalar and complete invariance) were tested for both measures. All forms of MI were satisfied for the SF-36 in Aboriginal and non-Aboriginal women, indicating that researchers can make valid comparisons of health across these two groups using the SF-36. Configural and metric invariance were established for the mini OQOL measure, which indicate that Aboriginal and non-

Aboriginal women have equivalent conceptualizations about osteoporosis quality of life. But scalar and complete invariance were not satisfied in these two groups, meaning that a valid comparison between groups is not possible.

The MI of the CES-D was investigated for Anglo and Latino adolescents (Crockett et al., 2005). Configural, metric, and scalar invariance were tested for the two ethnic groups. Configural invariance was established between groups of Anglo and Mexican Americans adolescents whereas for Cuban and Puerto Rican adolescents a lack of configural invariance was observed. Partial metric invariance was obtained for adolescents of Anglo and Mexican Americans. Scalar invariance was observed for all ethnic groups. Therefore, the authors concluded that the CES-D is useful in cross-ethnic research for assessment and treatment of depression.

The MI of the WHOQOL was examined for healthy and chronic disease populations and for different disease groups (Yao & Wu, 2005). In Yao and Wu's study, healthy groups of individuals were matched to individuals in each of five disease groups such as pulmonary, hypertension, peptic ulcer, sinusitis, and liver disease. Age and gender were used to match individuals in healthy and in a particular disease group. Multi-group CFA was employed on a four-factor (first-order factors) model with a general quality of life factor (second-order factor). The four factors were physical, psychological, social relationship and environment. This study examined configural and metric invariance between healthy and disease groups as well as between each pair of five disease groups. Configural invariance was satisfied for disease groups and their matched healthy groups, indicating that the four-factor model was acceptable in all groups. Metric invariance was also found for disease and healthy groups, suggesting that disease groups

and the matched healthy groups may have the same interpretation about the items of the WHOQOL. In addition, for each pair of five disease groups, metric invariance was satisfied; that is, the disease groups have the same perceptions about the WHOQOL questionnaire as other disease groups.

MI of another disease-specific HRQOL measure, the Mini Asthma Quality of Life Questionnaire (Mini AQLQ), was assessed in a sample of Latino and African-American asthmatic patients (Mora et al., 2009). CFA was adopted to test the MI of three-factor structure of this measure; three factors were symptom and emotional function, environmental stimuli, and activity limitation. CFA supported configural and metric invariance for Latino and African-American patients. This indicates that both the structure and meaning of the items of Mini AQLQ were same across these two groups of patients. Scalar and complete invariance were supported partially in the two groups. In particular, 11 out of 15 items of this measure showed scalar and complete invariance. Therefore, authors concluded that inclusion or exclusion may be needed for making unbiased comparison of HRQOL across different ethnic groups.

There are other studies that have also used CFA techniques to test for MI of HRQOL measures. For example, MI of the PedsQL^{TM} 4.0 Generic Core Scales for children with chronic health conditions and children without any chronic conditions (i.e., healthy) has been examined (Limbers et al., 2008). MI of the five-factor structure on the PedsQL^{TM} 4.0 Generic Core Scales was established for healthy and chronic health condition groups. The five factors are physical functioning, emotional functional, social functioning, school-related cognitive functioning, and missed school. Configural and metric invariance were tested for healthy and chronic health condition groups. Both

configural and metric invariance of the PedsQL™ 4.0 Generic Core Scales were achieved across healthy and chronic health condition groups. This implies that children in this study had a similar interpretation of the items on the PedsQL™ Generic Core Scales regardless of whether they were healthy or had a chronic health condition. Therefore, the authors concluded that when the PedsQL™ is used and differences in self-perceived HRQOL are found across chronic health condition and healthy groups, these are more likely to reflect real differences in HRQOL, rather than differences in the interpretation of the PedsQL™ items due to health status.

The longitudinal MI of the PedsQL™ 4.0 Generic Core Scales was examined over a one-year period for children five to 17 years of age. A total of 2,887 children from a statewide evaluation of the California State Children's Health Insurance Program (SCHIP; Varni et al., 2008) were included in the study. Multigroup CFA was used to study longitudinal MI of the PedsQL™ 4.0 Generic Core Scales. A five-factor structure was considered in the study, with the factors being physical functioning, emotional functioning, social functioning, school functioning, and missed school. Longitudinal MI was established by testing a set of MI hypotheses including invariance of covariance matrices, configural invariance, metric invariance, scalar invariance, and complete invariance. To assess the strictest form of MI, the model was further restricted by imposing the equality constraints of latent variances, latent covariances, and latent means sequentially across measurement occasions. An equivalent factor structure on the PedsQL™ 4.0 Generic Core Scales over time was supported by the study findings. The researchers concluded that children in this study interpreted items on the PedsQL™ 4.0

Generic Core Scales in a similar manner over time. This study as well as all other studies discussed in this section adopted ML as their methods of model **e**stimation.

2.4 Structural Equation Modeling for Complex Survey Design

Large health surveys often employ a complex design. For example, survey participants may be selected using methods such as clustering, stratification, unequal probability of selection, and post-stratification (Longford, 1995) instead of SRS. Multistage survey designs may also be used to select survey participants. Data collected using multistage survey designs often include information about clustering and stratification as well as unequal selection probabilities at all levels of sampling. The effects of clustering, stratification, and unequal probability of selection on the analysis of complex survey data are discussed below.

An assumption in conventional SEM is that observations are independent and identically distributed. When a complex survey design is adopted that employs a multistage survey method instead of SRS, observations exhibit some degree of dependence. Conventional estimation methods assume that the correlation of the errors across individuals is zero. When clustered data are used with a conventional estimation method, standard errors may be underestimated which subsequently result in inflated Type I error rates (Kish & Frankel, 1974). In the context of SEM, there may be an effect on the χ^2 statistic when clustered data are analyzed using conventional estimation methods (Muthén & Satorra 1995), which may lead to an improper rejection of the proposed model. In a simulation study (Muthén & Satorra, 1995) with clustered data, conventional ML and mean-corrected ML (Satorra & Bentler, 1988) estimation methods were compared using the χ^2 statistic for a factor model. The mean-corrected ML takes the

survey design into account on the analysis whereas conventional ML did not. This study demonstrated the superiority of the mean-corrected method over the conventional method for estimating the χ^2 statistic. That is, the values of χ^2 statistic were larger for conventional ML than for mean- corrected ML estimation procedure.

Stratification is another design issue for modeling data from a complex survey design. Stratification helps to obtain more precise or efficient estimates of population parameters. If stratification is the part of survey design employed for data collection and conventional ML method is used for model estimation, then unbiased parameters estimates may not be obtained (Kalton, 1983b).

Often in large-scale surveys unequal probability of selection of observations is part of the complex survey design. In a multistage survey design, the probability proportionate to size (PPS) method is commonly used. In this method, higher probability is assigned to select the larger clusters. This method is useful when a fixed number of observations are selected from each cluster. When unequal selection probability is ignored to select sample observations, parameter estimates may be biased if variables are correlated with the probability of selection. Previous SEM research has found biased parameter estimates when unequal selection probability was ignored (Asparouhov, 2005; Kaplan & Ferguson, 1999). Unequal selection probabilities are useful to calculate sample weights. Unequal selection probabilities can be included in the analysis by incorporating sample weights. A sample weight is the inverse of the selection probability.

Let us consider a particular complex survey design as follows: a two-stage survey design is considered in which clusters are selected in stage one and then, in stage two,

observations are selected from the clusters with equal or unequal probabilities. The sample weight for the *jth* cluster (*j*=1, 2, …, *J*) is

$$w_j = \frac{1}{p_j}, \tag{5}$$

where p_j is the probability that cluster *j* is included in the sample. The sample weights for observations within clusters are

$$w_{ji} = \frac{1}{p_{i/j}}, \tag{6}$$

where $p_{i/j}$ is the probability that observation *i* is selected in cluster *j*, given that cluster *j* is selected. The multiplication of the weights w_j and w_{ji} produces the final weight for each observation.

The issue of including sample weights in SEM analysis was addressed using a simulation study (Kaplan & Ferguson, 1999). In this study, observations were selected from two strata with unequal selection probabilities but within each stratum the selection probability was equal for all observations. Therefore, there were only two sample weights, w_1 and w_2, in the analysis. Sample weights were utilized in the calculation of weighted covariance matrices. Then both the weighted and unweighted covariance matrices were analyzed using the ML estimation procedure to investigate the effects of sample weights on goodness-of-fit indices such as RMSEA and the χ^2 statistic. The results indicated that the values of the χ^2 statistic obtained from the procedures of using normalized weights and ignoring sample weights were close in the analysis.

Sample weights can not be used in ML estimation (Kaplan & Ferguson 1999; Muthén & Muthén, 2007; Stapleton, 2008). Therefore, the choice of estimation methods

for SEM analysis depends on whether sample weights are used or not. In SEM, PML allows the use of not only sample weights but also characteristics of the survey design in the analysis simultaneously. Skinner (1989) introduced the PML method that can be used for complex survey data including stratification, cluster, and unequal probability of selection. In fact this method is applicable to a more general survey design which includes stratified multistage sampling with unequal probability of selection at all stages of the survey.

To analyze complex survey data pseudo log-likelihood values are used to perform the LRT. When data are obtained from a complex survey design the LRT statistic is defined as

$$\chi^2 = 2\,(L^*_C - L^*_{UC}), \tag{7}$$

where L^*_C and L^*_{UC} are pseudo log-likelihood values for the constrained and unconstrained models, respectively (Asparouhov & Muthén, 2005). The distribution of this test statistic depends on the survey design, including the sample weights, the stratification, and the cluster sampling. The LRT statistic has approximately a χ^2 distribution with the degrees of freedom equal to the difference in the number of parameters between constrained and unconstrained models (Asparouhov & Muthén, 2005). This adjustment was done similarly to the adjustments of the Satorra-Bentler (2001) robust χ^2 tests. The adjusted LRT statistic, based on pseudo log likelihood values, is defined as

$$\chi^2_{PML} = 2c(L^*_C - L^*_{UC}), \tag{8}$$

where, the correction factor is,

$$c = \frac{d_C - d_{UC}}{Tr((l''_C)^{-1}\sigma^2(l'_C)) - Tr((l''_{UC})^{-1}\sigma^2(l'_{UC}))}, \quad (9)$$

and l'_C, l'_{UC} and l''_C, l''_{UC} are the first and second derivatives of the pseudo log-likelihood functions L^*_C and L^*_{UC}; d_C and d_{UC} are the number of parameters in the constrained and unconstrained models, respectively; σ^2 is variance; Tr stands for Trace which sums the elements on the main diagonal of a square matrix.

Similar to equation 4, the corresponding CFI for this adjusted method (equation 8) can be defined as

$$CFI_{PML} = \frac{\max[(\chi^2_{PMLC} - df_{PMLC}),\ 0]}{\max[(\chi^2_{PMLUC} - df_{PMLUC}), (\chi^2_{PMLC} - df_{PMLC})]}, \quad (10)$$

where χ^2_{PMLC} and χ^2_{PMLUC} are χ^2 statistics for constrained and unconstrained models, respectively and df_{PMLC} and df_{PMLUC} are the corresponding degrees of freedom. This adjustment to the LRT not only corrects for complex survey design but also for distributional misspecifications such as non-normality. Therefore, the LRT statistic (equation 8) allows researchers to conduct robust SEM analysis when observed variables are not normally distributed. A simulation study (Lei & Lomax, 2005) found that parameter estimates and the χ^2 statistic were sensitive to the non-normality of observed variables when conventional ML method was used.

In SEM literature, there are not many studies available that have investigated the properties of a complex survey design on estimating χ^2 and CFI statistics for different estimation methods. Some studies (Asparouhov & Muthén, 2005; Stapleton, 2006;

Stapleton, 2008) compared estimation methods using simulation study to justify the effects of complex survey design on calculation of the χ^2 statistic.

Asparouhov & Muthén (2005) demonstrated the effects of various complex survey designs on the estimation of χ^2 statistic (equation 8) for the PML method by including and ignoring survey design in the analysis. A simulation study was conducted using a single outcome variable in the model and two hypotheses were tested using the χ^2 statistic for PML method. The first hypothesis was that the means of the outcome variable in two groups were equal and the second hypothesis was that the variances of the outcome variable were unequal in two groups. Four different approaches were considered for computing the χ^2 statistic such as including both stratification and clustering, including stratification and ignoring clustering, including clustering and ignoring stratification, and ignoring both clustering and stratification. The results indicated that the sampling features in complex survey design can affect the distribution of the χ^2 statistic. In particular, the Type I error rates for the χ^2 statistic were approximately 5% when α=0.05 with stratification and clustering in the analysis and all other methods produced erroneous results. There were almost no Type I errors for the χ^2 statistic when cluster was included but stratification was ignored. The method of including stratification but ignoring clustering effects produced incorrect Type I error rates (52%) for the χ^2 statistic. As well, Type I error rates for the χ^2 statistic were high (38%) for the method of ignoring both clustering and stratification information. In terms of Type II errors, the error rates for the χ^2 statistic was 24% when the effects of both stratification and clustering were taken into account; however this rate converged to 0% as sample size was increased. The

highest rates (50%) of Type II error was observed when stratification was ignored. On the other hand, negligible Type II error rates were seen in the methods of ignoring only cluster effects (Type II error rate of 1%) and ignoring both stratification and cluster effects (Type II error rate of 2%).

Stapleton (2006) investigated the PML method for estimating the χ^2 statistic in the context of SEM analysis. She examined the effects of complex survey design by including and ignoring clusters and stratification information in the analysis. Six different survey designs were investigated in single-stage, two-stage and three-stage sampling procedures. When PML method ignoring survey design was used, the χ^2 statistic rejected the model too often compared to when PML method including survey design characteristics was used instead of SRS (Stapleton, 2006). With PML method ignoring survey design, Stapleton found that for a two-stage survey design, rejection rates were 0.50 to 0.60 whereas for a three-stage design they were 0.75, where the nominal level of significance was 0.05. This implies that the more complex the survey design, the greater the probability of making a Type I error about the null hypothesis of overall goodness of fit if survey design is ignored. Stapleton also investigated a design effect adjusted χ^2 obtained by dividing the conventional χ^2 by the average design effect, when clustering was a characteristic of the survey design and there was homogeneity within clusters. The design effect is defined as the ratio of the correct sampling variance of a statistic under the complex survey design to the sampling variance obtained under SRS (Kish, 1965). This method resulted in inappropriately low χ^2 values, that is, the χ^2 statistic was overcorrected using this method. It has been shown that the χ^2 test is fairly robust (i.e.,

Type I error rate is close to α) when PML estimation was employed with complex survey design instead of ignoring complex survey design (Stapleton, 2006).

An adjusted χ^2 statistic has been proposed to take survey design effects into account, obtained by dividing the χ^2 statistic obtained from conventional CFA by the average design effect for the estimates of the model parameters (Stapleton, 2008). Stapleton compared the design effect adjusted χ^2, which was estimated using re-sampling techniques (JRR, BRR, and bootstrap), with the χ^2 statistics for PML methods with and without taking complex survey design into account (i.e., the same methods for PML that was used in Stapleton, 2006). Type I error rates for the χ^2 test statistics in design effect adjusted and PML method with complex survey designs were near the nominal level of significance of 0.05, whereas this rates for the conventional test were near 0.75. However, these results are only for a selected three-stage complex survey design and high intraclass correlation (ICC).

2.5 Software for Structural Equation Modeling

Software available for conducting SEM analyses includes Amos 6.0 (Arbucle, 2005), SAS (SAS Institute Inc., 2009), EQS 6 (Bentler & Wu, 2002), LISREL 8.8 (Joreskog & Sorbom, 1996) and Mplus 5.1 (Muthén & Muthén, 2007). However, conducting SEM analyses for complex survey data is challenging because of lack of availability of appropriate analysis procedures in these software packages. Only LISREL and Mplus can implement PML estimation. PML estimation was implemented in Mplus beginning with version 3.11 (Asparouhov, 2004) and in LISREL with version 8.8 (Asparouhov & Muthén, 2006). The performance of the LRT statistic based on PML estimation in Mplus and LISREL was assessed using a simulation study (Asparouhov &

Muthén, 2006). The Type I error rates for the LRT statistic described by equation 8 indicated that Mplus performed better than LISREL software. Specifically, the Type I error rates for the LRT statistic implemented in Mplus were close to the nominal 5% value. On the other hand, the LRT statistic implemented in LISREL produced large error, that is, the Type I error were 0.65 to 0.67 when the nominal level of significance was 0.05. Therefore, Mplus is the recommended choice for conducting multi-group CFA when data are sampled from a complex survey design.

2.6 Summary of Literature Review

In this chapter, previous literature has been summarized. The main points are highlighted here: A measurement model establishes the relationship between observed and latent variables, and must be fit to one's data and evaluated using CFA techniques before testing MI of HRQOL measures. A range of statistical fit indices is recommended to assess the overall fit of the measurement model such as χ^2 test, TLI, RMSEA, and SRMR Different fit indices are used because each of them address different aspects of model misspecification. Many forms of MI can be tested using CFA techniques. In order to establish MI of HRQOL measures across different groups, two nested models, constrained and unconstrained, must be specified and identified. Measures used to compare HRQOL across different groups include SF-36, SIP, WHOQOL, PedsQL 4.0 Generic Core Scale and CES-D. CFA techniques have been employed to test MI of these HRQOL measures when data were collected using SRS. In complex survey designs, observations may be selected using methods such as clustering, stratification, unequal probability of selection, and post-stratification instead of SRS. These results in a lack of independence amongst the observations, which can affect the results of tests of MI based

on SEM. Moreover, sample weights are generated and included in the data set if unequal probability of selection is a part of complex survey design, and ignoring the sample weights has an impact on the estimation of model parameters and goodness-of-fit indices on SEM analysis.

In conventional CFA, it is assumed that observations are independent. ML methods are employed to estimate model parameters and test hypotheses. There are currently a number of methods available to undertake SEM analyses when data are collected using a complex survey methodology. These strategies include estimating model parameters using ML, design effect adjusted ML, PML, and computer-intensive re-sampling techniques (JRR, BRR and bootstrap methods). Researchers who conduct SEM analysis with complex survey data may often choose to ignore the survey design in their analysis and adopt ML estimation. But reviewing recent literatures that used the SEM techniques, it has been found that the ML method resulted in inflated rates of rejection of the LRT statistic in complex survey data. The design effect adjusted method was also applied in χ^2 test when data were collected using complex survey design. In this method, adjusted χ^2 was obtained by dividing the conventional χ^2 by the average design effect. However, this method resulted in inappropriately low χ^2 values, that is, χ^2 was overcorrected using this method. The performance of the PML estimation method has also been investigated for different complex survey designs. This method appears to provide robust χ^2 tests. The estimation method that is the most often suggested in SEM analysis with complex survey data is PML (Asparouhov & Muthén, 2005; Muthén & Satorra, 1995; Stapleton, 2006).

CHAPTER THREE: SIMULATION STUDY

A simulation study was conducted to investigate the performance of the LRT and CFI statistics for testing hypotheses about four forms of MI. The estimation methods, ML and PML, were compared to compute the test statistics. First, the conventional method was used to compute the LRT (equation 3) and CFI (equation 4) statistics, based on ML estimation. Second, a robust method was used to compute the LRT (equation 8) and CFI (equation 10) statistics, based on PML estimation. In this method, results were produced for two approaches using only sample weights as well as using both sample weights and clusters in the analysis. The approach that used only sample weight was called PML1 and the other approach that used both sample weights and clusters was called PML2. Therefore three methods were used to investigate the performance of the LRT and CFI statistics. However, when PML is mentioned that indicates both PML1 and PML2 methods.

In the simulation study, data were generated for a population with two groups. The population contained homogeneous clusters of different sizes and observations from two groups were included in each cluster. The samples were drawn using a two-stage design. First, clusters were selected using the PPS method and then observations from the selected clusters were selected using SRS.

The simulation parameters were: (1) magnitude of latent variable (i.e., factor) loadings, (2) intercepts of observed variables, (3) magnitude of correlation between latent variables, (4) standard deviations of observed variables, (5) size of the ICC, (6) cluster size, and (7) total sample size. These characteristics of the simulation are summarized in Table 1 and are described in greater detail in subsequent sections of this chapter.

3.1 Simulation Study Methods

3.1.1 Study Design

This simulation study adopted a survey design that has many similarities to the design of the Canadian Community Health Survey (CCHS) cycle 3.1, which is the focus of the numeric example presented in Chapter 4. Three sampling frames, area frame, list frame of telephone numbers, and random digit dialing frame, were used in the CCHS cycle 3.1 to select the sample of households or dwellings. In the area frame, a multistage stratified cluster design was used. In the first stage, homogeneous strata were created based on criteria such as geography, socio-economic status, and demography. Each stratum was comprised of dwellings or households. Within each stratum, dwellings are regrouped to create clusters. Clusters, or primary sampling units, were selected from each stratum using the PPS method, in which the probability of selecting a sample unit is proportional to the size of the population. Cluster sizes varied from 150 to 250 households. In the second stage, dwelling or households lists were prepared for each cluster and a systematic sampling design was implemented to select households from each cluster. Systematic sampling is a method of selecting sample units from a sampling frame according to a random starting point and a fixed, periodic interval. The product of the probabilities for each of the two stages of selections represents the overall probability of selection. The inverse of this probability is used as the initial weight. Several adjustments (e.g., sample increase, nonresponse, removal of out-of-scope dwellings), were made to create the final sample weights when area frame was used (Statistics Canada, 2006).

