THE EFFECT OF PARTIAL MEASUREMENT INVARIANCE ON PREDICTION

by

BARBARA HAYES DONAHUE

(Under the Direction of Deborah Bandalos)

ABSTRACT

In education it is common to make judgments about the progress or achievement of different groups of students; it is also common to compare the mean scores of different groups and make predictions about individuals within groups. Thus, it is important to understand how underlying measurement differences, if any, affect such predictions and comparisons. Under partial measurement invariance, some model parameters are invariant while others are allowed to vary across groups. This allows the use of a scale in which there may be some difference in measurement between the groups, while still considering the overall comparison to be meaningful. The purpose of this study is to investigate the amount of partial measurement noninvariance that can be tolerated while still allowing for comparable predictions across groups. Specifically, varying degrees of size of factor loadings, model size, sample size, amount of partial measurement invariance, factor loading differences across groups, and differing levels of predictive influence across groups were examined from the standpoint of their effects on power and accuracy in prediction. The results for partial measurement invariance suggest that level of noninvariance and factor loading differences affect goodness of fit indices while the size of the

factor loading has more of an effect on parameter estimates and bias. It appears that model size affects all of the dependent variables presented here.

INDEX WORDS: Partial measurement invariance, Structural equation modeling

THE EFFECT OF PARTIAL MEASUREMENT INVARIANCE ON PREDICTION

by

BARBARA HAYES DONAHUE

B.A., Purdue University, 1992

M.A., University of Georgia, 1997

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2006

THE EFFECT OF PARTIAL MEASUREMENT INVARIANCE ON PREDICTION

by

BARBARA HAYES DONAHUE

Major Professor:	Deborah Bandalos
Committee:	Noel Gregg
	Robert Vandenberg
	Joseph Wisenbaker

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2006

DEDICATION

To the quarryman from Tipperary and his wife.

ACKNOWLEDGEMENTS

I have been blessed to have excellent professors who provided me with wonderful opportunities in educational research. Particularly, Steve Cramer, Jeri Benson, Noel Gregg, Knut Hagtvet and Steve Olejnik.

My committee members past and present: my chair, Debbi Bandalos, who got me by default and has been extremely helpful getting me to this point. Robert Vandenberg whose work in measurement invariance inspired me to explore the area further. Noel Gregg for giving me a context for bringing together measurement and learning disabilities. Joe Wisenbaker for being a calm, friendly face as I tried to find my way.

The staff at the College of Education and Graduate School has always been extremely helpful. They do not get enough credit for what they do. I am particularly grateful to Demetrius Smith, Laura Watson and Brenda Davis for their help and providing a good laugh when I needed it. I am also grateful to Stephanie Bales for her work on formatting my dissertation and reminding me I'm nearly there.

I had the fortune to meet Tom Benton as I was working on the study. He helped me by creating macros that made the job easier and listening as I went over (and over) my results. He remains a friendly face, quick wit and brilliant mind and I enjoy our chats about measurement and statistics immensely.

My father, Tom Donahue, told me of his PhD experiences and gave me sound advice that always came at the right time.

My mother, Mary Hayes Donahue, gave me amazing emotional support and a place I could work near the end. I cannot thank her enough. Now we can do genealogy! My second mom, Judie Farrow, was also instrumental in her support. I am grateful to them both.

Finally, I could not have gotten through these last three years without the help and support of Debbie Crouch and her family.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

In education it is common to make judgments about the progress or achievement of different groups of students; it is also common to compare the mean scores of different groups and make predictions about individuals within groups. Implicit in the comparison of any groups' scores is the assumption that the construct under consideration is measured the same way for individuals classified differently. Thus, it is important to understand how the underlying measurement differences, if any, affect such predictions and comparisons. The testing of measurement equivalence/invariance (ME/I) has recently been the focus of increased scholarly attention in the form of review articles (Steenkamp & Baumgartner, 1998; Vandenberg, 2002; Vandenberg & Lance, 2000) and empirical research (Cheung & Rensvold, 2002; Hutchinson & Young, 2003; Meade & Lautenschlager, 2004). The aim of the present study is to examine measurement invariance and its effect on prediction. This chapter will summarize relevant ME/I literature.

Measurement Invariance

A great deal of research has been conducted on the topic of measurement equivalence/invariance (ME/I) over the last 30 years. Broadly speaking, the concept of ME/I is concerned with the testing of whether the same items mean the same thing to members of different groups. In other words, the key question centers on the extent to which items measure the same underlying construct in categorically different populations. There are several levels of ME/I, which can be thought of on a continuum from theory to practice. These include conceptual

(e.g. Hui & Triandis, 1985), measurement and structural levels. The conceptual level refers to the extent to which a construct can be conceptualized in the same way within the groups of interest. The measurement level, also called the scale level, refers to the ME/I tests that examine how items work together in different groups. Finally, the structural level refers to the ME/I tests that examine how the whole structure of interest functions across groups, for example, examining latent construct mean differences across groups.

Conceptual Level

Within the conceptual level of invariance three subsections have been suggested (Hui & Triandis, 1985): conceptual equivalence, operationalization and item level equivalence. Conceptual equivalence refers to whether or not the construct under investigation is meaningful across groups (Hui & Triandis, 1985). This is particularly salient in cross-cultural research. In the operationalization subsection, researchers are interested in the extent to which the construct is being measured in the same way across groups. For example, is the construct of interest meaningful in the same way across different groups? Once conceptual equivalence and operationalization are established, item level equivalence would indicate that the instruments being used to measure the construct across groups have the same items and that these have the same meaning. If item level equivalence does not hold numerical comparison between groups is not possible (Hui & Triandis, 1985; Poortinga, 1989). According to Hui & Triandis (1985), "each item should mean the same thing to subjects from Culture A as it does to those from Culture B" (p. 134). This notion of equivalence does not allow for the possibility of partial measurement invariance. Partial measurement invariance will be given further treatment below.