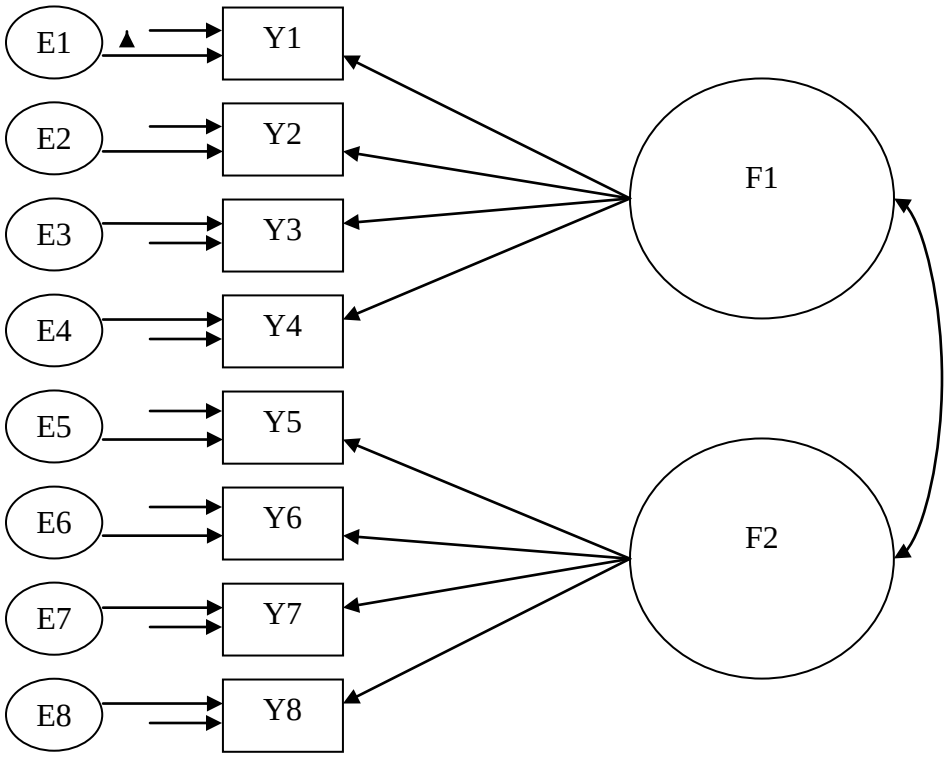
In the list frame of telephone numbers, a list frame stratum was created for each health region. Then the SRS design was applied to select the telephone numbers (i.e., household) from the stratum. Hence, the probability of selection corresponds to the ratio of the number of sampled households to the number of telephone numbers in the list frame stratum. The inverse of the probability of selection was used as the initial weight. A random digit dialing frame was employed only in four health regions. In this frame, random digit dialing stratum was formed as an aggregation of Area Code Prefixes, each of which contains valid banks of one hundred numbers. So, the probability of selection was the ratio of the number of sampled households to one hundred times the number of banks within the stratum. Therefore, the inverse of this probability was used as the initial weight. Similar to the area frame several adjustments were done to create one set of initial weights for list frame and random digit dialing frame. One final set of weights was created through integration and post-stratification from the two initial sets of weights.

In the simulation study, observations were drawn from the population using a two-stage survey design. In the first stage, clusters were selected using the PPS method and then in the second stage, observations were selected from each cluster using SRS. Both the simulation study and CCHS used a two-stage survey design, however there were some differences. The same PPS method was used to select clusters in both the simulation study and the CCHS. However, the simulation study used the SRS method to select the observations instead of a combination of systematic sampling and SRS, which was used in the CCHS design. The probability of selection in the simulation study was produced by multiplying the probabilities of selection in the two stages.

3.1.2 Measurement Model

This simulation study used a measurement model (Figure 1) with two latent variables, F1 and F2, eight observed variables represented by Y1 to Y8 and eight error variances represented by E1 to E8.

Figure 1: The measurement model for the simulation study



Note to Figure 1: The circles and rectangles indicate the latent variables (i.e., factors) and observed variables (i.e., indicators), respectively. The lines with single arrowheads from latent to observed variables represent the latent variable effects. The single-headed lines, described by E1 to E8, to the observed variables represent the error variances. The only single-headed lines to the observed variables represent the intercepts of observed variables. Finally, the double-headed curves indicate the correlation between the latent variables.

This model was chosen because it is the most common model that has been used to describe the measurement of the SF-36 (Stadnyk, Calder, & Rockwood, 1998), which was the HRQOL measure investigated in the numeric example in Chapter 4. In this model, it is assumed that the latent variables are correlated, which has been observed in previous research (Cheung & Rensvold, 2002; French & Finch, 2006). The latent variable, F1, was measured by the observed variables Y1 to Y4 and F2 was measured by the observed variables Y5 to Y8. A single arrowhead line (e.g., F1→ Y1) from a latent variable to an observed variable represents the direct effect, or factor loading on the observed variable. As well, a single arrowhead line (e.g., E1→Y1) from an error variance to an observed variable indicates the combined effect of all other sources of influence on the observed variable.

3.1.3 Data Generation

The population data were generated from a multivariate normal distribution. A data matrix, **A,** of eight standard normal variables was generated using the SAS RANNOR procedure (SAS Institute Inc., 2009). **A** was multiplied by the factor pattern matrix, **P,** to introduce correlation among the observed variables. In the resultant matrix, ***M,*** the variables were correlated. **P** is defined as the square root of the covariance matrix, which was calculated from the population parameters. Then the observations of the data matrix ***M*** were transformed using

$$\boldsymbol{M}_t = \boldsymbol{M} * \boldsymbol{SD} + \boldsymbol{\mu}, \tag{11}$$

where * indicates matrix multiplication, $\boldsymbol{\mu}$ and ***SD*** are diagonal matrices containing means and standard deviations, respectively. The simulation data were generated for two

groups. The same methodology was applied to generate the data in each group. Another data matrix, say $\mathbf{O}_{1,}$ of 8 standard normal variables with single observation was generated using the SAS generator RANNOR (SAS Institute Inc., 2009) and it was transformed to have zero mean and pre-specified standard deviations. This observation was then added to all observations in each cluster to induce correlations among the observations within each cluster.

A population size of 60,000 observations was generated. Previous studies have also generated similar population sizes (e.g., Flora & Corran, 2004). The population was comprised of 1200 clusters. Each cluster contained observations from both groups. Specifically, 40% of the observations in each cluster were from Group1 and 60% of the observations in each cluster were from Group2. The clusters contained different numbers of observations; specifically, the number of observations ranged from 25 to 75; the average size of each cluster was 50 observations.

3.1.4 Simulation Parameters

Six parameters were manipulated in the simulation study. These were: (a) magnitude of latent variable (i.e., factor) loadings, (b) intercepts of observed variables, (c) magnitude of correlation between latent variables, (d) magnitude of ICC, (e) cluster size and (f) total sample size (Table 1).

Previous research (Guadagnoli & Velicer, 1988) suggested that the size of the factor loadings was important in determining the stability of the factor analysis solution. Specifically, the authors found that solutions were stable when factor loadings were equal to 0.80 even with small sample size. Stable solutions were obtained when factor loadings were close to 0.60 but the sample size was greater than 150. However, for small factor

loadings, for example 0.40, sample sizes of 300 to 400 observations were needed to obtain a stable solution. Given the different combinations of factor loadings and sample sizes, three different patterns of the factor loadings were considered. In the first pattern, Pattern A, the factor loadings of the eight indicators were set equal to 0.70 in both groups. In the second pattern, Pattern B, the factor loadings of the observed variables were set equal to values of 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.9. The factor loadings were equal in the two groups for both of these patterns. In the third pattern, Pattern C, the factor loadings were not equal in the two groups. Specifically, the factor loadings in Group 1 were equal to 0.4, 0.5, 0.6, 0.6, 0.7, 0.7, 0.8, and 0.8 and in Group 2 they were equal to 0.4, 0.6, 0.6, 0.6, 0.7, 0.7, 0.7, and 0.8. In previous applied research about the SF-36, estimated factor loadings ranging from 0.4 to 0.9 were found for the observed variables (Keller et al. 1998; Peek, Ray, Patel, Stoebner-May, & Ottenbacher, 2004; Wu, Lee, & Yao, 2006). These studies also found a range of values (63 to 95) for the intercepts of the observed variables of the SF-36. In this study, the intercepts of the observed variables were equal and unequal in the two groups. In the equality condition, the value of the intercepts was 70 in both groups but in the inequality condition the values of intercepts were 75 and 80 in Group 1 and Group 2, respectively. Other simulation studies (Chen, 2007; Muthén & Satorra, 1995) were also conducted using similar values for the intercepts.

The correlation, ρ, between the two latent variables was set at values of 0.20, 0.50 and 0.80. Different values of ρ were also investigated in other studies (Cheung & Rensvold, 2002; Muthén & Satorra, 1995; Wu et al., 2006). The standard deviations, σ, of the observed variables were set to be equal in the two groups. The values of σ were

assumed to be 15 in the two groups. The values of σ of 8 indicators of the SF-36 were found in the range of 12 to 36 (Keller et al. 1998; Peek et al., 2004; Wu et al., 2006).

The ICC was also manipulated in this simulation study. The ICC can simply be expressed as

$$\text{ICC} = \frac{\sigma_{\text{B}}^2}{\sigma_{\text{B}}^2 + \sigma_{\text{W}}^2}, \tag{12}$$

where σ_{B}^2 is the between cluster variance and σ_{W}^2 is the within cluster variance. Different values of the ICC were assumed by changing the standard deviations in the generation of data matrix $\mathbf{O_1}$ in section 3.1.3. Three values of standard deviations, 0, 2 and 4, were used to generate this data matrix in order to produce different value of the ICC. The values of the ICC were equal to 0.00, 0.27 and 0.61 when the values of 0, 2 and 4 were used to generate $\mathbf{O}_1$ and σ equal 15 was used to generate the data matrix **M** in section 3.1.3. All eight indicators in the CFA model were assumed to have the same ICC. Different values of ICC were assumed to be investigated to cover a range of survey applications. The values of ICC were considered from low to high to take the effect of various levels of ICC on estimation methods. Previous research has investigated values for the ICC ranging from 0.05 to 0.50 (Muthén & Satorra, 1995; Stapleton, 2008).

The cluster size was varied, because the size of clusters is known to strongly affect parameter estimation (Muthén & Satorra, 1995). Different cluster sizes also reflect the different survey situations. In this study, four different sizes of clusters were considered, including clusters of size 25, 40, 60, and 75.

The number of clusters in the sample must be equal or more than the number of free parameters in the model to be estimated. In some models (for example, configural

invariance), the number of free parameters is 50, which is the maximum number of free parameters to be estimated in one model. Therefore, the number of clusters was assumed 54 and this number was held fixed throughout the simulation study.

Several rules of thumb have been proposed to decide about the sample size for a particular study, for example, 5 to 10 observations per parameters, 50 observations per variable, no less than 100 observations, and so on (Floyd & Widaman, 1995; Joreskog & Sorbom, 1989; Muthén & Muthén, 2002). Three different samples of sizes N = 400, 650, and 1000 were investigated in this study. These sample sizes are consistent with other simulation studies conducting MI tests (Cheung & Rensvold, 2002; French & Finch, 2006; Lubke & Muthen, 2004). Various sample sizes were investigated because the χ^2 test statistic is sensitive to sample size.

A total of 972 conditions were investigated in the simulation study: 3 patterns of factor loadings x 3 correlations between two factors x 3 levels of ICC x 4 cluster sizes x 3 total sample sizes x 3 estimation methods. For each set of conditions, 1000 samples of data were generated. For each sample, four MI tests were conducted and information about rejection of the LRT and CFI tests was recorded as well as the estimates of the factor loadings. In cases where there was no difference between the two groups in the simulation parameters, rejection of the LRT and CFI tests represents a Type I error (i.e., erroneously rejecting a true null hypothesis). In cases where there was a different between the two groups in the simulation parameters, rejection of the LRT and CFI tests represents a correct decision and enables an investigation of statistical power. To investigate the power of estimation methods both factor loadings and intercepts were

different between the two groups for all power conditions. For all conditions four cluster sizes were used.

Table 1: Simulation study parameters.

Parameter	Parameter values
λ	• Pattern A: 0.70 (all are equal in Group 1 & Group 2) • Pattern B: 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.90 (Group 1 & Group 2) • Pattern C: 0.40, 0.50, 0.60, 0.80, 0.50, 0.80, 0.50, 0.60 (Group 1) & 0.40, 0.60, 0.60, 0.70, 0.50, 0.80, 0.70, 0.60 (Group 2)
τ	• 70-70 (Group 1 & Group 2) • 75-80 (Group 1 & Group 2)
ρ	• 0.20 • 0.50 • 0.80
σ	• 15
ICC	• 0.00 • 0.27 • 0.61
Cluster size	• 25 • 40 • 60 • 75
# of clusters	• 54
N	• 400 • 650 • 1000

Note: λ = factor loading; τ = intercept; ρ = correlation between two factors; σ = standard deviation; ICC = intraclass correlation; *N* = sample size.

3.1.5 Sample Selection Design

From the generated population, samples were drawn in two stages. In stage 1, 54 clusters were selected using a PPS method from 1200 clusters and the sample weights for selected clusters were calculated. In stage 2, a SRS design was applied to select the three samples of sizes of 400, 650, and 1000 observations. Samples were drawn from the 54

clusters selected in stage 1. The sample weights for the selected observations were also calculated. The final sample weights for selected observations were obtained by multiplying the sample weights calculated in the two stages. Unequal number of observations was selected from the two groups in each of the three samples.

3.1.6 Simulation Software

SAS software version 9.2 (SAS Institute Inc., 2009) was used to generate the data for the population and to select the observations. The data generation program was written in SAS/IML. The SURVEYSELECT procedure was used to sample the observations from the population. Mplus software, version 5.1 was used to estimate the model parameters (Muthén & Muthén, 2007) and conduct the tests of statistical significance for each form of MI.

3.1.7 Measurement Invariance Tests

Four tests of MI were conducted for each simulated dataset: configural, metric, scalar and complete invariance. Let M1, M2, M3, & M4 represent the configural, metric, scalar, and complete invariance models, respectively. The χ^2 and CFI statistics were calculated according to equations 3 and 4, respectively, for the ML method, and according to equations 8 and 10, respectively, for the PML method. The estimation methods were evaluated based on the Type I error and power rates of χ^2 and CFI for configural invariance. The Type I error and power rates of $\Delta\chi^2$ (i.e., LRT) and ΔCFI were investigated for metric, scalar, and complete invariance.

First of all, configural invariance was established if the two-factor model (Figure 1) showed acceptable fit in each of the two groups. The statistical significance of

the χ^2 statistic was assessed at the value of α = 0.05. The χ^2 test rejected the model M1 when the *p*-value was less or equal to 0.05. The CFI statistic rejected the model M1 when its value was less than 0.95. To examine metric invariance, the difference in χ^2 values (i.e., $\Delta\chi^2$) was calculated between the models of M2 & M1 for the ML estimator. For the PML method $\Delta\chi^2$ was calculated for models of M2 & M1 based on the Satorra-Bentler χ^2 scaled difference test (Satorra & Bentler, 2001). These differences were then tested with the degrees of freedom (df) equal to the difference in df between the models M2 and M1. Non-significance of the difference in χ^2 implies the invariance of factor loadings across two groups, i.e., metric invariance is established. The difference in the CFI (i.e., ΔCFI) values of the models M2 and M1 was also obtained for ML and PML methods. If ΔCFI was greater than 0.01, it was suggested that the null hypothesis of the equality of the factor loadings in two groups should be rejected (Bentler, 1990).

Scalar invariance of the measurement model across two groups was examined by calculating $\Delta\chi^2$ and ΔCFI between the models M3 and M2 for the ML and PML methods. Again, Satorra-Bentler scaled χ^2 difference tests were applied to calculate $\Delta\chi^2$ when PML was employed. The difference tests were conducted as in metric invariance. Finally, the complete invariance test was carried out by calculating $\Delta\chi^2$ and ΔCFI for the models M4 and M3. Again, the difference tests were conducted as in scalar invariance. The $\Delta\chi^2$, ΔCFI, and model parameter estimates were calculated for 1000 replications.

Type I error and power rates of $\Delta\chi^2$ and ΔCFI were calculated for each of MI tests procedures and for all three estimation methods, ML, PML1 and PML2. Moreover,

the percent bias of the standardized factor loadings was calculated for each replication using the formula as

$$\text{Bias} = \frac{\hat{\theta} - \theta}{\theta} \times 100, \tag{13}$$

where θ is the population parameter and $\hat{\theta}$ is the estimate. The bias was same for the PML1 and PML2 because estimated factor loadings were same for these two methods. However, the standard errors of the factor loadings were not the same for the two estimation methods PML1 and PML2.

3.2 Results

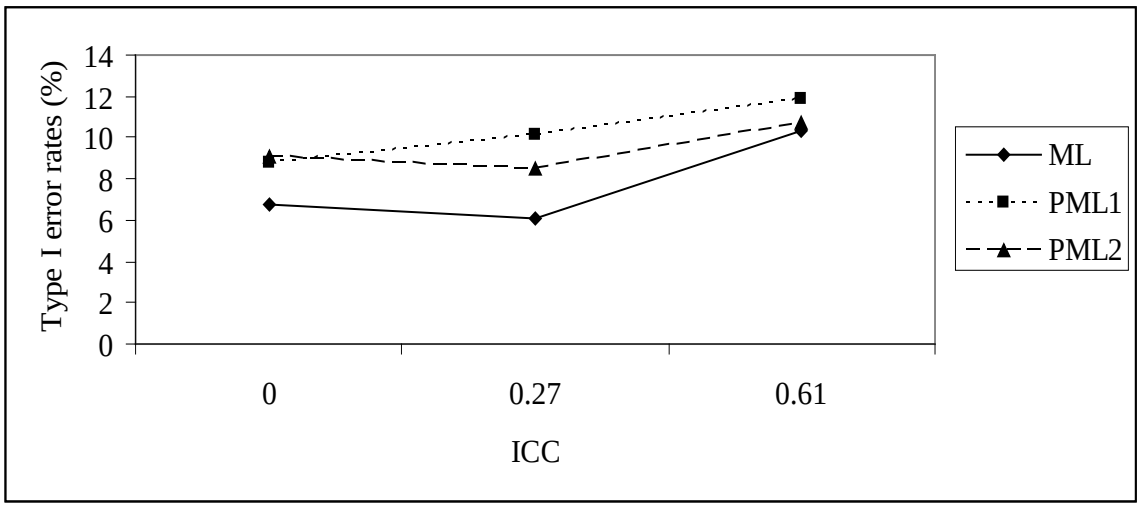
The simulation study compared the ML, PML1, and PML2 estimation methods by investigating the performance of the fit statistics LRT and CFI for MI tests using CFA techniques. The average Type I error and power rates of the LRT and ΔCFI for MI tests are provided by the patterns of factor loadings, sample size, ICC, and methods of estimation. In Appendix A, the Tables A-1 and A-2 described the average Type I error rates of the LRT and CFI statistics, respectively; while Tables A-3 and A-4 described the average power rates of these statistics, respectively. The average percentage biases of the standardized factor loadings are found in Tables A-5 to A-7 in Appendix A. Biases were presented for the ML and PML2 because the bias was the same for the PML1 and PML2 methods. Type I error rates, power rates, and biases were summarized separately in a number of figures in this chapter.

3.2.1 Type I Error Rates

For the factor loadings of Pattern A and Pattern B, the rejection rates of the LRT for MI tests are described in Table A-1 by sample size, ICC and method of estimation.

The Type I error rates of the LRT for configural invariance with factor loading Pattern A were higher for PML1 and PML2 than that of ML method when sample size and ICC were relatively small (Figure 2). On the other hand, ML and PML1 methods had higher Type I error rates than that of PML2 method in the case of high ICC and large sample size (Figures 3 and 4).

Figure 2: Type I error rates of the likelihood ratio test by intraclass correlation (ICC) and estimation method for $N = 400$ with Pattern A factor loadings.



An increasing trend was observed in the Type I error rates of the LRT for the methods of ML and PML1 with increasing sample sizes but this trend did not exist for the method of PML2. The Type I error rates of the LRT for configural invariance were similar for factor loading Patterns A and B.

Figure 3: Type I error rates of the likelihood ratio test by intraclass correlation (ICC) and estimation method for N = 650 with Pattern A factor loadings.

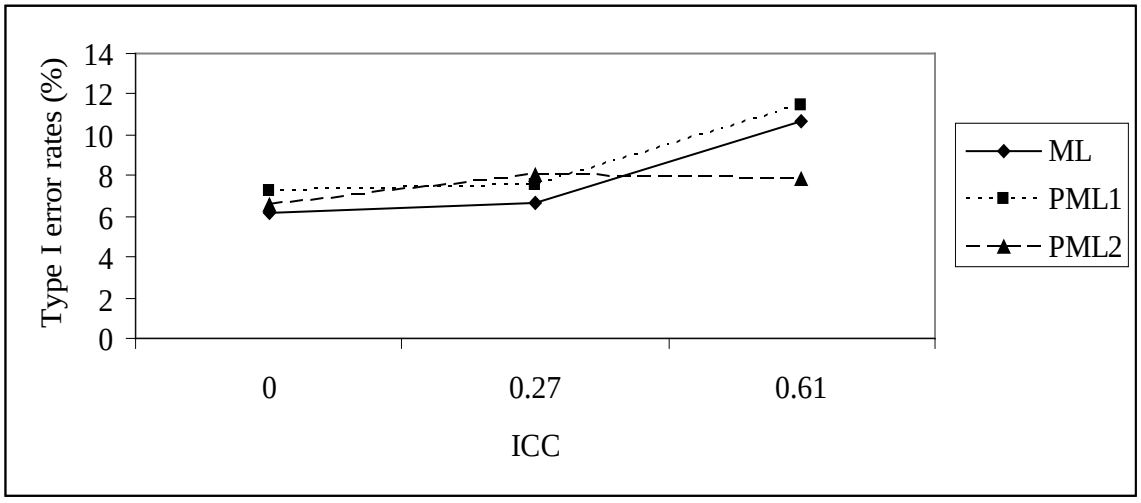
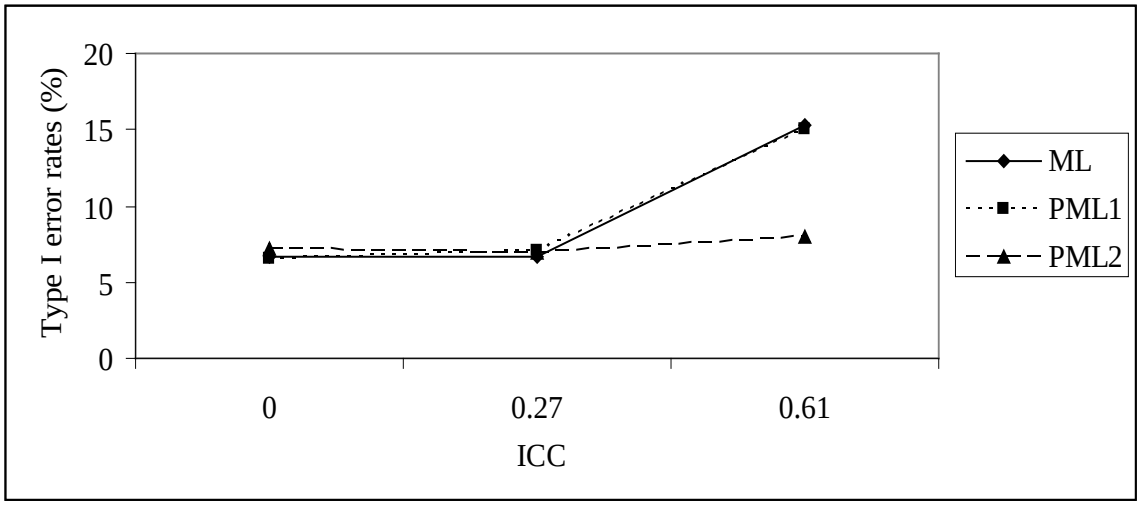


Figure 4: Type I error rates of the likelihood ratio test by intraclass correlation (ICC) and estimation method for N = 1000 with Pattern A factor loadings.



For metric invariance, no particular trend for the Type I error rates of LRT was found for different sample sizes and ICC with factor loading Patterns A and B (Table A-1 in Appendix A). The Type I error rates of the LRT for all estimation methods were close to the nominal 5% level, except for the condition of sample size of 650, ICC of 0.61 in Pattern A. For this particular condition the rejection rate (11.23%) in the PML1 method was almost double compared to the other methods. The results about the Type I error

rates of the LRT were similar to metric invariance for scalar invariance as well (Table A-1). However, the Type I error rates of LRT for PML1 method was approximately double that of the other methods when sample size was small (i.e., 400 observations) with factor loadings Pattern A in complete invariance. For the factor loadings Pattern B, the null hypothesis of complete invariance was rejected in the same way by all three methods of estimation.

The rejection rate was very low for the CFI for all three methods of estimation when configural invariance was tested (Table A-2 in Appendix A). This was true for both scenarios of factor loadings Pattern A and Pattern B. In terms of the Type I error rates of ΔCFI, an increasing tendency appeared when ICC was increased. This tendency was apparent when sample size was small because for high sample size the Type I error rates were close to zero. This fact was true except for the method of PML1 with factor loadings Pattern A and sample size of 400. Compared to the ML and PML2 methods, PML1 had rejected the null hypothesis of metric, scalar and complete invariance more frequently for factor loadings Pattern A. For ML and PML2 methods the Type I error rates of ΔCFI were less than the nominal 5% level for all forms of MI.

3.2.2 Power Rates

Table A-3, in Appendix A, describes the power rates of the LRT for MI tests by sample size, ICC and methods of estimation with factor loadings Pattern C (i.e., factor loadings were unequal within and between two groups). Moreover, intercepts were unequal in the two groups of observations.

In order to test the metric invariance, ML, PML1, and PML2 methods rejected the LRT in an increasing rate for increasing sample size but in a decreasing rate with increasing values of ICC (Figures 5, 6 and 7).

Figure 5: Power rates of the likelihood ratio test by intraclass correlation (ICC) and estimation method for metric invariance with $N = 400$.

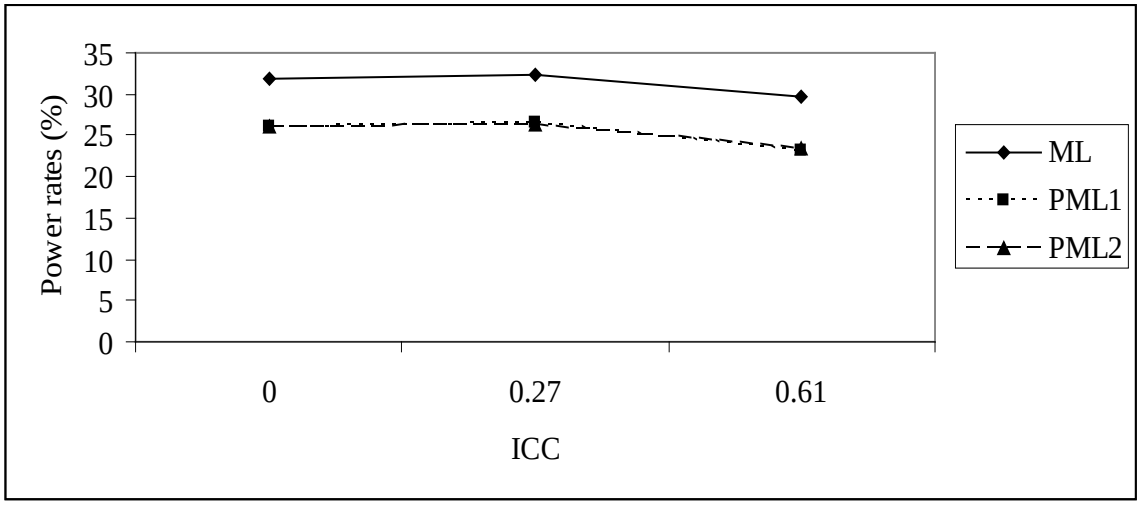
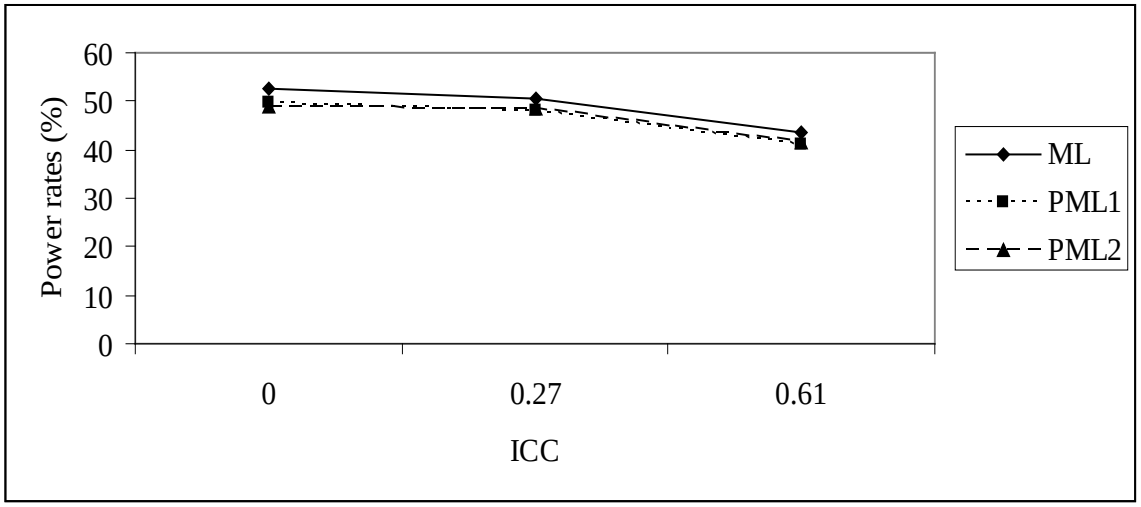


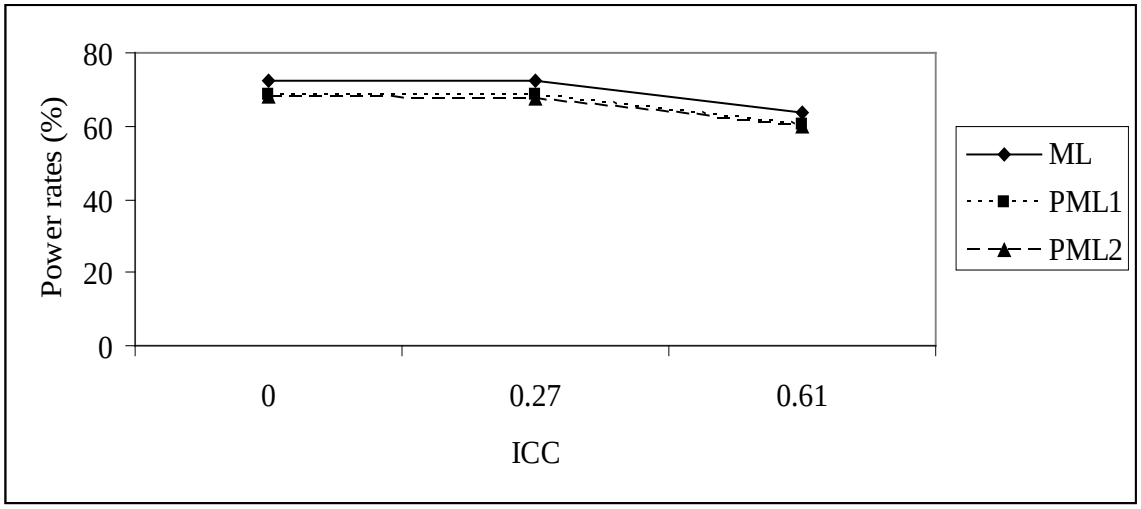
Figure 6: Power rates of the likelihood ratio test by intraclass correlation (ICC) and estimation method for metric invariance with $N = 650$.



The power rates of the LRT were higher for ML than for the PML1 and PML2 methods for factor loading Pattern C with unequal intercepts in the two groups. For

increased sample size, power rates were very similar for all three methods. Similar patterns of the power rates of LRT were found for testing scalar invariance and complete invariance. These patterns of power rates were also similar to metric invariance.

Figure 7: Power rates of the likelihood ratio test by intraclass correlation (ICC) and estimation method for metric invariance with $N = 1000$.



The power rates of ΔCFI were described in Table A-4 in Appendix A for factor loadings pattern C with unequal intercepts in the two groups. For testing metric invariance, an increasing pattern was found in the power rates of ΔCFI with increasing values of ICC for all methods of estimation (Figure 8). The rates for ΔCFI were higher for the PML1 and PML2 methods compared to the ML method of estimation. This same relationship was also found for scalar invariance. For complete invariance, the power rates ofΔCFI were higher for PML2 th an that of ML and PML1 (Figure 9). The differences were more evident when ICC was high. However, a decreasing pattern emerged in the power rates of ΔCFI with increasing values of ICC, which was opposite to

metric or scalar invariance. Moreover, there was a decreasing pattern in the power rates when sample sizes were increased for all forms of MI.

Figure 8: Power rates for the difference in comparative fit indices for nested models by intraclass correlation (ICC) and estimation method for metric invariance with N = 1000.

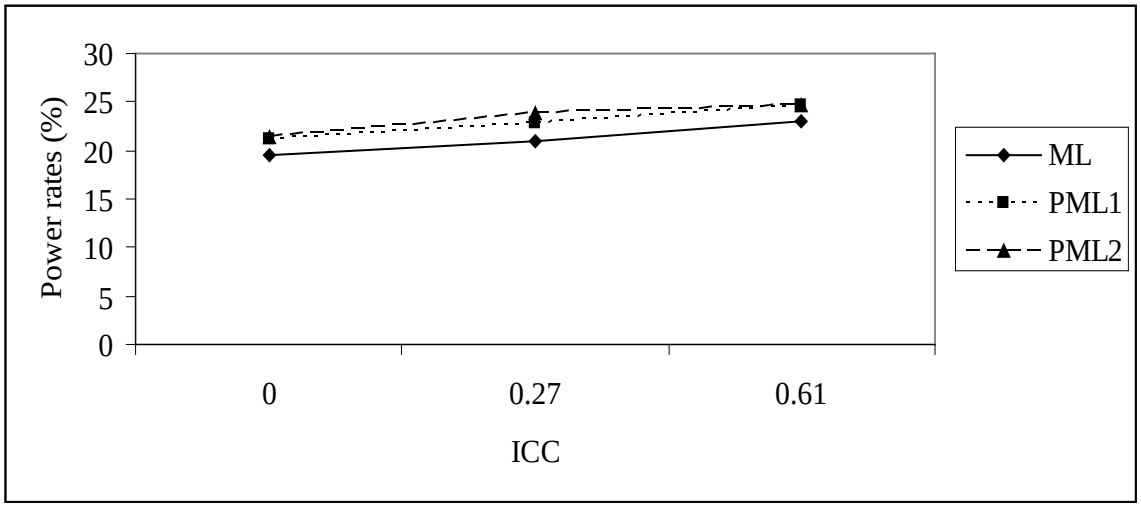
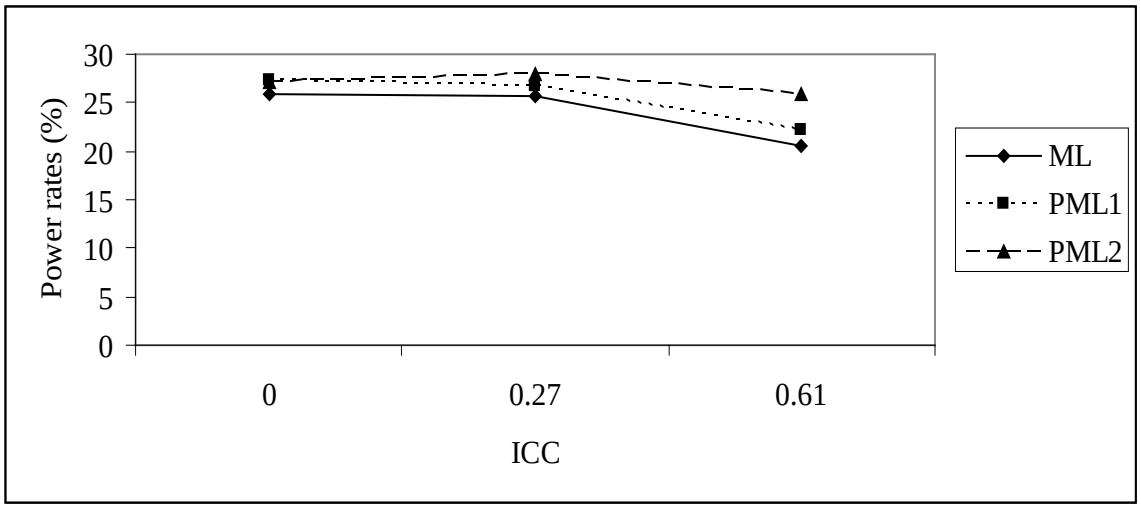


Figure 9: Power rates for the difference in comparative fit indices for nested models by intraclass correlation (ICC) and estimation method for complete invariance with N = 1000.

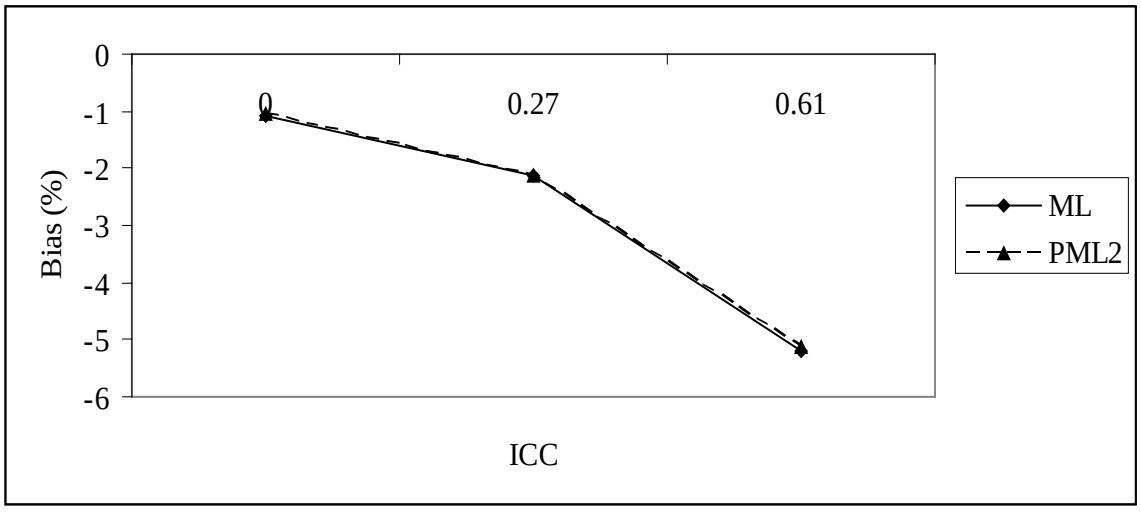


3.2.3 Bias of Standardized Factor Loadings

The average percentage biases of standardized factor loadings are described in Table A-5 for the Pattern A factor loadings and in Table A-6 for the Pattern B factor loadings in Appendix A. The results are summarized in Figures 10-12.

All the biases of factor loadings were negative when the value of ICC was 0.27 and 0.61 except for a few conditions. Moreover, biases increased negatively with increasing values of ICC for all forms of MI tests. But all biases were greater or equal to -5.2% for the factor loadings of Pattern A. It was not apparent that there was a particular trend in the biases of factor loadings for the methods of ML and PML2. As an example, Figure 10 illustrated the biases for these methods when Pattern A factor loadings were used for Group 2 in the configural invariance model. The biases for Group 1 were also similar for the two methods of estimation. For complete invariance, estimation methods also produced similar biases to configural invariance.

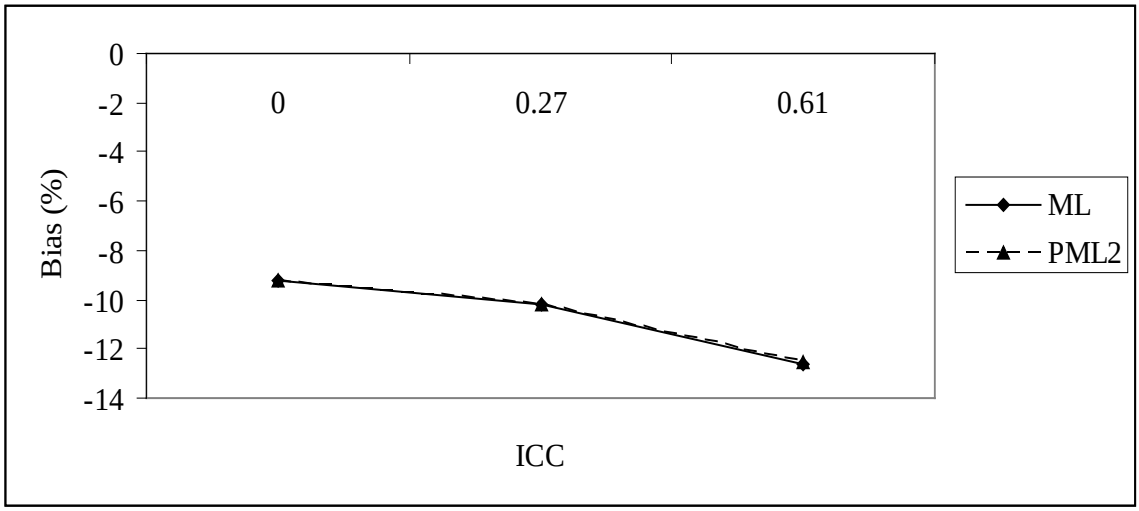
Figure 10: Bias (B_{31}) of standardized factor loadings (Pattern A) by intraclass correlation (ICC) and estimation method for configural invariance, Group 2 with $N = 400$.



For Pattern B factor loadings, biases tended to be high and positive in Group 1 for testing configural invariance with sample sizes of 400 and 650 and low factor loadings (from 0.40 to 0.60) (Table A-6). In particular, for a factor loading of 0.40, the bias was 492.7% when PML2 method was employed with sample size of 400. Comparatively, the ML method had smaller bias than the PML2 method for this specific sample size and low factor loadings. On the other hand, for a sample size of 650 and factor loading of 0.40, the bias was 80.0% when ML was used with an ICC of 0.27 whereas the bias was 78.1% when PML2 was used with an ICC of 0.61. The tendency for increased bias disappeared when sample size was increased to 1000 with the same factor loadings. Both methods, ML and PML2, produced similar results of bias for Group 2 in the configural invariance model. As well, for complete invariance, there was no particular trend of biases for these two methods of estimation.

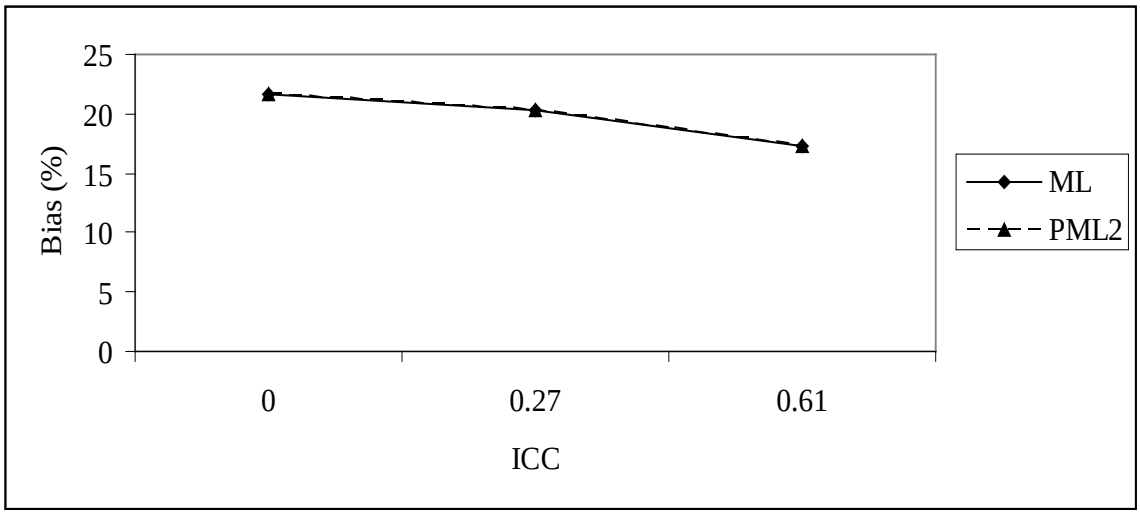
The biases of standardized factor loadings are provided in Table A-7 in Appendix A for Pattern C factor loadings. As in the other factor loading patterns, negative biases were observed and they increased with increasing values of the ICC. In contrast, positive biases decreased with increasing values of ICC. For configural invariance, the biases were similar for the two methods ML and PML2 regardless of sample size. This was also true for other patterns (Pattern A and Pattern B) of factor loadings because there was no impact of parameter constraints for testing configural invariance.

Figure 11: Bias (B_{41}) of standardized factor loadings (Pattern C) by intraclass correlation (ICC) and estimation method for complete invariance, Group 1 with N = 1000.



For testing complete invariance, biases were both positive and negative because some of the factor loadings were unequal in two groups but the estimated factor loadings were constrained to be equal. No substantial differences were evident between ML and PML2 estimation methods (Figures 11 and 12). Estimation methods ML and PML2 also produced similar biases for different sample sizes.

Figure 12: Bias (B_{72}) of standardized factor loadings (Pattern C) by intraclass correlation (ICC) and estimation method for complete invariance, Group 1 with N = 1000.



CHAPTER FOUR: CANADIAN COMMUNITY HEALTH SURVEY DATA ANALYSIS

This chapter demonstrates the application of the estimation methods and CFA techniques that were used in the simulation study in the previous chapter. However, although three estimation methods were investigated in the simulation study, only two of them were used to analyze the CCHS data. Specifically, in the CCHS data analysis, ML and PML1 were compared for testing the MI of the SF-36 (the SF-36 questionnaire is attached in Appendix C). The other method, PML2 which requires information about clustering or stratification variables, was not used because this information is not available in the CCHS cycle 3.1 data.

This chapter is organized as follows: data source and study sample, study measures and data analysis are described in the Methods section. The Results section includes a description of the characteristics of the sample and subsamples, the measurement model, and the measurement invariance test results.

4.1 Methods

4.1.1 Data Source and Study Sample

The CCHS is a national health survey that covers approximately 98% of the Canadian population in the provinces, including immigrants, aged 12 years or older. Statistics Canada conducts the CCHS to provide regular and timely cross-sectional estimates of health determinants, health status, and health system utilization for a total of 136 health regions in Canada, including 10 regions in Manitoba. CCHS data are generally collected at two-year intervals. The CCHS excludes individuals living on Indian Reserves

and on Crown Lands, institutional residents, full-time members of the Canadian Forces and the residents of some remote areas.

The CCHS cycle 3.1 data were collected from January 2005 to January 2006. In this cycle, there were 7,352 respondents 12 years of age or older from the province of Manitoba. All respondents aged 20 years or older were included in the analysis, to retain a focus on the adult population in whom chronic disease is more likely to be a significant issue.

For all of Canada, the sample size for CCHS cycle 3.1 was 132,947 respondents. Manitoba had a response rate of 83.3% in this cycle 3.1. The national response rate was 78.9%. The provincial response rates varied from 76.3% in Quebec to 85.7% in Newfoundland and Labrador (Statistics Canada, 2006).

4.1.2 Study Measures

The CCHS cycle 3.1 included the SF-36 as optional content, meaning that the data was only gathered in those provinces and territories who decided to participate in data collection. The SF-36 was chosen by the province of Manitoba as optional content. The other province that selected this measure in cycle 3.1 was Newfoundland and Labrador.

The SF-36 is a well-known measure of HRQOL (Alonso et al. 2004). The items in the SF-36 were drawn from the 245-item Medical Outcomes Study (Ware & Sherbourne, 1992). The Medical Outcomes Study is used to evaluate whether variations in patient outcomes are explained by differences in system of care, clinician specialty, and clinician's technical and interpersonal styles and to develop more practical tools for the routine monitoring of patient outcomes in medical practices. Medical outcomes may

include clinical end points; physical, social and role functioning in everyday living, patient's perceptions of their general health and well-being; and satisfaction with treatment.

The SF-36 is composed of 36 items that are summarized into multi-item scales (i.e., indicators); each scale is designed to measure one of eight generic health domains: physical functioning (PF), role limitations due to physical health problems (RP), bodily pain (BP), general health perceptions (GH), vitality (VT), social functioning (SF), role limitation due to emotional problems (RE), and mental health (MH). The PF scale includes ten items regarding the limitations of physical activities (vigorous and moderate) including lifting and carrying groceries, climbing stairs, bending, kneeling, walking moderate distances, and bathing or dressing. Two scales, RP and RE, were defined in the SF-36 to distinguish between role limitations due to physical and mental health problems. Four items regarding the problems with work or other regular daily activities as a result of physical health were included in RP; e.g., cut down the amount of time spent on work, accomplished less work than expected, difficulties performing work or other activities. RE contains three items about problems with work or regular activities because of depression or anxiety. BP has two items concerning the frequency of bodily pain or discomfort and measuring the extent of interference with normal activities due to pain. There are five items in the GH perception scale about personal health. VT, a scale of four items, measures energy level and fatigue. The SF scale has two items that assess the impact of physical health or emotional problems on social activities. The five-item MH scale consists of items from anxiety, depression, loss of behavioral or emotional control, and psychological wellbeing. The eight scales are summarized into two component

scores: physical health composed of PF, RP, BP, and GH, and mental health composed of VT, SF, RE, and MH. To facilitate comparisons across the SF-36 scales, each scale's raw scores are transformed to a 0 (worst measured health) to 100 (best measured health). The transformed scale scores are derived using the formula:

Transformed scale = [(Actual score - Lowest possible score)/Possible score range] x 100.

In the CCHS cycle 3.1, self-reported information on ethnic or cultural background was used to identify Aboriginal and non-Aboriginal respondents. The term Aboriginal may have different meanings, depending on the context. The respondents were first asked about their ancestor's ethnic or cultural background as follows: "*To which ethnic or cultural groups did your ancestors belong?*" There were two questions which were used to define Aboriginal population. The first question which was used from January to May 2005 was: "*People living in Canada come from many different cultural and racial backgrounds. Are you… Aboriginal peoples of North America (North American Indian, Métis, Inuit/Eskimo)?*" The second question which was used from June 2005 to January 2006 was: "*Are you an Aboriginal person, that is, North American Indian, Métis or Inuit?*" The respondents who answered yes to either of these two questions were considered to be Aboriginal. All other respondents are considered as non-Aboriginal population.

CCHS respondents were asked to provide information about several chronic health conditions. Chronic conditions include both physical and mental health conditions. A chronic condition was defined in the survey as any long-term condition that had lasted or was expected to last six months or more and that had been diagnosed by a health professional. The results of the chronic condition questions were used by CCHS

methodologists to define a single indicator of health status, called any chronic condition. Specifically, questions were asked about the following 32 chronic conditions: food allergies, other allergies, asthma, fibromyalgia, arthritis/rheumatism, high blood pressure, back problems, migraine headaches, chronic bronchitis, emphysema, chronic obstructive pulmonary disease, diabetes, epilepsy, heart disease, cancer, stomach or intestinal ulcers, stroke, urinary incontinence, bowel disorders, Alzheimer's Disease or any other dementia, cataracts, glaucoma, thyroid condition, chronic fatigue syndrome, multiple chemical sensitivities, schizophrenia, mood disorder, anxiety disorder, autism or any other developmental disorder, learning disability, eating disorder such as anorexia or bulimia, and any other long-term physical or mental health condition that has been diagnosed by a health professional. For each of the chronic conditions, there was a dichotomous response. Moreover, a single dichotomous chronic health condition variable was created which represents the presence or absence of any of the chronic health conditions mentioned above.

Information about age, sex, and marital status of respondents was also collected. Marital status was categorized as married, common-law, widowed, separated, divorced, and single. Respondents to the CCHS 3.1 were asked to provide their best estimate of the total income, before taxes and deductions, of all household members from all sources in the past 12 months. Respondents' total household income was assigned to one of six categories: less than $10,000, $10,000-$29,999, $30,000-$49,999, $50,000-$79,999, $80,000-$99,999 and $100,000+. Respondents were also classified as whether they lived in an urban or rural area. Urban areas are those which are continuously built-up having a

population concentration of 1,000 or more, as well the population density of 400 or more per square kilometre.

4.1.3 Data Analysis

The study sample was used to derive two subsamples based on the criteria of whether survey respondents self-reported having at least one chronic condition (ALOCC) or did not have any chronic conditions (NCC). The data were summarized by frequencies and percentages on the following variables: ethnicity, age group (20-44 years, 45-64 years, and 65+ years), sex, marital status, region, income, and chronic conditions. Moreover, in each of the subsamples, percentages and frequencies were separately generated for Aboriginal and non-Aboriginal respondents. Unweighted frequencies and percentages were produced. The distributional assumptions of the SF-36 indicators were assessed using univariate and multivariate measures of skewness and kurtosis. The correlations among the SF-36 indicators were also calculated for the entire sample as well as for the two subsamples of ALOCC and NCC, by ethnicity (see Appendix B).

As this study focused on the MI of the SF-36 in Aboriginal and non-Aboriginal groups for each of the subsamples of ALOCC and NCC, it was necessary to find a suitable measurement model of the SF-36 that fits well to Aboriginal and non-Aboriginal groups in both subsamples. In order to obtain a well-fitted measurement model, the standard two-factor (physical health and mental health) model (i.e., Figure 1) was initially fit to the SF-36 data for Aboriginal and non-Aboriginal groups in both ALOCC and NCC subsamples. Goodness-of-fit indices that were used to evaluate the measurement model included the χ^2 test, CFI, RMSEA, TLI, and SRMR. The model χ^2 statistic was assessed at the significance level of $\alpha=0.05$. The CFI and TLI values of

0.90 or more were used as indicators of good fit. Values of the RMSEA less than 0.10 were reflective of a well-fitting model. For the SRMR, the criterion of 0.08 was adopted as an indication of a well-fitting. These criterions for evaluating measurement model fit have been adopted in previous research. (Lix et al., 2009; Zimprich, Allemand, & Hornung, 2006).

Model modification indices were calculated and used as a guide to re-specify the measurement model to improve model fit. The correlation between two factors, physical and mental, was also estimated for the fitted measurement model. An estimated correlation greater than 0.85 has been used previously as an indicator of collinearity of the factors (Kline, 2005).

The measurement model for which a good fit was obtained for each of Aboriginal and non-Aboriginal groups in the subsamples of ALOCC and NCC was used as a final model to test the MI hypotheses. The four most common types of MI hypotheses (Vandenberg & Lance, 2000) were tested using CFA techniques: configural, metric, scalar, and complete invariance.

Configural invariance was established if the measurement model shows acceptable fit in each subsample of ALOCC and NCC. Configural invariance was assessed based on several goodness of fit indices including χ^2 test, RMSEA, CFI, TLI, and SRMR. Each of these indicators is sensitive to different aspects of model misspecification (Loehlin, 2004), which is why multiple indicators are recommended for evaluation of MI.

The LRT statistic was used to test the null hypothesis for each form of MI. Because of the sensitivity of the LRT to sample size, it is recommended to also use the

absolute difference in CFI values for two models to assess MI across groups (Cheung & Rensvold, 2002). An absolute difference in CFI values less than or equal 0.01 for two models is evidence that model parameters are likely to be equal across groups (Bentler, 1990).

If configural invariance was satisfied, then the hypothesis of metric invariance was tested next. To examine metric invariance, constrained (M_{MI}) and an unconstrained (M_{UC}) models were fit to the data. In the former model, factor loadings were assumed to be equal across Aboriginal and non-Aboriginal groups, while in the latter model they were freely estimated. The difference in χ^2 values (i.e., $\Delta\chi^2$) for M_{MI} and M_{UC} was used to test the invariance of factor loadings across the two groups. The $\Delta\chi^2$ was compared to a critical value from a χ^2 distribution with the degrees of freedom equal to the difference in degrees of freedom between the two nested models, M_{MI} and M_{UC}. Non-significance of $\Delta\chi^2$ statistic implies the invariance of factor loadings across two groups (i.e., complete metric invariance is established). The absolute difference in CFI values (i.e., ΔCFI) between M_{MI} and M_{UC} models was calculated and compared to the criterion of 0.01. If the value of ΔCFI was less than or equal to 0.01 then the null hypothesis of metric invariance was not rejected. If there was a conflict between the two statistics $\Delta\chi^2$ and ΔCFI, i.e., one rejects the null hypothesis and the other do not, then the decision was taken based on the ΔCFI statistic because the $\Delta\chi^2$ statistic is sensitive to sample size (Meredith & Teresi, 2006).

Scalar invariance was assessed only if metric invariance was established across the two groups. Scalar invariance of the measurement model was assessed by considering another constrained model (M_{SI}) in which factor loadings and intercepts were constrained

to be equal across two groups. The $\Delta\chi^2$ and ΔCFI statistics for models M_{MI} and M_{SI} were used in the same way as for the test of metric invariance to determine whether scalar invariance was established. If scalar invariance was established then complete invariance was evaluated. For this form of invariance, the last model (M_{CI}) was considered in which factor loadings, intercepts, and error variance were constrained to be equal across groups. Again, the $\Delta\chi^2$ and ΔCFI statistics for models M_{SI} and M_{CI} were examined to determine whether complete invariance of the measurement model was achieved.

Separate analyses were conducted for each of the subsamples, ALOCC and NCC, to establish the MI for Aboriginal and non-Aboriginal groups. This is because the presence of a chronic condition may be a confounding variable in the analysis.

The ML and PML1 methods were employed to estimate model parameters and therefore to test the hypotheses about MI. The PML2 method was not employed in this analysis due to the unavailability of information about clusters or strata used in the CCHS survey design.

The CCHS data set contained sample weights, which were developed by the survey methodologists. Sample weights were used in the analysis so that parameter estimates produced from the data were representative of the population from which sample was selected.

SAS software version 9.2 was used to conduct the descriptive analyses of the samples (SAS Institute Inc., 2009). All MI analyses were carried out using Mplus software version 5.1 (Muthén & Muthén, 2007). Mplus is appropriate for conducting multi-group CFA when groups are of unequal size and data are obtained from a complex survey design.

4.2 Results

4.2.1 Characteristics of Sample

There were 7,352 Manitoba respondents to the CCHS cycle 3.1. A total of 6,437 respondents were adults 20 years of age or older. The characteristics of the study sample are described in Table 2. Off-reserve Aboriginal respondents comprised 8.2% (n = 518)

Table 2: Characteristics of Manitoba adult respondents, Canadian Community Health Survey, cycle 3.1 (2005/2006).

		N	%
Ethnicity	Aboriginal	518	8.2
	Non-Aboriginal	5802	91.8
Sex	Male	2918	45.3
	Female	3519	54.7
Age (years)	20-44	2502	38.9
	45-64	2148	33.4
	65+	1787	27.8
Marital Status	Married/Common-law Partner	3693	57.5
	Single	1234	19.2
	Widowed/Separated/Divorced	1500	23.3
Region	Urban	4214	65.5
	Rural	2223	34.5
Chronic Conditions	Any chronic condition	4711	73.5
	Anxiety disorder	292	4.6
	Arthritis	1749	27.2
	Asthma	510	7.9
	Bowel disorder	295	4.6
	Cancer	127	2.0
	Chronic bronchitis	173	2.7
	Chronic obstructive pulmonary disease (COPD)	65	1.2
	Diabetes	462	7.2
	Heart disease	385	6.0
	High blood pressure	1468	22.9
	Mood disorder	395	6.2
	Stroke	125	1.9
Annual Income	<$10,000	188	3.5
	$10,000-29,999	1535	28.4
	$30,000-49,999	1291	23.9
	$50,000-79,999	1321	24.4
	$80,000-99,999	463	8.6
	$100,000+	610	11.3

of the adult respondents. The percentage of females (54.7%) was higher than that of males. The average age of the adult respondents was 51.9 years (standard deviation (SD)=18.7) with the highest percentage of respondents in the youngest age group (20-44 years) and the lowest percentage in the age group of 65+ years. More than half (57.5%) of the respondents were married or living common-law and less than one fifth (19.2%) was single. Almost two third (65.5%) of the CCHS 3.1 respondents lived in urban areas.

Overall, there were 4711 (73.2%) respondents who reported having at least one chronic condition, with arthritis (27.2%) and high blood pressure (22.9%) being the most frequently reported chronic diseases. Chronic obstructive pulmonary disease (COPD), stroke, cancer, and chronic bronchitis had the lowest prevalence in the sample.

The highest percentage (28.4%) of respondents was in the income range of $10,000 to $29,999, whereas the lowest percentages (3.5%) of respondents were in the lowest income range of less than $10,000 (Table 2). More than ten percent of respondents were in the highest income range ($100,000+).

Descriptive information about the SF-36 indicators is contained in Table 3. The highest mean score was observed for the indicator of RE while the indicator of VT had the lowest mean score. The univariate measures of skewness were found to be negative whereas measures of kurtosis were found to be positive for all of the indicators. The skewness and kurtosis for the indicators of RP, BP, GH, and VT were close to zero which implied that indicators were approximately normally distributed. The skewness or kurtosis for the other indicators indicated that there was a departure from the assumption of normal distribution of the indicators.

Table 3: Means, standard deviations, skewness and kurtosis of indicators of the SF-36 for all Manitoba adult respondents, Canadian Community Health Survey, cycle 3.1 (2005/2006).

	PF	RP	BP	GH	VT	SF	RE	MH
$\overline{X}$	84.72	81.46	76.83	73.13	64.12	90.49	92.83	83.59
σ	23.57	34.19	24.33	19.85	21.18	19.16	22.53	14.52
γ_1	-1.89	-1.59	-0.92	-0.94	-0.74	-2.47	-3.24	-1.72
γ_2	2.80	0.88	0.17	0.60	0.09	6.11	9.41	3.68

Note: PF = physical functioning; RP = role limitations due to physical health problem; BP = bodily pain; GH = general health perception; VT = vitality; SF = social functioning; RE = role limitations due to emotional problems; MH = mental health; $\overline{X}$ = mean; σ = standard deviation; γ_1 = skewness; γ_2 = kurtosis.

4.2.2 Characteristics of Subsamples

In the Manitoba adult population, 73.2% respondents reported having at least one chronic condition. Approximately 8.1% of respondents reported being of Aboriginal ethnicity in both the ALOCC (*n* = 376 respondents) and NCC (*n* = 138 respondents) subsamples.

Table 4 described the unweighted percentages and frequencies of ALOCC respondents by ethnicity; the corresponding results for NCC respondents are reported in Table 5. In the NCC subsample (Table 5), there was a higher percentage of non-Aboriginal than Aboriginal males. The mean ages of Aboriginal and non-Aboriginal groups in the ALOCC subsample were 43.9 (SD = 15.8) and 56.1 (SD = 18.3) years, respectively. The corresponding mean ages in the NCC subsample were 34.1 (SD = 12.2) and 43.1 (SD = 15.8) years, respectively. The proportion of Aboriginal respondents in the ALOCC subsample was almost double (53.5%) that of non-Aboriginal respondents (29.4%) in the lowest age group (20-44 years). On the other hand, in the highest age group (65+ years), the proportion of non-Aboriginal respondents in the ALOCC subsample was more than three times (35.4%) that of Aboriginal respondents (10.9%)

(Table 4). In terms of marital status, in the ALOCC subsample, about half of Aboriginal respondents were married or had a common-law partner, which was similar to the percentage for non-Aboriginal respondents.

Table 4: Characteristics of Manitoba adult respondents with at least one chronic condition by ethnicity, Canadian Community Health Survey, cycle 3.1 (2005/2006).

		Aboriginal		Non-Aboriginal	
		N	(%)	*N*	(%)
Total		376	8.1	4251	91.9
Sex	Male	144	38.3	1778	41.8
	Female	232	61.7	2473	58.2
Age (years)	20-44	201	53.5	1252	29.5
	45-64	134	35.6	1492	35.1
	65+	41	10.9	1507	35.5
Marital Status	Married/Common-law partner	193	51.5	2438	57.4
	Single	108	28.8	661	15.6
	Widowed/Separated /Divorced	74	19.7	1148	27.0
Region	Urban	246	65.4	2726	64.1
	Rural	130	34.6	1525	35.9
Chronic Condition	Anxiety disorder	42	11.2	248	5.9
	Arthritis	118	31.6	1587	37.4
	Asthma	49	13.0	447	10.5
	Bowel disorder	23	6.1	267	6.3
	Cancer	-	-	116	2.7
	Chronic bronchitis	26	6.9	140	3.3
	Chronic obstructive pulmonary disease (COPD)	–	–	60	1.6
	Diabetes	49	13.0	405	9.5
	Heart disease	25	6.7	346	8.2
	High blood pressure	92	24.5	1343	31.6
	Mood disorder	47	12.5	342	8.1
	Stroke	-	-	114	2.7
Annual Income	<$10,000	27	8.4	129	3.6
	$10,000-29,999	120	37.2	1108	30.6
	$30,000-49,999	67	20.7	877	24.2
	$50,000-79,999	66	20.4	825	22.8
	$80,000-99,999	21	6.5	305	8.4
	$100,000+	22	6.8	379	10.5

Notes: Data in columns with – are suppressed due to small cell size.

In the NCC subsample, there were differences between Aboriginal and non-Aboriginal respondents in the proportion of who were married or common-law or were single. The proportion of Aboriginal respondents (73.9%) with NCC was higher compare to non-Aboriginal respondents (68.7%) in urban area (Table 5) but these proportions were similar for the respondents with ALOCC (Table 4).

Anxiety disorders, asthma, chronic bronchitis, diabetes, and mood disorders were higher and arthritis, heart disease, and high blood pressure were lower in the Aboriginal than in the non-Aboriginal group in the ALOCC subsample (Table 4).

Table 5: Characteristics of Manitoba adult respondents with no chronic condition by ethnicity, Canadian Community Health Survey, cycle 3.1 (2005/2006).

		Aboriginal		Non-Aboriginal	
		N	(%)	*N*	(%)
Total		138	8.2	1535	91.8
Sex	Male	53	38.4	855	55.7
	Female	85	61.6	680	44.3
Age (years)	20-44	116	84.1	897	58.4
	45+	22	15.9	638	41.6
Marital Status	Married/Common-law partner	77	55.8	925	60.4
	Single	61†	44.2†	389	25.4
	Widowed/Separated /Divorced			217	14.2
Region	Urban	102	73.9	1054	68.7
	Rural	36	26.1	481	31.3
Annual Income	<$30,000	54	44.3	279	21.1
	$30,000-49,999	24	19.7	316	23.9
	$50,000-79,999	24	19.7	401	30.4
	$80,000+	20	16.4	324	24.5

Notes: † represents the combined percentages and frequencies for single, widowed, separated and divorced. These categories were combined due to small sample sizes.

Based on the total household income, there were proportionately more Aboriginal respondents than non-Aboriginal respondents in lower income categories in both the

ALOCC and NCC subsamples. However, in the highest income groups, there was a higher proportion of non-Aboriginal than Aboriginal respondents in both the NCC and ALOCC subsamples.

Table 6: Means, standard deviations, skewness and kurtosis of indicators of the SF-36 for Manitoba adult respondents by chronic disease status and ethnicity, Canadian Community Health Survey, cycle 3.1 (2005/2006).

	PF	RP	BP	GH	VT	SF	RE	MH
ALOCC								
Aboriginal								
$\overline{X}$	81.96	75.62	71.62	66.24	60.54	86.96	85.79	78.67
σ	23.68	37.40	26.74	21.35	21.69	21.15	30.91	17.72
γ_1	-1.55	-1.16	-0.64	-0.68	-0.53	-1.84	-2.00	-1.20
γ_2	1.67	-0.34	-0.41	0.05	-0.31	2.92	2.50	1.04
Non-Aboriginal								
$\overline{X}$	80.26	76.96	72.53	69.60	61.45	88.43	91.66	82.72
σ	25.75	36.91	25.04	20.57	21.85	20.99	24.20	15.23
γ_1	-1.51	-1.25	-0.70	-0.79	-.064	-2.16	-2.93	-1.66
γ_2	1.30	-0.12	-0.21	0.17	-0.15	4.35	7.40	3.27
NCC								
Aboriginal								
$\overline{X}$	95.53	94.59	89.13	78.52	69.50	95.02	95.78	83.91
σ	11.07	19.48	17.68	13.07	17.89	11.28	16.57	11.38
γ_1	-4.27	-4.06	-1.90	-0.48	-0.67	-2.97	-4.38	-0.98
γ_2	23.53	16.20	3.94	-.026	0.10	10.25	19.71	1.02
Non-Aboriginal								
$\overline{X}$	96.50	93.72	88.54	83.81	71.79	96.50	97.48	87.24
σ	10.00	20.99	16.84	12.72	17.12	10.96	12.87	10.55
γ_1	-5.35	-3.59	-1.68	-0.83	-0.92	-4.42	-5.84	-1.69
γ_2	35.71	12.05	3.24	0.59	1.01	23.24	35.99	4.71

Note: PF = physical functioning; RP = role limitations due to physical health problem; BP = bodily pain; GH = general health perception; VT = vitality; SF = social functioning; RE = role limitations due emotional problems; MH = mental health; ALOCC = at least one chronic condition; NCC = no chronic condition; $\overline{X}$ = mean; σ = standard deviation; γ_1 = skewness; γ_2 = kurtosis.

The mean scores of the eight SF-36 indicators of SF-36 were higher for non-Aboriginal than Aboriginal respondents in the ALOCC subsample (Table 6). In the NCC

subsample, the mean scores of the eight SF-36 indicators, with the exception of RP and BP, were higher for non-Aboriginal respondents than their Aboriginal counterparts. Comparing Aboriginal respondents who had and did not have a chronic condition, the mean scores were higher for the latter. A similar pattern of results was observed for non-Aboriginal respondents. Although all the measures of skewness were negative, the values were close to zero for the indicators of BP, GH, and VT in the ALOCC subsample, and for the indicators of GH and VT in the NCC subsample for each of the two groups (Table 6). The univariate measures of kurtosis for almost all of these indicators were positive (Table 6). The values of kurtosis for the indicators of RP, BP, GH and VT were close to zero in the two groups of Aboriginal and non-Aboriginal respondents in the ALOCC subsample. Large values of kurtosis were observed for the indicators of PF, RP, SF, and RE in the NCC subsample.

4.2.3 Measurement Model

Prior to testing different forms of MI, a measurement model was fitted to the data for each of the Aboriginal and non-Aboriginal groups in the ALOCC and NCC subsamples. In the standard SF-36 model (similar to Figure 1), four indicators have high loadings on the physical health factor and the other four indicators have high loadings on the mental health factor (Ware, Snow, Kosinski, & Gandek, 1993). The fit of the measurement model for each of the Aboriginal and non-Aboriginal groups with and without additional specifications are summarized in Table 7. Fit was evaluated using both ML and PML1 estimation methods.

The χ^2 test results suggest that this measurement model did not fit the data for the Aboriginal and non-Aboriginal groups in the ALOCC and NCC subsamples when the

Table 7: Fit criteria for measurement models of the SF-36, Canadian Community Health Survey, cycle 3.1 (2005/2006).

Sub-sample		RMSEA (90% CI)	TLI	SRMR	χ^2	p-value	df	CFI	*r*
Two-factor Confirmatory Factor Analysis Model									
				PML1					
ALOCC	AB	0.09 (0.07-0.11)	0.83	0.06	75.77	0.00	19	0.89	0.81
	NAB	0.07 (0.07-0.08)	0.84	0.06	431.10	0.00	19	0.89	0.82
NCC	AB	0.13 (0.10-0.17)	0.71	0.09	63.96	0.00	19	0.80	0.46
	NAB	0.08 (0.07-0.09)	0.50	0.07	196.78	0.00	19	0.66	0.66
				ML					
ALOCC	AB	0.12 (0.10-0.14)	0.89	0.05	120.66	0.00	19	0.92	0.81
	NAB	0.14 (0.14-0.15)	0.83	0.07	1567.25	0.00	19	0.88	0.83
NCC	AB	0.10 (0.06-0.14)	0.83	0.08	43.43	0.00	19	0.88	0.38
	NAB	0.14 (0.13-0.15)	0.62	0.08	580.25	0.00	19	0.74	0.60
Two-factor Confirmatory Factor Analysis Model with Correlated Errors									
				PML1					
ALOCC	AB	0.07 (0.04-0.10)	0.90	0.04	34.94	0.00	13	0.96	0.94
	NAB	0.05 (0.04-0.05)	0.93	0.03	127.30	0.00	13	0.97	0.91
NCC	AB	0.09 (0.04-0.14)	0.88	0.05	25.93	0.02	13	0.94	0.83
	NAB	0.04 (0.03-0.06)	0.86	0.04	46.19	0.00	13	0.94	0.97
				ML					
ALOCC	AB	0.08 (0.06-0.11)	0.95	0.03	45.74	0.00	13	0.98	0.91
	NAB	0.09 (0.08-0.10)	0.93	0.03	456.53	0.00	13	0.97	0.92
NCC	AB	0.05 (0.00-0.11)	0.96	0.05	17.40	0.18	13	0.98	0.77
	NAB	0.10 (0.08-0.11)	0.83	0.05	188.10	0.00	13	0.92	0.96
One-factor Confirmatory Factor Analysis Model									
				PML1					
ALOCC	AB	0.10 (0.08-0.12)	0.78	0.07	97.11	0.00	20	0.84	
	NAB	0.08 (0.08-0.09)	0.78	0.07	598.64	0.00	20	0.84	
NCC	AB	0.29 (0.26-0.32)	-0.39	0.12	245.98	0.00	20	0.01	
	NAB	0.08 (0.07-0.09)	0.53	0.08	197.75	0.00	20	0.66	
				ML					
ALOCC	AB	0.16 (0.14-0.18)	0.81	0.06	195.51	0.00	20	0.87	
	NAB	0.16 (0.15-0.16)	0.78	0.07	2083.45	0.00	20	0.84	
NCC	AB	0.16 (0.13-0.20)	0.53	0.11	90.06	0.00	20	0.67	
	NAB	0.15 (0.15-0.16)	0.54	0.08	729.47	0.00	20	0.67	
One-factor Confirmatory Factor Analysis Model with Correlated Errors									
				PML1					
ALOCC	AB	0.06 (0.04-0.09)	0.92	0.04	34.79	0.00	14	0.96	
	NAB	0.05 (0.04-0.06)	0.92	0.04	157.03	0.00	14	0.96	
NCC	AB	0.08 (0.03-0.13)	0.89	0.06	26.24	0.02	14	0.95	
	NAB	0.04 (0.03-0.05)	0.88	0.04	45.75	0.00	14	0.94	
				ML					
ALOCC	AB	0.09 (0.06-0.11)	0.94	0.03	54.01	0.00	14	0.97	

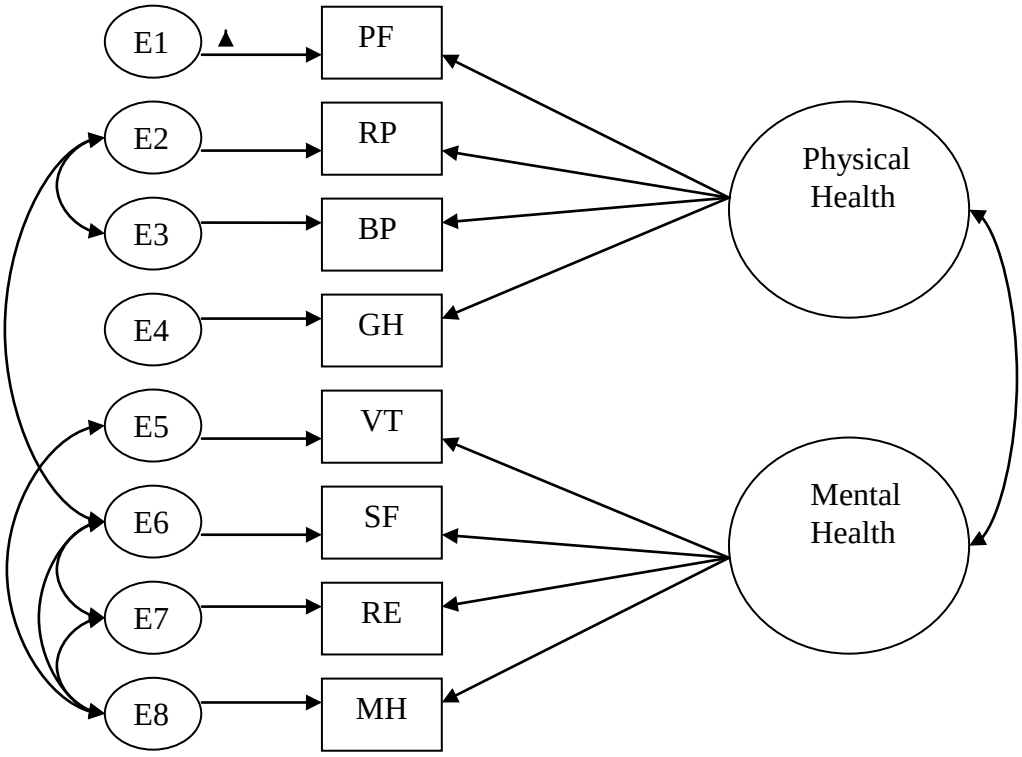
	NAB	0.09 (0.09-0.10)	0.92	0.04	527.22	0.00	14	0.96
NCC	AB	0.05 (0.00-0.11)	0.95	0.06	19.15	0.16	14	0.98
	NAB	0.09 (0.08-0.10)	0.84	0.05	188.93	0.00	14	0.92

Notes: RMSEA = root mean square error of approximation; CI = confidence interval; TLI = Tucker-Lewis index; SRMR = root mean squared residual; df = degrees of freedom; CFI = comparative fit index; *r* = estimated correlation between two latent variables; PML1 = pseudo-maximum likelihood with weights; ML = maximum likelihood; ALOCC = at least one chronic condition; NCC = no chronic condition; AB = Aboriginal; NAB = Non-Aboriginal.

PML1 estimation method was employed. This conclusion was supported by many of the other goodness-of-fit indices. The RMSEA provided support of approximately good fit of this model for non-Aboriginal but not for Aboriginal respondents in both ALOCC and NCC subsamples. Based on the SRMR values, this model can not be rejected, except for Aboriginal respondents in the NCC subsample. Therefore, no test criterion has shown good fit of this measurement model for all groups of Aboriginal and non-Aboriginal in ALOCC and NCC subsamples for PML1 estimator. When an ML estimator was applied for this measurement model, the CFI value indicated a good fit for Aboriginals in ALOCC subsample. The SRMR values also demonstrated the evidence of good fit for Aboriginal and non-Aboriginal groups in the ALOCC subsample. All other fit statistics rejected this model for all groups.

Careful investigation of the modification indices for this model suggested that a substantial improvement in fit could be obtained when the error variances of the indicators were allowed to correlate. Specifically the following correlations had high modification indices: RP and BP; RP and SF; VT and MH; SF and MH; SF and RE; RE and MH. Based on the modification indices a two-factor measurement model (Figure 13) with correlated errors was specified and goodness of fit was assessed using both PML1 and ML estimation methods.

Figure 13: The two-factor model for the SF-36



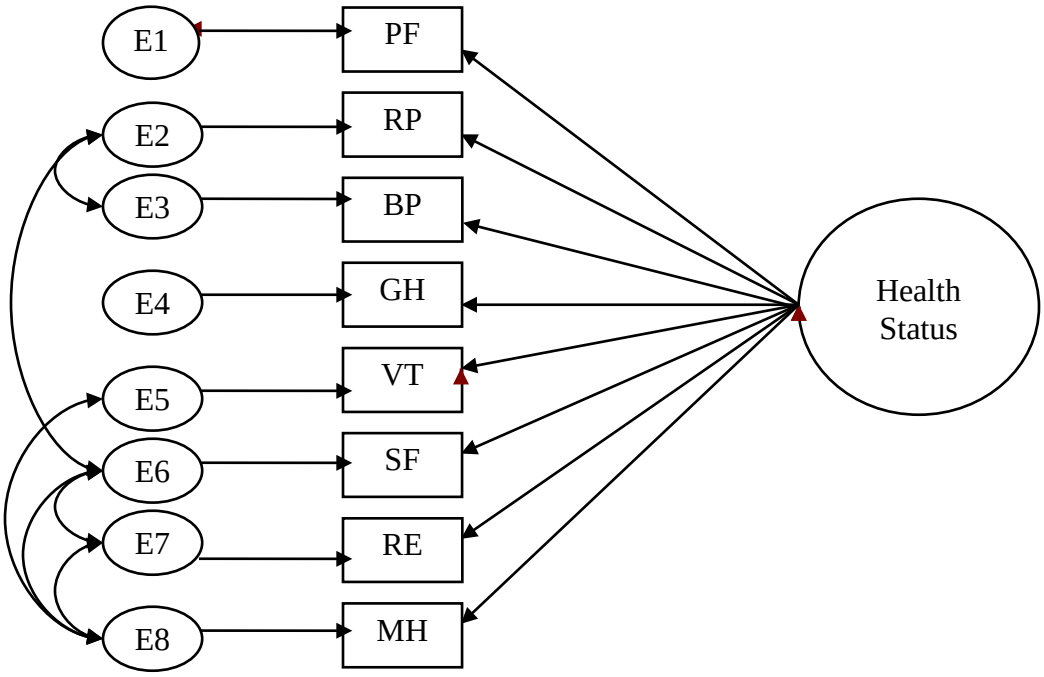
Note to Figure 13: The circles and rectangles indicate the factors (or latent variables) and observed variables (or indicators), respectively. The lines with single arrowheads from factors to observed variables represent the effects of factors on observed variables described by PF, RP, BP, GH, VT, SF, RE, and MH. The other single-headed lines to the observed variables represent the error variances described by E1 to E8. Finally, double-headed curves indicate the correlation between factors or error variances.

The CFI and SRMR statistics indicated an acceptable fit of this model to all groups for PML1 (Table 7). The TLI supported an adequate fit for Aboriginal and non-Aboriginal in the ALOCC subsample. For the RMSEA, a good fit of this model was found for all groups except the Aboriginal group in the NCC subsample. This model was rejected based on the χ^2 test for all groups. It is worthwhile to mention that a significant result for the LRT statistic does not necessary mean that the measurement model is not a good fit to the data because this statistic is sensitive to sample size. Again for the ML estimator, the evaluation of the values of SRMR and CFI obtained sufficiently good fit of this model for

all groups. The assessment of TLI values also suggested good fit except for non-Aboriginal groups in the NCC subsample. Another fit index RMSEA indicated approximately close fit of this model for the Aboriginal group in the NCC subsample; however, acceptable fit was also observed for the groups of Aboriginal and non-Aboriginal in the ALOCC subsample. While the χ^2 test accepted this model for the Aboriginal group in the NCC subsample, for other group this model was rejected by this test. Overall, this measurement model has an acceptable fit for the two groups in the ALOCC and NCC subsamples.

Further analysis revealed that the correlations between the physical and mental health factors were 0.94, 0.91, 0.83, and 0.97 in the Aboriginal with ALOCC, non-Aboriginal with ALOCC, Aboriginal with NCC, and non-Aboriginal with NCC groups, respectively when PML1 estimation was adopted. Kline (2005) suggests that if two factors have high correlations, then one should consider fitting a measurement model with a single factor. Thus, the two factors were combined into a single factor called "Health Status". First one factor measurement model without any correlation among error variances was fitted. None of the fit indices indicated a good fit of the model to the data (Table 7). Therefore, a one-factor measurement model with correlation among error variances (Figure 14) was fitted for Aboriginal and non-Aboriginal groups in subsamples of ALOCC and NCC. Again, the measurement model was evaluated based on fit indices. The CFI and SRMR indices suggested adequate fit of the measurement model (Figure 14) for all groups for both PML1 and ML estimators (Table 7). A good fit was also obtained

Figure 14: The one-factor model for the SF-36



Note to Figure 14: The circles and rectangles indicate the factors (or latent variables) and observed variables (or indicators), respectively. The lines with single arrowheads from factors to observed variables represent the effects of factors on observed variables described by PF, RP, BP, GH, VT, SF, RE, and MH. The other single-headed lines to the observed variables represent the error variances described by E1 to E8. Finally, double-headed curves indicate the correlation between error variances.

by the evaluation of TLI values for all groups except for non-Aboriginal respondents in the NCC subsample with PML1 and ML estimators. The fit criterion RMSEA indicated approximately close fit for non-Aboriginal respondents in both the ALOCC and NCC subsamples when PML1 was used and for Aboriginal respondents in the NCC subsample when ML was used. An acceptable fit was demonstrated by this fit criterion for other groups in the ALOCC and NCC subsamples with PML1 and ML estimators. As usual, the χ^2 test, which is sensitive to sample size, rejected this measurement model for all groups except for Aboriginal in NCC with ML estimators. Hence, the one-factor

measurement model (Figure 14) provided a better fit compared with other measurement models for all groups and estimation methods.

Finally, the one-factor measurement model with correlated error variances (Figure 14) was selected to conduct the MI tests across Aboriginal and non-Aboriginal groups in the ALOCC and NCC subsamples.

4.2.4 Measurement Invariance Tests

With the one-factor model selected as the final measurement model, MI tests were conducted for Aboriginal and non-Aboriginal groups. These analyses were conducted separately for the ALOCC and NCC subsamples using both ML and PML1 estimation methods. The results for the MI tests are summarized in Table 8.

Configural invariance or the baseline model appeared to be an acceptable fit to the data for Aboriginal and non-Aboriginal groups in the subsample ALOCC when the ML was used to estimate the model parameters. The fit indices RMSEA, TLI, SRMR, and CFI supported configural invariance. While the LRT statistic ($p < 0.05$) was statistically significant but it is known that this test is sensitive to sample size. Thus, the hypothesis of configural invariance was retained for the Aboriginal and non-Aboriginal groups in the subsample ALOCC.

Given that configural invariance was established for these two groups, the null hypothesis of metric invariance was tested next. When the ML method was used to estimate the model, the LRT (p=0.0001) resulted in rejection of the hypothesis of metric invariance across Aboriginal and non-Aboriginal groups in the ALOCC subsample. However, the value ofΔCFI, which was less than 0.01, did not support this conclusion. Therefore, the hypothesis of metric invariance was not rejected.

Table 8: Measurement invariance test results for Aboriginal and non-Aboriginal adult respondents, Canadian Community Health Survey, cycle 3.1 (2005/2006).

Invariance Hypothesis	RMSEA (90% CI)	TLI	SRMR	χ^2	df	CFI	$\Delta\chi^2$	Δdf	ΔCFI
ALOCC									
				ML					
Configural	0.09 (0.09, 0.10)	0.92	0.04	581.23*	28	0.96			
Metric	0.09 (0.08, 0.09)	0.94	0.05	610.15	35	0.96	28.92*	7	0.00
Scalar	0.08 (0.08, 0.09)	0.94	0.05	652.90	42	0.96	42.75*	7	0.00
Complete	0.08 (0.07, 0.08)	0.95	0.08	713.76	50	0.95	60.86*	8	0.01
				PML1					
Configural	0.05 (0.05, 0.06)	0.92	0.04	199.56*	28	0.96			
Metric	0.05 (0.04, 0.06)	0.93	0.05	221.52	35	0.96	11.32	7	0.00
Scalar	0.05 (0.04, 0.05)	0.94	0.06	237.20	42	0.96	18.99*	7	0.00
Complete	0.04 (0.04, 0.05)	0.95	0.08	242.27	50	0.96	14.66*	8	0.00
NCC									
				ML					
Configural	0.09 (0.08, 0.10)	0.85	0.05	208.09*	28	0.93			
Metric	0.09 (0.08, 0.10)	0.86	0.08	244.62	35	0.91	36.53*	7	0.02
Scalar	0.08 (0.07, 0.09)	0.87	0.07	272.80	42	0.90	28.18*	7	0.01
Complete	0.08 (0.07, 0.09)	0.88	0.10	301.12	50	0.89	28.32*	8	0.01
				PML1					
Configural	0.05 (0.03, 0.06)	0.89	0.04	76.31*	28	0.94			
Metric	0.06 (0.04, 0.07)	0.84	0.09	121.01	35	0.90	15.99*	7	0.04
Scalar	0.05 (0.04, 0.06)	0.86	0.09	130.95	42	0.90	13.50*	7	0.00
Complete	0.04 (0.03, 0.05)	0.90	0.18	123.43	50	0.91	3.98	8	0.02

Notes: RMSEA = root mean square error of approximation; CI = confidence interval; TLI = Tucker-Lewis index; SRMR = root mean squared residual; df = degrees of freedom; CFI = comparative fit index; $\Delta\chi^2$ = difference in χ^2 values between two nested models; Δdf = difference in df of two nested models; ΔCFI = difference in CFI values between two nested models; PML1 = pseudo-maximum likelihood with weights; ML = maximum likelihood; ALOCC = at least one chronic condition; NCC = no chronic condition. * indicates that the values are statistically significant at $\alpha = 0.05$.

Given that metric invariance was established, a test for scalar invariance was conducted for Aboriginal and non-Aboriginal groups in the ALOCC subsample. The LRT statistic was statistically significant ($p < 0.05$), however, the value of ΔCFI was less than 0.01 for the ML estimation method. Given this finding, there is support for the hypothesis of scalar invariance of the SF-36 for Aboriginal and non-Aboriginal groups. Given this result, the null hypothesis of complete invariance for Aboriginal and non-Aboriginal

groups in the ALOCC subsample was tested. According to the LRT statistic (= <0.0001), the null hypothesis was rejected when ML estimation was adopted. The ΔCFI value for this estimator was equal to 0.01. Thus, the results indicate the complete invariance of the SF-36 for Aboriginal and non-Aboriginal respondents in the ALOCC subsample when ML estimation method was used. Similar to the ML method, results were also obtained for all forms of MI across Aboriginal and non-Aboriginal groups in the ALOCC subsample when model parameters were estimated using PML1 method (Table 8). All fit indices except the LRT statistic ($p < 0.05$) indicate a good fit to the data for configural invariance. Therefore, the configural invariance was also established for PML1 method. The LRT statistic ($p = 0.1253$) was not statistically significant and the value of ΔCFI was also less than 0.01, so the hypothesis of metric invariance was supported by the PML1 method. Then a test for scalar invariance was conducted. Even though the LRT statistics ($p < 0.05$) were statistically significant but the value of ΔCFI was less than 0.01, supporting the hypothesis of scalar invariance for PML1. None of the LRT statistic ($p = 0.0661$) and ΔCFI test rejects the null hypothesis of complete invariance when PML1 estimator was adopted.

In the NCC subsample for Aboriginal and non-Aboriginal groups, configural invariance was established with ML method based on RMSEA, SRMR, and CFI although the TLI and LRT ($p < 0.05$) rejected. Both the LRT ($p < 0.05$) and ΔCFI > 0.01 resulted in rejection of the null hypothesis of metric invariance when ML method was used to estimate the model parameters. Therefore, tests of other MI hypotheses were not conducted for ML method. Configural invariance was tested using PML1 method as well. The results were consistent with ML method, i.e., configural invariance was established

for PML1 method. Given this result, metric invariance was tested for this method. Like ML method, for PML1 method the null hypothesis of metric invariance was rejected by the LRT ($p < 0.05$) and $\Delta CFI > 0.01$. Further tests of MI hypotheses were not conducted for PML1 method.

CHAPTER FIVE: DISCUSSION AND CONCLUSIONS

5.1 Summary and Discussion

This research investigated estimation techniques for testing MI properties in complex survey data by using a simulation study and demonstrated the application of these techniques by a numeric example. In the simulation study, estimation techniques ML, PML1, and PML2 were compared. In the numeric example the MI for the SF-36 were compared for the ML and PML1 estimation methods only. The simulation results suggest that the ML method offered good control of the Type I error rate for the LRT for testing configural invariance when sample size and ICC were small with complex survey data. On the other hand, for high values of the ICC and a large sample size, the PML2 provided better control of Type I error rates for configural invariance than the other methods, ML and PML1. The results are consistent with previous research (Stapleton, 2008) which compared the methods of PML1 and PML2 for SEM analysis of complex survey data. Stapleton found that the factors which influenced the Type I error rates of the LRT included the ICC and sample size, and higher values of the ICC were associated with higher Type I error rates. In the current study this trend was more apparent for ML and PML1 methods than for the PML2 method because the latter method takes the survey design into account but ML and PML1 do not. There was a sharp increase in the Type I error rates with increased sample size for ML and PML1 methods. But the Type I error rates were stable for different sample sizes when PML2 method was used. The Type I error rates were seriously affected for ML and PML1 methods when both ICC and sample size were relatively high. In particular, with the sample size of 1000 and ICC of 0.61, Type I error rates were about two times greater for the ML and PML1 methods than

PML2 method. When constraints such as equality of factor loadings, intercepts, and error variances were imposed between two groups for testing metric, scalar, and complete invariance, respectively, all three methods of estimation rejected the null hypothesis of MI tests similarly. The Type I error rates were close to α = 0.05 for all methods of estimation.

All three estimation methods had similar Type I error rates when using the CFI to test configural invariance. Type I error rates of the CFI were lower than α for all conditions. In terms of the Type I error rates of ΔCFI, all estimation methods produced similar results. The factors that influenced the Type I error rates of ΔCFI were ICC and sample size for testing metric, scalar, and complete invariance. Type I error rates tended to increase with increased values of ICC. In contrast, Type I error rates had a decreasing pattern for increasing sample sizes. Indeed, for a sample of size 400, error rates were 10.97% for PML1 method. On the other hand, with a sample of size 1000, this rate decreased to 0.07%. Previous research has also documented that sample size had an impact on the Type I error rates of LRT and ΔCFI for MI tests using CFA (French & Finch, 2006). However, in that study only the conventional ML method was used.

The findings about the power of the test statistics suggested that the power of the LRT was associated with ICC and sample size. No difference was found by estimation methods. The power of the LRT was decreased with increased value of ICC but it was increased when sample size was increased for metric, scalar, and complete invariance. With small sample size, ML had higher power whereas similar power was observed for all three methods when sample size was high. The power of ΔCFI indicated that the PML2 and PML1 methods had higher power compared with ML but that the PML2

method had the highest power among all methods when ICC was high. This was true for metric, scalar, and complete invariance. An increasing relationship was observed between the power of ΔCFI and ICC for metric and scalar invariance , i.e., the power of ΔCFI increased when ICC was increased, whereas this relationship was opposite for complete invariance, i.e., the power of ΔCFI decreased when ICC was increased.

The results regarding the bias of standardized factor loadings showed that bias was negative for almost all simulation conditions, which implies that on average the estimates of the standardized factor loadings were less than the population parameters. But bias was relatively small for all forms of MI tests when all factor loadings were high and equal in two groups. All estimation methods resulted in highly positive biased estimates of factor loadings when factor loadings were small and unequal in each group with small sample size. For this particular condition, ML was less biased than PML2 method for estimating factor loadings. The findings are consistent with previous research (Enders & Bandalos, 2001) as well. ICC was a factor that had a great impact on factor loadings estimates. Bias increased as ICC increased. Both ML and PML2 estimation methods produced similar magnitudes of biases when factor loadings or sample size were large.

Bias results were both positive and negative when intercepts and some of the factor loadings were unequal in the two groups. When biases were positive, a decreasing pattern was observed as ICC increased; a reverse pattern was found for negative biases. The magnitudes of biases were greater for unequal factor loadings than equal factor loadings in the two groups. However, biases were similar for estimation methods ML and PML2.

Some studies (Stapleton, 2006 & 2008) showed that the PML2 method was superior to the PML1 for SEM analysis using complex survey data. In this research CFA techniques were used to test for MI across two groups, unlike the Stapleton's research, which focused on fitting a single model to complex survey data. Stapleton used a larger sample size (14,400) compared to a maximum sample size of 1,000 in this research. Higher sample size is associated with higher Type I error rates of the LRT. Moreover, she used only one ICC of 0.50, which is relatively high. This research also found higher Type I error rates for the ML and PML1 than the PML2 when ICC was high (i.e., 0.61). In particular for configural invariance, with a sample size of 1,000 and ICC of 0.61, the Type I error rates for the ML and PML1 were about 2 times greater than for the PML2 method. However, for testing metric, scalar, and complete invariance; all three methods produced similar Type I error rates for the LRT. These results were observed, probably, due to the fact that the LRT statistic was calculated as the difference between two χ^2 values for constrained and unconstrained models (i.e., $\Delta\chi^2$), which was tested for MI. On the other hand, for configural invariance χ^2 value itself was tested.

The analysis of the Canadian Community Health Survey data was undertaken to compare different estimation methods for testing hypotheses about MI in a real dataset. Specifically, an analysis was conducted for MI of the SF-36 across Aboriginal and non-Aboriginal Manitoba populations with and without chronic health conditions. The descriptive analyses revealed that asthma, chronic bronchitis, diabetes, and heart disease were significantly higher in Aboriginal than non-Aboriginal groups. On the other hand, arthritis, bowel disorders, cancer, high blood pressure, and stroke were significantly lower in Aboriginal than non-Aboriginal groups. Ethnic comparisons revealed that

Aboriginal respondents with chronic condition had lower mean scores on all but one (PF) of the domains of the SF-36 than non-Aboriginal respondents with chronic conditions. This finding is consistent with other studies which have explored the relationship between HRQOL and specific chronic condition, for example, Thommasen and Zhang (2006) reported that mean scores of domains of the SF-36 for Aboriginal respondents with diabetes were lower than these for non-Aboriginal respondents.

The SF-36 is a widely used general measure of quality of life and an appropriate tool to compare HRQOL in different populations (Buchholz, Krol, Rist, Nieuwkerk, & Schippers, 2008; Schlenk et al., 1998). The MI of the SF-36 was assessed for two ethnic groups, Aboriginal and non-Aboriginal, in the CCHS. The aim was to determine whether the SF-36 has the same conceptualization or meaning across different ethnic groups. The well-established and commonly used CFA techniques were applied to test hypotheses about MI. Based on previous research (Keller et al., 1998), initially a two-factor measurement model was chosen for the data; the two factors were physical and mental health. This model did not fit the data well for Aboriginal and non-Aboriginal groups in each of the subsamples of ALOCC and NCC. In order to improve the fit of this measurement model, under the guidance of modification indices, correlations among error variances were included, which was consistent with previous research (Lix et al., 2009). There was a high correlation between the two factors. These two factors were combined into a single Health Status factor, which is recommended in the SEM literature (Kline, 2005). Therefore, a series of CFA models was conducted for testing MI of a one-factor structure of the SF-36 across Aboriginal and non-Aboriginal groups. The estimation methods PML1 and ML were used in the CCHS data analysis. The other

method PML2 was not used because this method requires cluster or strata information and neither of these were found in the data.

The results of CFA supported the configural invariance of the SF-36 for Aboriginal and non-Aboriginal respondents for both the ALOCC and NCC subsamples, regardless of the method of estimation. This implies that the same construct of the SF-36 was measured in each of four groups. For the ALOCC subsample, metric invariance was also established for Aboriginal and non-Aboriginal groups for the two estimation methods, which suggested that the questions or items of the SF-36 had equivalent meaning across the two groups. But for the NCC subsample, none of the estimation methods supported metric invariance across Aboriginal and non-Aboriginal groups, implying that ethnicity can influence interpretation of one's well-being in healthy populations. Further tests of MI were not conducted in the NCC subsample because invariance was not satisfied.

This study used strong criteria for establishing MI, by testing the equality of intercepts and error variances, respectively, across groups. Vandenberg and Lance (2000) found that only about 12% of studies tested these forms of invariance, which are known as scalar and strong invariance, respectively. Scalar invariance was supported only in the ALOCC subsample for the two groups regardless of whether the ML and PML1 were employed. This result indicated that the measurement of the latent variables as well as the means of indicators were equivalent across groups. For valid comparison of indicator means across groups, scalar invariance should be satisfied (Meredith & Teresi, 2006). Complete invariance, the strongest form of MI, was also established in the two groups within the ALOCC subsample for both ML and PML1 methods of estimation, which

allows making unbiased group comparisons on the SF-36 global or domain scores. The results indicate that there is no difference between ML and PML1 estimation methods, which is consistent with the simulation results for testing MI using CFA techniques in complex survey data. A recent study also found the MI of the SF-36 for Aboriginal and non-Aboriginal women (Lix et al., 2009).

5.2 Conclusions

The results of the simulation study suggest that the performance of the three estimation methods (ML, PML1, and PML2) was affected by ICC and sample size for testing MI using CFA techniques in complex survey data. Based on the Type I error rates of the LRT, the following conclusions can be made: ML is an appropriate method to adopt for testing configural invariance when sample size and ICC are small in complex survey data; for large sample sizes and high ICC, the PML2 estimation method is the recommended method for conducting the LRT. With other types invariance (metric, scalar, and complete), the three estimation methods ML, PML1, and PML2 had similar Type I error rates for the LRT.

The CFI behaved somewhat different for MI tests in complex survey data. For configural invariance, all three estimation methods produced similar Type I error rates for the CFI statistic. Larger sample sizes were associated with lower Type I error rates but the greater values of ICC were associated with higher Type I error rates for ΔCFI when metric, scalar, and complete invariances were tested. In addition, any one of the three methods appears to be a good candidate for the ΔCFI test to investigate metric, scalar, and complete invariance.

With respect to the power of the LRT, ML was more powerful method for detecting the difference in factor loadings and intercepts when sample size was small. However, for larger sample sizes, power rates were similar with all methods of estimation. The Power of the LRT decreased when the ICC increased in value. Larger sample size was associated with decreased power of the CFI for testing configural invariance and also decreased power of the ΔCFI statistic for testing hypotheses about metric, scalar and complete invariance. But a high ICC was associated with increased power of the ΔCFI statistic for testing metric and scalar invariance and decreased power for complete invariance. Compared with the ML, the PML1 and PML2 had higher power for the ΔCFI statistic to detect differences in factor loadings and intercepts between two groups. Biases of standardized factor loadings were similar for all three methods of estimation.

Overall the findings suggest that to test configural invariance using CFA techniques in complex survey data, ML is a good method with small ICC and small sample size, whereas PML2 is appropriate with high ICC and large sample size. Moreover, the PML1 method was similar to the PML2 method in the case of large sample sizes with low ICC. In addition, for testing metric, scalar, and complete invariance, no clear superiority was found for one method than the others.

MI results of the SF-36 in the CCHS cycle 3.1 for Aboriginal and non-Aboriginal respondents who reported having at least one chronic condition imply that the two groups have similar conceptualizations of their quality of life. But this conceptualization appears to be different in healthy populations. MI of the SF-36 was not supported for respondents who did not report having any chronic conditions, suggesting that ethnicity can influence

interpretation of one's well-being in healthy populations. In terms of estimation methods, the ML and PML1 performed similarly to test MI of the SF-36 when the data were collected using a complex survey design. It was not possible to apply the PML2 to the CCHS data because this method requires information about strata and cluster variables; neither strata nor cluster variable was available in the CCHS cycle 3.1 data.

5.3 Strengths of the Study

There were several strengths of the simulation study which compared the estimation methods of ML, PML1, and PML2 to establish MI of a two-factor structure in complex survey data. The strengths include the choice of factor structure, complexity of the study design, estimation methods, and simulation conditions investigated. The factor structure that was investigated for MI was similar to that of the SF-36, as evidenced in the published literature. The SF-36 has been validated and is a widely used general measure of quality of life (Bjorner, Kreiner, Ware, Damsgaard, & Bech, 1998; de Vet et al., 2005). In terms of the study design, a two-stage complex survey design which involved clusters and weights was adopted for the simulation study. The estimation methods that were chosen to compare for investigating MI were recommended methods from previous research and are not as computationally intensive as jackknife and bootstrap methods for researchers to implement. In particular, the ML method is commonly used in CFA and is available in many software packages including Mplus. The other methods PML1 and PML2 can take sample weights and survey design into account when these are applied in the context of SEM analysis for complex survey data. If weights or clusters are not incorporated in the analysis, bias may arise in the parameter estimates. Simulation conditions that were manipulated in this study were selected based

on previous research. The conditions included factor loadings, intercepts, correlation between two factors, ICC, size of clusters, and sample size, which affect parameter estimates and test statistics.

The CCHS is a population-based health survey which uses a representative sample of Canadian population to investigate health status. Large numbers of respondents participate in the CCHS. As a result, it was possible to test for MI among different ethnic groups. The presence of chronic health conditions may be a potential confounder in the analysis, therefore, separate analyses was conducted for ALOCC and NCC subsamples. Strong criteria were adopted to test MI across Aboriginal and non-Aboriginal groups in each subsample. In particular, MI of the SF-36 was established only when all four forms (i.e., configural, metric, scalar and complete) of MI were satisfied. Previous research suggests that only about 12% of studies test scalar or strong invariance (Vandenberg & Lance, 2000). A well-established CFA method was used for conducting MI tests across the two groups. The effects of complex survey design were included in the CFA model to obtain valid results of MI tests.

5.4 Limitations of the Study

There were some limitations of this study. The design of the simulation study reflects the design of the CCHS but there are some differences. In the first stage of selection, clusters were selected using the same PPS method that was used in the CCHS design. But in the second stage of selection, SRS was used instead of systematic sampling that was adopted in the CCHS. However, the survey design adopted in the simulation study was consistent with the designs used in previous research for complex survey data (Stapleton, 2006). An important aspect of this study was that a particular baseline model

was properly specified and simulated. It is not clear whether the results of this study can be applied in a situation when the baseline model is different. Imbalanced sample sizes in two groups may influence MI test results. This was assumed in the current simulation study due to disproportionate representativeness of Aboriginal and non-Aboriginal respondents in the CCHS. Sample size and the number of indicators per factor have been shown to influence the performance of the LRT (French & Finch, 2006). In this study, three sample sizes were considered, but the number of indicators per factor was held constant. Therefore the results may not generalize to other sample sizes when there are a different number of indicators per factor. French and Finch (2006) also commented that the number of factors has some impact on the power of the LRT as well as the CFI.

A limitation in the analysis of CCHS data is that there were differences in the age distribution of Aboriginal and non-Aboriginal groups. Therefore, age could be a confounder in the analysis of MI. The sample size in the Aboriginal group was small, which did not allow the models to be stratified by both age and ethnicity. MI was tested for a single general quality of life measure; the results can not be generalized to other measures that might have been applied in the study populations, such as the EuroQol (Kind, Brooks, & Rabin, 1996) or Sickness Impact Profile (Berger et al., 1981). This research used only a single statistical method, CFA, to test MI. There are other methods besides CFA that can be used to assess MI. For example, item response theory (Bjorner et al., 1998) has been proposed for testing MI. However, these methods have not been investigated for testing MI in complex survey data. In addition, some other methods which may be applied to test MI in complex survey data, including JRR, BRR, and bootstrapping, were not investigated (Stapleton, 2008).

The CCHS data contain information only on self-reported ethnicity. This fact may influence the validity of the results because not all individuals may correctly self-report their ethnicity. The CCHS cycle 3.1 data excluded individuals living on Indian Reserves, which affects the generalizability of the results to the entire Aboriginal population in Manitoba. As well, the study did not investigate differences between Aboriginal and non-Aboriginal respondents for all of Canada because the SF-36 was an optional content in the CCHS data collection and was not administered in all provinces and territories. The non-Aboriginal population included respondents from many different ethnic groups, that is, who were not Aboriginal. The results may be influenced by the fact that all ethnic groups in Aboriginal and non-Aboriginal populations were assumed to be homogeneous, and we did not test for MI for different ethnic sub-groups within the non-Aboriginal population.

A one-factor measurement model was used to test MI of the SF-36. Some studies have demonstrated that a two-factor measurement model was not a good fit to the SF-36 data. For example, a study in a Jamaican population with chronic sickle cell disease found that a three-factor measurement model provided a better fit to the data: physical health, mental health, and role limitations (Asnani, Lipps, & Reid, 2007). Therefore, it was important to carefully evaluate the measurement model for the CCHS data. However, this one-factor model may not be valid for Aboriginal and non-Aboriginal populations in other provinces.

5.5 Future Research

This research investigated the MI of a single HRQOL measure. In a future study, it would be important to test the MI of other HRQOL measures. In particular, MI should

be evaluated for other general HRQOL measures that could be applied in population-based studies such as the CCHS. As well, other methods of testing MI in complex survey data, including empirical resampling methods should be compared to the conventional ML method.

Studies about the psychometric properties of HRQOL measures have drawn increased attention in recent years for cross-sectional as well as longitudinal studies (Limbers et al., 2008; Lix et al., 2009; Varni et al., 2008). Like cross-sectional studies, in longitudinal studies there is increasing recognition of the importance of evaluating the measurement properties of HRQOL measures over time to ensure that changes in HRQOL measures reflect true differences in the population. MI over time is also known as response shift (Sprangers & Schwartz, 1999). There is no published research which has examined the response shift in complex survey data that employs clustered or stratified sampling methodologies. Although estimation methods such as PML, JRR, BRR, and bootstrap methods (Stapleton, 2008) were recommended for SEM with cross-sectional complex survey data, these methods have never been investigated in the context of assessing response shift in longitudinal using CFA techniques. Therefore, in future research, the investigation of these methods for response shift in SEM analysis would be a potential topic for making substantive contribution to the SEM literature on measurement of HRQOL.

In the CCHS data analysis, MI of the SF-36 for Aboriginal and non-Aboriginal groups was tested in each of the subsamples of ALOCC and NCC. For a large sample size, MI can be tested for a sample with a single chronic health condition, e.g., diabetes. Also, the data for the SF-36 do not appear to exhibit a normal distribution for all

indicators. In particular, for the CCHS 3.1 data, univariate measures of kurtosis for some of the indicators of the SF-36 indicate substantial departures from normality. Therefore, there is a need for future research about estimation methods for non-normal data.

REFERENCES

Ahmed, S., Mayo, N. E., Corbiere, M., Wood-Dauphinee, S., Hanley, J., & Cohen, R. (2005). Change in quality of life of people with stroke over time: True change or response shift? *Quality of Life Research*, *14*, 611–627.

Alonso, J., Ferrer, M., Gandek, B., Ware, Jr., J. E., Aaronson, N. K., Mosconi, P., … Leplege, A. (2004). Health related quality of life associated with chronic conditions in eight countries: Results from the International Quality of Life Assessment (IQOLA) Project. *Quality of Life Research*, *13*, 283-298.

Arbucle, J. (2005). Amos 6.0 User's Guide. Chicago: IL, Amos Development Corporation.

Asnani, M., Lipps, G., & Reid, M. (2007). Component structure of the SF-36 in Jamaicans with sickle cell disease. *West Indian Medical Journal*, *56*(6), 491-497.

Asparouhov, T. (2004). Stratification in multivariate modeling. Retrieved on August 8, 2004, from http://www.statmodel.com/mplus/examples/webnotes/MplusNote921.pdf

Asparouhov, T. (2005a). Sampling weights in latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*, 411-434.

Asparouhov, T., & Muthén, B. (2005). Multivariate statistical modeling with survey data. *Proceedings of the Federal Committee on Statistical Methodology (FCSM) Research Conference.* Retrieved on September 26, 2006, from http://www.fcsm.gov/05papers/Asparouhov_Muthen_IIA.pdf

Asparouhov, T., & Muthén, B. (2006). Comparison of estimation methods for complex survey data analysis. *Mplus Web Notes.* Retrieved on June 20, 2009 from http://www.statmodel.com/resrchpap.shtml

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, *107*, 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Bentler, P. M, & Wu, E. (2002). EQS 6 for Windows User's Guide. Multivariate Software, Inc., Encino, CA.

Bergner, M. (1989). Quality of life, health status and clinical research. *Medical Care*, *27,* S148-S156.

Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The Sickness Impact Profile: development and final revision of a health status measure. *Medical Care*, *19*, 787-805.

Bjorner, J. B., Kreiner, S., Ware, J. E., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish transition of the SF-36. *Journal of Clinical Epidemiology*, *51*, 1189-1202.

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

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (eds). *Testing structural equation models* (pp. 136-162). Newbury Park: Sage Publications.

Buchholz, A., Krol, A., Rist, F., Nieuwkerk, P., & Schippers, G. (2008). An assessment of factorial structure and health-related quality of life in problem drug users using the Short Form 36 health survey. *Quality of Life Research, 17,* 1021-1029.

Camilleri-Brennan, J., & Steele, R. (1999). Measurement of quality of life in surgery. *Journal of the Royal College of Surgeons of Edinburgh, 44,* 252-259.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*, 464-504.

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

Crockett, L. J., Shen, Y. L., Randall, B. A., Russell, S. T., & Driscoll, A. K. (2005). Measurement equivalence of the Center for Epidemiological Studies Depression Scale for Latino and Anglo adolescents: A National Study. *Journal of Consulting and Clinical Psychology, 73*, 47-58.

de Vet, H. C. S., Ader, H. J., Berwee, C. B., & Pouwer, F. (2005). Are factor analytic techniques used appropriately in the validation of health status questionnaires? A systematic review on the quality of factor analysis of the SF-36. *Quality of Life Research, 14,* 1203-1218.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 8*, 430-457.

Feldt, T., Lintula, H., Suominen, S., Koskenvuo, M., Vahtera, J., & Kivimaki, M. (2007). Structural validity and temporal stability of the 13-item sense of coherence scale:

Prospective evidence from the population-based HeSSup study. *Quality of Life Research, 16,* 483–493.

Flora, D. B., & Corran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466-491.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and Research refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286-299.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 13,* 378-402.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 11,* S78-S94.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103,* 265-275.

Guyatt, G., Mitchell, A., Irvine, E. J., Singer, J., Williams, N., Goodacre, et al. (1989). A new measure of health status for clinical trials in inflammatory bowel disease. *Gastroenterology, 96,* 804-810.

Guyatt, G. H., Feency, D. H., & Patrick, D. L. (1993). Measuring health related quality of life. *Annals of Internal Medicine, 118,* 622-629.

Hunt, S. M., McEwen, J., & McKenna, S. P. (1985). Measuring health stats: a new tool for clinicians and epidemiologists. *Journal of the Royal College of General Practitioners, 35,* 185-188.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 1-55.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology. *Journal of Cross-Cultural Psychology, 16,* 131-152.

Jordan-Marsh, M. (2002). The SF-36 Quality-of-Life Instrument: Updates and Strategies for Critical Care Research. *Critical Care Nurse, 22,* 35-43.

Joreskog, K.G., & Sorbom, D. (1996). LISREL 8: User's reference guide. Chicago, IL: Scientific Software International, Inc.

Kaplan, D., & Ferguson, A. J. (1999). On the utilization of sample weights in latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 305-321.

Keller, S. D., Ware, Jr, J. E., Bentler, P. M., Aaronson, N. K., Alonso, J., Apolone, G. et al. (1998). Use of structural equation modeling to test the construct validity of the SF-36 health survey in ten countries: Results from the IQOLA project. *Journal of Clinical Epidemiology, 51,* 1179-1188.

Kind, P., Brooks, R., & Rabin, R. (1996). *EQ-5D concepts and methods: A development history*. Netherland: Springer.

Kish, L., & Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B, 36,* 1-37.

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. 2nd ed. New York: Giulford Press.

Krause, E. D., Kaltman, S., Goodman, L. A., & Dutton, M. A. (2007). Longitudinal factor structure of posttraumatic stress symptoms related to intimate partner violence. *Psychological Assessment, 19,* 165–175.

Lam, C. L. K., & Lauder, I. J. (2000). Impact of chronic diseases on the health-related quality of life (HRQOL) of Chinese patients in primary care. *Family Practice, 17*(2), 159-166.

Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 12,* 1-27.

Limbers, C. A., Newman, D. A., & Varni, J. W. (2008). Factorial invariance of child self-reports across healthy and chronic health condition groups: A confirmatory factor analysis utilizing the PedsQL™ 4.0 Generic Core Scale. *Journal of Pediatric Psychology, 33,* 630-639.

Lix, L. M., Metge, C., & Leslie, W. D. (2009). Measurement equivalence of osteoporosis-specific and general quality-of-life instruments in Aboriginal and non-Aboriginal women. *Quality of Life Research, 18,* 619-627.

Loehlin, J. C. (2004). *Latent variable models: An introduction to factor, path, and structural equation analysis*. 4th ed. Mahwah, N. J.: Lawrence Erlbaum Associates.

Longford, N. T. (1995). *Model-based methods for analysis of data from 1990 NAEP trial state assessment.* Washington DC: National Centre for Education Statistics, 95-696.

Lubke, G., & Muthén, B. (2004). Factor-analyzing Likert scale data under the assumption of multivariate normality complicates a meaningful comparison of observed groups or latent classes. *Structural Equation Modeling: A Multidisciplinary Journal, 11*, 514-534.

Makikangas, A., Feldt, T., Kinnunen, U., Tolvanen, A., Kinnunen, M. L., & Pulkkinen, L. (2006). The factor structure and factorial invariance of the 12-item General Health Questionnaire (GHQ-12) across time: Evidence from two community-based samples. *Psychological Assessment, 18,* 444–451.

Marsh, H. W., & Yeung, A. S. (1996). The distinctiveness of affects in specific school subjects: An application of confirmatory factor analysis with the National Education Longitudinal Study of 1988. *American Educational Research Journal, 33*, 665-689.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(Suppl 3), 69-77.

Mora, P. A., Contrada, R. J., Berkowitz, A., Musumeci-Szabo, T., Wisnivesky, J., & Halm, E.A. (2009). Measurement invariance of the Mini Asthma Quality of Life Questionnaire across African–American and Latino adult asthma patients. *Quality of Life Research, 18*, 371–380.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 599-620.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus User's Guide*. 5th ed. Los Angeles, CA: Muthen & Muthen.

Muthén, B. O. & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267-316.

O'Boyle C. A. (1992). Assessment of quality of life in surgery. *British Journal of Surgery, 79*, 395-398.

Olschewski, M., Schilgen, G., Schumacher, M., & Altman, D. G., (1994). Quality of life assessment in clinical cancer research. *British Journal of Cancer, 70*, 1-5.

Peek, M. K., Ray, L., Patel, K., Stoebner-May, D., & Ottenbacher, K. (2004). Reliability and validity of the SF-36 among older Mexican Americans. *The Gerontologist, 44*, 418-425.

Pfeffermann, D. (1993). The role sampling weights when modeling survey data. *International Statistical Review, 61*, 317-337.

SAS Institute Inc. (2009). *SAS/STAT user's guide, version 9.2*. Cary, NC: Author.

Satorra, A., & Bentler, P.M. (1988). Scaling Corrections for Chi-Square Statistics in Covariance Structure Analysis. *Proceedings of the Business and Economic Statistics Section of the American Statistical Association*, 308-313.

Satorra, A., & Bentler, P. M. (2001). A scaled difference $\chi 2$ test statistic for moment structure analysis. *Psychometrika, 66*, 507-514.

Schlenk, E. A., Erlen, J. A., Dunbar-Jacob, J., McDowell, J., Engberg, S., Sereika, ... Bernier, M. J. (1998). Health-related quality of life in chronic disorder: a comparison across studies using the MOS SF-36. *Quality of Life Research, 7,* 57-65.

Schultz, S. E., & Kopec, J. A. (2003). Impact of chronic conditions. *Health Reports, 14*(4), 41-53.

Sheila, S. T. (2005). Implications for quality of life research in Latino populations. *Journal of Transcultural Nursing, 16,* 136-141.

Skinner, C. J. (1989). Domain Means, Regression and Multivariate Analysis. In C. J. Skinner, D. Holt, & T. M. F. Smith (eds) *Analysis of Complex Surveys* (pp. 59-88). Chichester: Wiley.

Sprangers, M. A. G., de Regt, E. B., Andries, F., van Agt, H. M. E., Bijl, R. V., de Boer, J. B. et al. (2000). Which chronic conditions are associated with better or poorer quality of life? *Journal of Clinical Epidemiology, 53,* 895-907.

Sprangers, M. A., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Social Science &Medicine, 48*(11), 1507-1515.

Stadnyk, K., Calder, J., & Rockwood, K. (1998). Testing the measurement properties of the Short Form-36 health survey in a frail elderly population. *Journal of Clinical Epidemiology, 51,* 827-835.

Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling: A Multidisciplinary Journal, 13,* 28-58.

Stapleton, L. M. (2008). Variance estimation using replication methods in structural equation modeling with complex sample data. *Structural Equation Modeling: A Multidisciplinary Journal, 15*, 183-210.

Statistics Canada (2006). Canadian Community Health Survey Cycle 3.1. Public Use Microdata File User Guide.

Stewart, A. L., Greenfield, S., Hays, R. D., Wells, K., Rogers, W. H., Berry, S. D., et al. (1989). Functional status and well-being of patients with chronic conditions. *Journal of the American Medical Association, 262,* 907-913.

Testa, M. A., & Simonson, D. C. (1996). Assessment of quality of life outcomes. *New England Journal of Medicine, 334,* 835-840.

Thommasen, H. V., & Zhang, W. (2006). Impact of chronic disease on quality of life in the Bella Coola Valley. *Rural and Remote Health, 6,* 528.

The WHOQOL Group (1998a). Development of the World Health Organization WHOQOL-BREF quality of life assessment. *Psychological Medicine, 28,* 551-558.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1-10.

Turner-Bowker, D. M., Bartley, P. J., Ware, J. E. (2002). SF-36® Health Survey & "SF" Bibliography: 3rd Ed. Lincoln (RI): QualityMetric Incorporated.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70.

Varni, J., Limbers, C., Newman, D., & Seid, M. (2008). Longitudinal factorial invariance of the PedsQLTM 4.0 Generic Core Scales Child Self-Report Version: One year prospective evidence from the California State Children's Health Insurance Program (SCHIP). *Quality of Life Research, 17,* 1153-1162.

Ware, Jr. J. E. (1987). Standards for validating health measures: definition and content. *Journal of Chronic Diseases, 40,* 473-480.

Ware, Jr., J. E., & Sherbourne, C. (1992). The MOS 36-item Short-Form Health Survey 1: conceptual framework and item selection. *Medical Care, 30,* 473-483.

Ware, Jr., J. E., Snow, K. K., Kosinski, M., & Gandek, B. (1993). SF-36 health survey: Manual and interpretation guide. Boston M A: The Health Institute, New England Medical Center.

Wee, H. L., Cheung, Y. B., Li, S. C., Fong, K. Y., & Thumboo, J. (2005). The impact of diabetes mellitus and other chronic medical conditions on health-related quality of life: Is the whole greater than the sum of its parts? *Health and Quality of Life Outcomes, 3,* 1-11.

Wood-Dauphinee, S. (1999). Assessing quality of life in clinical research: from where have we come and where are we going? *Journal of Clinical Epidemiology, 52,* 355-363.

Wu, C., Lee, K., & Yao, G. (2006). Examining the hierarchical factor structure of the SF-36 Taiwan version by exploratory and confirmatory factor analysis. *Journal of Evaluation in Clinical Practice, 13,* 889-900.

Yao, G., & Wu, C. H. (2005). Factorial invariance of the WHOQOL-BREF among disease groups. *Quality of Life Research, 14,* 1881-1888.

Yu, C. H. Y., & Zinman, B. (2007). Type 2 diabetes and impaired glucose tolerance in aboriginal populations: A global perspective. *Diabetes Research and Clinical Practice, 78*(2), 159-170.

Yuan, K-H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40*, 115-148.

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica, 67,* 361-370.

Zimprich, D., Allemand, M., & Hornung, R. (2006). Measurement invariance of the abridged sense of coherence scale in adolescents. *European Journal of Psychological Assessment, 22*(4), 280–287.

APPENDIX A: SIMULATION STUDY RESULTS

Table A- 1: Average Type I error rates (%) of the likelihood ratio test for four forms of measurement invariance.

		Configural			Metric			Scalar			Complete		
N	ICC	ML	PML1	PML2	ML	PML1	PML2	ML	PML1	PML2	ML	PML1	PML2
Factor loadings: Pattern A													
400	0.00	6.80	8.83	9.07	6.60	5.83	5.40	5.17	6.10	3.47	5.33	9.93	4.03
	0.27	6.07	10.17	8.53	6.13	3.90	5.00	6.17	10.10	4.03	5.43	11.13	4.50
	0.61	10.30	11.83	10.73	4.93	4.33	4.47	5.40	3.07	4.13	5.37	11.57	4.27
650	0.00	6.20	7.27	6.57	6.70	6.00	6.17	6.10	4.63	5.30	4.40	4.83	4.70
	0.27	6.70	7.57	8.03	5.37	6.00	5.30	5.67	4.83	5.63	4.87	4.97	4.67
	0.61	10.70	11.43	7.80	5.87	11.23	5.63	4.67	11.37	4.97	5.40	11.97	5.37
1000	0.00	6.73	6.53	7.23	5.00	3.90	4.30	5.93	4.30	5.03	4.60	4.47	3.80
	0.27	6.73	7.00	6.93	6.20	5.23	5.93	5.33	4.33	4.23	4.77	4.03	4.37
	0.61	15.27	15.00	7.97	4.83	4.03	4.67	4.10	3.67	4.33	4.40	4.27	4.47
Factor loadings: Pattern B													
400	0.00	6.13	10.00	9.53	7.80	3.77	6.17	6.07	4.53	4.40	5.87	4.03	4.33
	0.27	6.13	10.27	10.50	7.23	4.03	5.73	5.70	3.53	4.57	5.27	5.00	4.50
	0.61	10.13	14.20	11.87	6.50	4.57	5.10	6.13	3.30	5.00	5.70	4.80	5.47
650	0.00	5.70	7.83	7.20	6.43	5.33	5.83	5.43	4.47	4.57	4.97	5.10	5.10
	0.27	6.23	8.30	6.80	6.27	5.57	6.00	6.20	5.17	5.87	5.27	5.47	5.43
	0.61	10.73	13.13	8.27	6.43	5.40	6.17	4.97	4.83	5.17	5.20	4.30	5.83
1000	0.00	6.50	6.13	6.80	5.87	8.00	5.33	5.00	4.07	4.37	5.50	4.30	4.83
	0.27	6.63	6.53	6.60	6.87	4.67	5.87	4.83	5.10	4.20	4.93	4.43	4.17
	0.61	17.47	16.87	9.10	5.50	4.57	5.13	4.57	3.73	4.93	4.97	3.87	5.13

Note: ICC = intraclass correlation; ML = maximum likelihood; PML1 = pseudomaximum likelihood with weights; PML2 = pseudomaximum likelihood with weights and clusters.

Table A- 2: Average Type I error rates (%) of differences in comparative fit indices between two nested models for four forms of measurement invariance.

		Configural			Metric			Scalar			Complete		
N	ICC	ML	PML1	PML2	ML	PML1	PML2	ML	PML1	PML2	ML	PML1	PML2
Factor loadings: Pattern A													
400	0.00	0.00	0.00	0.00	1.83	8.80	2.87	0.87	9.37	1.73	1.90	10.97	3.20
	0.27	0.00	0.10	0.03	1.70	5.30	3.20	1.40	5.93	2.23	1.77	3.77	3.63
	0.61	0.00	0.13	0.13	1.90	7.70	2.97	1.93	2.50	3.43	2.87	15.87	4.67
650	0.00	0.00	0.07	0.00	0.20	2.00	0.43	0.13	0.17	0.10	0.27	0.13	0.33
	0.27	0.00	0.07	0.00	0.40	1.97	0.27	0.13	0.20	0.23	0.23	0.23	0.30
	0.61	0.00	0.03	0.00	0.47	3.67	0.57	0.27	4.13	0.40	0.47	4.97	0.93
1000	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.03	0.00	0.07	0.00
	0.27	0.00	0.00	0.00	0.07	0.00	0.03	0.00	0.00	0.03	0.00	0.03	0.03
	0.61	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.03	0.07
Factor loadings: Pattern B													
400	0.00	0.00	0.00	0.00	1.13	1.47	2.30	0.93	1.43	1.37	0.83	1.63	1.70
	0.27	0.00	0.03	0.00	1.60	2.20	2.97	1.00	1.40	1.73	1.00	2.43	2.33
	0.61	0.00	0.07	0.00	2.43	3.83	4.23	1.57	2.10	2.90	2.40	3.80	5.17
650	0.00	0.00	0.00	0.00	0.10	0.20	0.20	0.03	0.10	0.13	0.13	0.07	0.13
	0.27	0.00	0.00	0.00	0.27	0.43	0.37	0.07	0.17	0.20	0.03	0.13	0.17
	0.61	0.00	0.00	0.00	0.37	0.47	1.03	0.13	0.40	0.57	0.40	0.20	0.70
1000	0.00	0.00	0.00	0.00	0.07	0.00	0.03	0.00	0.00	0.03	0.00	0.00	0.00
	0.27	0.00	0.00	0.00	0.00	0.03	0.00	0.03	0.00	0.03	0.00	0.00	0.00
	0.61	0.00	0.00	0.00	0.13	0.03	0.20	0.07	0.17	0.07	0.00	0.00	0.07

Note: ICC = intraclass correlation; ML = maximum likelihood; PML1 = pseudomaximum likelihood with weights; PML2 = pseudomaximum likelihood with weights and clusters.

Table A- 3: Average power rates (%) of the likelihood ratio test for four forms of measurement invariance.

		Configural*			Metric			Scalar			Complete		
N	ICC	ML	PML1	PML2	ML	PML1	PML2	ML	PML1	PML2	ML	PML1	PML2
						Factor loadings: Pattern C							
400	0	5.90	8.80	8.60	31.90	26.17	26.17	29.43	22.20	22.20	26.67	22.83	23.33
	0.27	6.80	10.07	9.60	32.40	26.67	26.27	27.97	21.97	22.30	25.20	20.80	21.80
	0.61	7.30	11.60	10.30	29.57	23.20	23.50	25.10	19.43	20.60	21.30	17.53	18.03
650	0	5.63	6.33	6.47	52.47	49.57	48.97	47.90	44.40	44.17	42.60	41.40	41.40
	0.27	7.47	8.27	8.40	50.50	47.97	48.33	45.47	42.33	42.17	40.67	39.13	39.50
	0.61	9.20	11.27	8.03	43.63	41.23	41.57	45.07	42.47	43.13	32.70	30.53	31.30
1000	0	5.27	6.80	6.80	72.53	68.40	68.13	69.07	64.87	64.37	64.87	61.27	61.00
	0.27	5.67	7.53	6.77	72.33	68.67	67.70	66.40	61.60	62.33	61.47	57.73	57.73
	0.61	11.60	12.07	7.83	63.60	60.27	60.13	61.27	57.87	58.43	48.77	44.50	44.97

Note: ICC = intraclass correlation; ML = maximum likelihood; PML1 = pseudomaximum likelihood with weights; PML2 = pseudomaximum likelihood with weights and clusters; * the rejection rates for configural model are Type I errors because there was no constraint between the two groups for testing this model.

Table A- 4: Average power rates (%) of differences in comparative fit indices between two nested models for four forms of measurement invariance.

		Configural*			Metric			Scalar			Complete		
N	ICC	ML	PML1	PML2	ML	PML1	PML2	ML	PML1	PML2	ML	PML1	PML2
						Factor loadings: Pattern C							
	0	0.27	0.77	0.77	23.97	28.87	28.33	25.60	26.93	26.33	30.07	34.03	34.17
400	0.27	0.27	1.20	1.27	25.33	28.33	28.03	25.93	28.10	28.07	29.37	33.50	34.30
	0.61	1.00	2.33	2.33	25.63	29.47	29.77	26.27	28.83	29.70	29.10	33.03	33.93
	0	0.00	0.00	0.00	22.33	24.50	24.20	24.60	25.63	25.10	28.30	30.13	29.93
650	0.27	0.00	0.00	0.00	23.87	25.20	24.93	25.60	26.40	26.73	27.90	30.00	29.83
	0.61	0.03	0.03	0.03	24.13	24.97	25.13	29.60	29.50	30.23	24.30	26.03	28.30
	0	0.00	0.00	0.00	19.53	21.13	21.47	21.90	22.67	22.70	25.97	27.43	27.20
1000	0.27	0.00	0.00	0.00	21.03	22.80	23.90	22.83	23.20	23.90	25.63	26.70	27.90
	0.61	0.00	0.00	0.00	23.00	24.67	24.67	25.60	26.30	29.60	20.47	22.17	25.90

Note: ICC = intraclass correlation; ML = maximum likelihood; PML1 = pseudomaximum likelihood with weights; PML2 = pseudo-maximum likelihood with weights and clusters; * the rejection rates for configural model are Type I errors because there was no constraint between the two groups for testing this model.

Table A- 5: Average percentage bias of standardized factor loadings (Pattern A) for configural and complete invariance

		B_{21}		B_{31}		B_{41}		B_{62}		B_{72}		B_{82}	
N	ICC	ML	PML2	ML	PML2	ML	PML2	ML	PML2	ML	PML2	ML	PML2
						Configural							
						Group 1							
	0	-0.68	-0.86	1.01	1.00	-0.80	-0.76	-0.11	-0.21	-0.23	-0.22	-0.05	0.00
	0.27	-1.59	-1.59	-0.03	-0.03	-1.70	-1.67	-0.37	-0.35	-1.26	-1.29	-0.62	-0.53
	0.61	-4.11	-4.13	-2.78	-2.81	-3.96	-3.96	-2.78	-2.82	-3.73	-3.72	-3.95	-3.96
						Group 2							
400	0	-0.90	-0.87	-1.08	-1.03	-0.12	-0.19	-0.18	-0.12	-0.11	-0.14	-0.82	-0.88
	0.27	-1.59	-1.62	-2.12	-2.13	-1.03	-1.05	-0.61	-0.59	-1.06	-1.19	-1.72	-1.78
	0.61	-3.88	-3.91	-5.20	-5.14	-4.04	-4.02	-2.98	-3.06	-3.79	-3.84	-4.80	-4.83
						Complete (Group 1 and Group 2)							
	0	-0.58	-0.68	-0.41	-0.40	-0.32	-0.32	0.06	0.04	-0.24	-0.30	-0.46	-0.48
	0.27	-1.37	-1.49	-1.38	-1.37	-1.37	-1.32	-0.11	-0.12	-1.08	-1.17	-1.26	-1.31
	0.61	-3.92	-3.96	-4.19	-4.18	-4.01	-3.95	-2.55	-2.53	-3.95	-3.94	-4.52	-4.48
						Configural							
						Group 1							
	0	-0.70	-0.71	0.89	0.92	-0.47	-0.43	0.36	0.29	-0.28	-0.28	0.55	0.49
	0.27	-1.66	-1.70	0.11	0.14	-1.43	-1.47	-0.49	-0.57	-0.98	-0.95	-0.64	-0.68
	0.61	-4.28	-4.26	-2.88	-2.80	-3.86	-3.88	-2.39	-2.49	-3.69	-3.73	-3.81	-3.83
						Group 2							
650	0	-0.80	-0.80	-1.18	-1.17	-0.14	-0.09	0.06	0.07	-0.26	-0.23	-0.77	-0.70
	0.27	-1.54	-1.50	-2.08	-2.07	-1.19	-1.19	-0.70	-0.70	-1.23	-1.27	-1.85	-1.85
	0.61	-3.91	-3.91	-4.92	-4.90	-3.75	-3.77	-3.01	-3.04	-3.66	-3.66	-4.66	-4.60
						Complete (Group 1 and Group 2)							
	0	-0.53	-0.51	-0.43	-0.43	-0.22	-0.20	0.51	0.46	-0.13	-0.16	-0.11	-0.09
	0.27	-1.36	-1.31	-1.19	-1.18	-1.24	-1.24	-0.20	-0.25	-1.09	-1.10	-1.21	-1.19
	0.61	-3.99	-3.95	-4.25	-4.21	-3.77	-3.79	-2.46	-2.52	-3.77	-3.79	-4.20	-4.19
1000						Configural							

Group 1												
0	-0.64	-0.68	1.01	1.02	-0.50	-0.54	0.20	0.18	-0.06	-0.06	0.52	0.48
0.27	-1.68	-1.71	0.14	0.12	-1.45	-1.40	-0.54	-0.55	-1.02	-1.07	-0.70	-0.75
0.61	-4.11	-4.14	-2.72	-2.72	-4.01	-3.92	-2.68	-2.72	-3.52	-3.57	-3.76	-3.81
Group 2												
0	-0.64	-0.65	-1.09	-1.08	0.04	0.02	0.18	0.20	-0.19	-0.20	-0.66	-0.68
0.27	-1.54	-1.56	-2.07	-2.08	-1.05	-1.06	-0.68	-0.69	-1.17	-1.19	-1.75	-1.70
0.61	-3.94	-3.96	-4.97	-4.97	-3.61	-3.61	-2.87	-2.90	-3.78	-3.86	-4.66	-4.64
Complete (Group 1 and Group 2)												
0	-0.38	-0.37	-0.42	-0.39	-0.23	-0.25	0.54	0.58	-0.10	-0.10	-0.03	-0.08
0.27	-1.38	-1.41	-1.31	-1.29	-1.21	-1.22	-0.26	-0.28	-1.10	-1.14	-1.21	-1.26
0.61	-3.93	-3.96	-4.11	-4.08	-3.73	-3.71	-2.42	-2.41	-3.71	-3.80	-4.15	-4.20

Note: B_{jk} = average percentage bias of standardized factor loadings for observed variable *j* on the *k*th latent variable (*j*=2, 3, & 4 with *k*=1; *j*=6, 7, & 8 with *k*=2); ICC = intraclass correlation; ML = maximum likelihood; PML2 = pseudomaximum likelihood with weights and clusters.

Table A- 6: Average percentage bias of standardized factor loadings (Pattern B) for configural and complete invariance.

		B_{21}		B_{31}		B_{41}		B_{62}		B_{72}		B_{82}	
N	ICC	ML	PML2	ML	PML2	ML	PML2	ML	PML2	ML	PML2	ML	PML2
						Configural							
						Group 1							
	0	81.74	248.08	69.47	202.83	54.9	166.19	0.08	0.09	-0.12	-0.12	0.22	0.25
	0.27	-2.81	413.49	1.91	334.78	-1.73	275.68	-0.64	-0.67	-0.91	-0.94	-0.72	-0.70
	0.61	410.37	492.70	330.39	397.05	273.5	329.37	-3.17	-3.18	-3.44	-3.50	-3.61	-3.61
						Group 2							
400	0	-2.59	-2.62	-3.40	-3.25	0.06	0.17	0.15	0.13	0.04	0.02	-0.10	-0.09
	0.27	-3.10	-2.92	-4.68	-4.71	-1.68	-1.49	-0.62	-0.68	-0.99	-1.02	-1.20	-1.24
	0.61	-6.09	-6.20	-7.91	-7.94	-3.25	-3.25	-3.03	-3.07	-3.48	-3.50	-4.00	-4.08
						Complete (Group 1 and Group 2)							
	0	-1.06	-1.32	-0.50	-0.29	-0.30	-0.21	0.64	0.66	-0.04	-0.08	0.01	0.04
	0.27	-2.32	-2.23	-1.27	-1.40	-1.38	-1.33	-0.06	-0.05	-0.85	-0.86	-1.05	-1.04
	0.61	-5.07	-5.20	-5.16	-5.15	-3.52	-3.32	-2.61	-2.57	-3.42	-3.44	-3.89	-3.90
						Configural							
						Group 1							
	0	-1.77	-1.73	2.56	2.63	-0.58	-0.50	0.14	0.11	-0.04	-0.06	0.27	0.26
	0.27	80.00	-3.16	67.98	1.33	54.2	-1.29	-0.65	-0.67	-0.92	-0.91	-0.79	-0.80
	0.61	-5.20	78.15	-1.31	65.26	-4.03	51.37	-3.19	-3.24	-3.38	-3.40	-3.58	-3.56
						Group 2							
650	0	-1.71	-1.65	-3.21	-3.29	-0.37	-0.48	0.15	0.10	-0.01	-0.01	-0.17	-0.17
	0.27	-2.58	-2.80	-4.57	-4.57	-1.25	-1.37	-0.68	-0.70	-0.95	-0.96	-1.21	-1.20
	0.61	-5.24	-5.17	-7.55	-7.53	-3.44	-3.53	-3.06	-3.07	-3.51	-3.54	-4.06	-4.03
						Complete (Group 1 and Group 2)							
	0	-0.72	-0.60	-0.35	-0.34	-0.33	-0.38	0.83	0.77	0.00	-0.02	-0.03	-0.03
	0.27	-2.07	-2.04	-1.73	-1.68	-0.93	-0.98	-0.02	-0.08	-0.86	-0.89	-1.17	-1.17
	0.61	-4.62	-4.58	-4.64	-4.61	-3.53	-3.62	-2.56	-2.61	-3.40	-3.43	-3.94	-3.93
						Configural							
1000						Group 1							
	0	-1.97	-2.16	2.30	2.30	-0.54	-0.62	0.02	0.01	-0.09	-0.10	0.28	0.26

0.27	-2.96	-3.09	1.71	1.69	-1.47	-1.39	-0.68	-0.64	-0.92	-0.93	-0.70	-0.73
0.61	-5.07	-5.20	-1.35	-1.24	-4.09	-3.96	-3.09	-3.11	-3.39	-3.45	-3.53	-3.53
Group 2												
0	-2.15	-2.06	-3.18	-2.97	-0.43	-0.41	0.21	0.21	-0.02	-0.02	-0.11	-0.12
0.27	-2.65	-2.71	-4.30	-4.20	-1.36	-1.33	-0.57	-0.58	-0.95	-0.94	-1.19	-1.21
0.61	-4.81	-4.87	-7.19	-7.06	-3.94	-3.83	-3.02	-3.05	-3.41	-3.44	-3.99	-4.02
Complete (Group 1 and Group 2)												
0	-1.16	-1.25	-0.55	-0.53	-0.18	-0.26	0.65	0.66	-0.01	-0.02	0.00	-0.02
0.27	-1.94	-2.05	-1.26	-1.21	-1.18	-1.17	-0.07	-0.07	-0.88	-0.89	-1.03	-1.06
0.61	-4.31	-4.36	-4.26	-4.19	-3.92	-3.87	-2.54	-2.58	-3.34	-3.39	-3.87	-3.90

Note: B_{jk} = average percentage bias of standardized factor loadings for observed variable *j* on the *k*th latent variable (*j*=2, 3, & 4 with *k*=1; *j*=6, 7, & 8 with *k*=2); ICC = intraclass correlation; ML = maximum likelihood; PML2 = pseudomaximum likelihood with weights and clusters.

Table A- 7: Average percentage bias of standardized factor loadings (Pattern C) for configural and complete invariance.

		B_{21}		B_{31}		B_{41}		B_{62}		B_{72}		B_{82}	
N	ICC	ML	PML2	ML	PML2	ML	PML2	ML	PML2	ML	PML2	ML	PML2
						Configural							
						Group 1							
	0	-0.99	-1.07	1.15	1.12	-0.23	-0.08	0.23	0.19	0.53	0.66	0.34	0.28
	0.27	-2.25	-2.13	0.29	0.34	-1.04	-1.05	-0.46	-0.44	-0.56	-0.57	-0.68	-0.78
	0.61	-4.29	-4.40	-2.57	-2.45	-3.55	-3.56	-2.62	-2.67	-3.12	-3.14	-3.80	-3.87
						Group 2							
	0	-1.13	-1.15	-2.09	-2.12	-0.09	-0.21	0.57	0.58	-0.18	-0.16	-0.93	-1.16
	0.27	-1.68	-1.82	-3.18	-3.05	-1.24	-1.22	-0.12	-0.18	-1.04	-1.09	-2.19	-2.05
400	0.61	-4.24	-4.18	-6.09	-5.73	-3.42	-3.59	-2.50	-2.49	-3.66	-3.77	-5.26	-5.39
						Complete							
						Group 1							
	0	13.4	13.23	-0.07	-0.03	-9.36	-9.44	-4.60	-4.59	21.84	21.77	-2.32	-2.36
	0.27	12.2	12.03	-0.79	-0.81	-10.3	-10.31	-5.33	-5.28	20.50	20.29	-3.60	-3.55
	0.61	9.27	9.18	-3.78	-3.60	-12.7	-12.70	-7.68	-7.72	17.23	17.10	-6.59	-6.72
						Group 2							
	0	-5.47	-5.64	-0.07	-0.03	3.59	3.50	-4.60	-4.59	-12.9	-13.02	-2.32	-2.36
	0.27	-6.48	-6.65	-0.79	-0.81	2.51	2.50	-5.33	-5.28	-13.9	-14.08	-3.60	-3.55
	0.61	-8.94	-9.01	-3.78	-3.60	-0.22	-0.22	-7.68	-7.72	-16.3	-16.36	-6.59	-6.72
						Configural							
						Group 1							
	0	-1.33	-1.35	1.59	1.61	-0.18	-0.13	0.25	0.22	0.43	0.41	0.61	0.53
	0.27	-1.72	-1.73	0.54	0.56	-1.23	-1.26	-0.49	-0.53	-0.97	-1.01	-0.82	-0.80
650	0.61	-4.29	-4.05	-2.85	-2.90	-3.54	-3.61	-2.60	-2.67	-2.56	-2.84	-3.99	-4.04
						Group 2							
	0	-0.98	-1.00	-2.17	-2.09	0.17	0.11	0.46	0.41	-0.28	-0.23	-1.10	-1.11
	0.27	-1.69	-1.68	-3.19	-3.18	-0.87	-0.95	-0.35	-0.40	-1.18	-1.19	-2.36	-2.37
	0.61	-4.35	-4.26	-6.10	-6.06	-3.34	-3.41	-2.53	-2.54	-3.79	-3.84	-5.26	-5.21
						Complete							

						Group 1							
	0	13.3	13.36	0.19	0.23	-9.13	-9.15	-4.41	-4.48	21.59	21.56	-2.20	-2.20
	0.27	12.3	12.38	-0.80	-0.79	-10.1	-10.11	-5.40	-5.45	20.34	20.31	-3.53	-3.52
	0.61	9.27	9.38	-3.73	-3.70	-12.5	-12.56	-7.56	-7.65	17.47	17.39	-6.44	-6.44
						Group 2							
	0	-5.60	-5.53	0.19	0.23	3.85	3.83	-4.41	-4.48	-13.2	-13.17	-2.20	-2.20
	0.27	-6.43	-6.35	-0.80	-0.79	2.76	2.73	-5.40	-5.45	-14.0	-14.06	-3.53	-3.52
	0.61	-8.94	-8.85	-3.73	-3.70	-0.02	-0.07	-7.56	-7.65	-16.1	-16.15	-6.44	-6.44
						Configural							
						Group 1							
	0	-0.88	-1.07	1.06	1.17	-0.29	-0.26	0.26	0.28	0.65	0.60	0.37	0.31
	0.27	-1.92	-2.05	0.30	0.32	-1.29	-1.24	-0.46	-0.46	-0.53	-0.61	-0.83	-0.89
	0.61	-4.12	-4.35	-2.75	-2.66	-3.72	-3.60	-2.71	-2.65	-2.97	-2.94	-4.14	-4.13
						Group 2							
	0	-1.01	-0.99	-1.98	-1.96	0.14	0.15	0.36	0.37	-0.05	-0.08	-0.96	-0.93
	0.27	-1.60	-1.67	-3.17	-3.17	-1.18	-1.11	-0.29	-0.28	-1.21	-1.22	-2.15	-2.21
1000	0.61	-3.84	-3.94	-6.26	-6.25	-3.66	-3.63	-2.55	-2.54	-3.69	-3.77	-5.25	-5.17
						Complete							
						Group 1							
	0	13.3	13.29	0.07	0.11	-9.23	-9.25	-4.57	-4.57	21.72	21.68	-2.19	-2.21
	0.27	12.2	12.21	-0.95	-0.92	-10.2	-10.22	-5.29	-5.30	20.31	20.24	-3.41	-3.50
	0.61	9.51	9.47	-3.90	-3.87	-12.6	-12.54	-7.61	-7.60	17.35	17.28	-6.65	-6.69
						Group 2							
	0	-5.56	-5.60	0.07	0.11	3.73	3.72	-4.57	-4.57	-13.1	-13.09	-2.19	-2.21
	0.27	-6.47	-6.49	-0.95	-0.92	2.59	2.61	-5.29	-5.30	-14.1	-14.12	-3.41	-3.50
	0.61	-8.74	-8.77	-3.90	-3.87	-0.09	-0.04	-7.61	-7.60	-16.2	-16.23	-6.65	-6.69

Note: B_{jk} = average percentage biases of standardized factor loadings for observed variable *j* on the *k*th latent variable (*j*=2, 3, & 4 with *k*=1; *j*=6, 7, & 8 with *k*=2); ICC = intraclass correlation; ML = maximum likelihood; PML2 = pseudomaximum likelihood with weights and clusters; $B_{31,}$ B_{62} and B_{82} are equal in Group 1 and Group 2 for complete invariance as the corresponding factor loadings were equal.

APPENDIX B: CANADIAN COMMUNITY HEALTH SURVEY DATA ANALYSIS

Table B- 1: Correlations of indicators of the SF-36 for all Manitoba adult respondents, Canadian Community Health Survey, cycle 3.1 (2005/2006).

	PF	RP	BP	GH	VT	SF	RE	MH
PF	1.00							
RP	0.55	1.00						
BP	0.55	0.60	1.00					
GH	0.57	0.44	0.49	1.00				
VT	0.46	0.46	0.49	0.55	1.00			
SF	0.50	0.57	0.52	0.47	0.54	1.00		
RE	0.17	0.27	0.21	0.24	0.33	0.42	1.00	
MH	0.22	0.25	0.29	0.39	0.55	0.47	0.50	1.00
N	6073	6102	6111	5906	5944	6022	6100	5971

Note: PF = physical functioning; RP = role limitations due to physical health problem; BP = bodily pain; GH = general health perception; VT = vitality; SF = social functioning; RE = role limitations due emotional problems; MH = mental health.

Table B- 2: Correlations of indicators of the SF-36 for Manitoba adult respondents by chronic disease status and ethnicity, Canadian Community Health Survey, cycle 3.1 (2005/2006).

	PF	RP	BP	GH	VT	SF	RE	MH	*N*
ALOCC									
PF		0.54	0.53	0.56	0.47	0.49	0.14	0.19	4033
RP	0.58		0.59	0.46	0.47	0.58	0.26	0.25	4059
BP	0.60	0.59		0.48	0.48	0.51	0.18	0.26	4073
GH	0.56	0.39	0.47		0.57	0.48	0.22	0.37	3936
VT	0.46	0.46	0.50	0.54		0.55	0.33	0.53	3964
SF	0.55	0.50	0.54	0.49	0.57		0.40	0.46	4022
RE	0.33	0.32	0.32	0.29	0.37	0.53		0.51	4056
MH	0.38	0.31	0.36	0.48	0.60	0.59	0.58		3981
N	364	365	367	356	359	360	366	359	
NCC									
PF		0.24	0.20	0.26	0.14	0.19	0.03	0.03	1488
RP	0.52		0.43	0.05	0.21	0.40	0.11	0.07	1492
BP	0.56	0.35		0.14	0.33	0.32	0.13	0.20	1492
GH	0.25	0.16	0.12		0.31	0.11	0.03	0.25	1456
VT	0.34	0.34	0.21	0.32		0.31	0.22	0.54	1459
SF	0.29	0.21	0.27	0.13	0.26		0.40	0.34	1473
RE	0.11	0.08	0.05	0.18	0.17	0.41		0.33	1493
MH	0.11	0.04	0.09	0.20	0.43	0.42	0.41		1468
N	133	134	134	130	131	133	134	132	

Note: Lower and upper diagonal correlations are for Aboriginal and non-Aboriginal respondents, respectively, for each of the subsamples of at least one chronic condition (ALOCC) and no chronic condition (NCC). PF = physical functioning; RP = role limitations due to physical health problem; BP = bodily pain; GH = general health perception; VT = vitality; SF = social functioning; RE = role limitations due emotional problems; MH = mental health.

APPENDIX C: SF-36 QUESTIONNAIRE

1. In general, would you say your health is:

- Excellent
- Very Good
- Good
- Fair
- Poor

2. Compared to one year ago, how would you rate your health in general now?

- Much better now than one year ago
- Somewhat better now than one year ago
- About the same
- Somewhat worse now than one year ago
- Much worse than one year ago

The following items are about activities you might do during a typical day. Does your health now limit you in any of the following activities?

3. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports:

- Limited a lot
- Limited a little
- Not at all limited

4. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf:

- Limited a lot
- Limited a little
- Not at all limited

5. Lifting or carrying groceries:

- Limited a lot
- Limited a little
- Not at all limited

6. Climbing several flights of stairs:

- Limited a lot
- Limited a little
- Not at all limited

7. Climbing one flight of stairs:

- Limited a lot
- Limited a little
- Not at all limited

8. Bending, kneeling, or stooping:

- Limited a lot
- Limited a little
- Not at all limited

9. Walking more than a mile:

- Limited a lot

- Limited a little
- Not at all limited

10. Walking several blocks:
    - Limited a lot
    - Limited a little
    - Not at all limited
11. Walking one block:
    - Limited a lot
    - Limited a little
    - Not at all limited
12. Bathing or dressing yourself:
    - Limited a lot
    - Limited a little
    - Not at all limited

Because of your physical health during the past 4 weeks, did you:

13. Cut down the amount of time you spent on work or other activities?
    - Yes
    - No
14. Accomplish less than you would like?
    - Yes
    - No
15. Limit in the kind of work or other activities?

- Yes
- No

16. Had difficulty performing the work or other activities (for example, it took extra effort)?
    - Yes
    - No

Because of emotional problems, during the past 4 weeks, did you:

17. Cut down the amount of time you spent on work or other activities?
    - Yes
    - No
18. Accomplish less than you would like?
    - Yes
    - No
19. Not do work or other activities as carefully as usual?
    - Yes
    - No
20. During the past 4 weeks, how much has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbours or groups?
    - Not at all
    - A little bit
    - Moderately

- Quite a bit
- Extremely

21. During the past 4 weeks, how much bodily pain have you had?
    - None
    - Very Mild
    - Mild
    - Moderate
    - Severe
    - Very Severe
22. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?
    - Not at all
    - A little bit
    - Moderately
    - Quite a bit
    - Extremely

These questions are about how you feel and how things have been with you during the last 4 weeks. For each question, please give the answer that comes closest to the way you have been feeling.

During the past 4 weeks, how much of the time:

23. Did you feel full of pep?
    - All of the time

- Most of the time
- A good bit of the time
- Some of the time
- A little bit of the time
- None of the time

24. Have you been a very nervous person?

- All of the time
- Most of the time
- A good bit of the time
- Some of the time
- A little bit of the time
- None of the time

25. Have you felt so down in the dumps that nothing could cheer you up?

- All of the time
- Most of the time
- A good bit of the time
- Some of the time
- A little bit of the time
- None of the time

26. Have you felt calm and peaceful?

- All of the time
- Most of the time
- A good bit of the time

- Some of the time
- A little bit of the time
- None of the time

27. Did you have a lot of energy?
    - All of the time
    - Most of the time
    - A good bit of the time
    - Some of the time
    - A little bit of the time
    - None of the time
28. Have you felt downhearted and blue?
    - All of the time
    - Most of the time
    - A good bit of the time
    - Some of the time
    - A little bit of the time
    - None of the time
29. Did you feel worn out?
    - All of the time
    - Most of the time
    - A good bit of the time
    - Some of the time
    - A little bit of the time

- None of the time

30. Have you been a happy person?
    - All of the time
    - Most of the time
    - A good bit of the time
    - Some of the time
    - A little bit of the time
    - None of the time
31. Did you feel tired?
    - All of the time
    - Most of the time
    - A good bit of the time
    - Some of the time
    - A little bit of the time
    - None of the time
32. During the past 4 weeks, how much of the time has your health limited your social activities (such as visiting with friends or close relatives)?
    - All of the time
    - Most of the time
    - A good bit of the time
    - Some of the time
    - A little bit of the time
    - None of the time

How true or false is each of the following statements for you?

33. To get sick a little easier than other people

- Definitely true
- Mostly true
- Don't know
- Mostly false
- Definitely false

34. As healthy as anybody I know

- Definitely true
- Mostly true
- Don't know
- Mostly false
- Definitely false

35. Expect health to get worse

- Definitely true
- Mostly true
- Don't know
- Mostly false
- Definitely false

36. Health is excellent

- Definitely true
- Mostly true

- Don't know
- Mostly false
- Definitely false