Measurement Level

The measurement level of invariance examines aspects of the measurement model, specifically, the relationships between observed variables and their underlying latent constructs (Byrne, 1998; Vandenberg & Lance, 2000). As suggested by Jöreskog (1971), the first test of concern is an omnibus test for equality of covariance matrices across groups. If this test is not significant, there is no difference in the covariance matrices of the two groups and no need to proceed further within a multiple group context. If, however, the omnibus test is significant, it is incumbent upon the researcher to determine where the differences in the model occur across groups. The next logical step is called a test of configural invariance. When examining configural invariance (Vandenberg & Lance, 2000) the concern is with the pattern of fixed and free parameters and the equivalence of this pattern across groups. Metric invariance (Horn & McArdle, 1992; Vandenberg & Lance, 2000), the next test, checks for factor loading equivalence across groups. The fourth test of measurement level invariance is scalar invariance (Meredith, 1993; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000), which identifies whether the item intercepts are equivalent across groups. Error variance refers to the measurement error inherent in any measurement, and the final test examines the equivalence of the measurement error variances (and typically any measurement error covariances) across groups. The order of the preceding tests is not standardized, but researchers tend to follow the Jöreskog tradition of nested models (Vandenberg & Lance, 2000).

Structural Level

Once the measurement level tests have been performed and the researcher is satisfied that measurement invariance holds, it is possible to examine the structural invariance of the construct under consideration. Researchers traditionally perform tests of the equivalence of factor

variances and covariances as well as tests of latent means. These tests take place at the latent or unobserved level. Particularly, the test of factor variances and covariances would be used to examine the equivalence of latent factor variance and covariance across groups (Vandenberg & Lance, 2000). The test of latent means would be of interest if the researcher wanted to examine latent construct mean differences across groups. Traditionally, researchers would go no further in the series of invariance tests if a lack of invariance were noted at any level of the hierarchical nested model tests noted above. Please see Table 1 for a list of the measurement and structural level tests.

In sum, it is useful to conceptualize ME/I along a continuum from the more conceptual to the more applied measurement and structural levels. The covariance modeling methods used to detect ME/I are presented below.

Table 1

List of Invariance Tests as Presented in Chapter 1

1. An omnibus test for equality of covariance matrices across groups
2. A test of the pattern of fixed and free parameters and their equivalence across groups (configural invariance)
3. A test of factor loading equivalence across groups (metric invariance)
4. A test of item intercept equivalence across groups (scalar invariance)
5. A test of equivalence of the measurement error variances across groups
6. A test of equivalence of latent factor variance and covariance across groups
7. A test of equivalence of latent means

Detecting Measurement Invariance

The key question of ME/I centers on the extent to which items measure the same underlying construct in categorically different populations. The likelihood ratio test is the most frequently used to test ME/I because it is based on the χ^2 and has known distributional properties (Vandenberg & Lance, 2000). In the context of invariance testing, the likelihood ratio test is a

comparison of two models – one nested within the other – such that a significant likelihood ratio would indicate a significant decrement in model fit (Hutchinson, 2002).

Cheung and Rensvold (2002) suggested levels of change in other goodness of fit indices that might be useful in the detection of measurement invariance. Specifically, they examined the properties of 20 change (Δ) in goodness-of-fit indices under both an unconstrained model and one having measurement invariance constraints. They generated 48 models by varying the number of factors, the values of factor variance, the correlations between factors, the number of items per factor and the factor loadings. The effects of these variations were examined through analysis of variance and invariance tests. Based on the results of this study, they proposed cutoff values for ΔCFI, ΔGamma Hat, and ΔMcDonald's NCI. While these results provided a good first step, more work is necessary in order to further explicate the usefulness of these goodness of fit indices in a noninvariance context.

Partial Measurement Invariance

Muthén and Christoffersson (1981) first introduced the concept of partial measurement invariance. As noted previously, many researchers recommend terminating invariance testing at the first step in which noninvariance is found. That is, if factor loadings are found to be noninvariant, tests of intercepts, etc. would not be conducted. Under partial measurement invariance, if some model parameters (i.e., factor loadings, factor covariances, error variances) are found to be noninvariant while others are invariant these parameters are allowed to vary across groups, and testing is continued. This allows the use of a scale in which there may be some difference between the groups, while still considering the overall comparison to be meaningful. Byrne, Shavelson and Muthén (1989) put this notion into practice using self-concept data from high and low academically tracked high school adolescents. These data included

responses on the Self Description Questionnaire III, the Affective Perception Inventory, the Self-Esteem Scale, and the Self-Concept of Ability Scale. Byrne et al. hypothesized a confirmatory factor analysis (CFA) model in which self-concept responses could be explained by four factors: general self-concept, academic self-concept, English self-concept, and mathematics self-concept. However, the hypothesized model did not fit properly and showed invariance across groups therefore the researchers resorted to a series of exploratory models, based upon their substantive knowledge as well as the statistical analysis. Based on this information, they allowed for two parameters in the factor loading matrix and three parameters in the factor variance/covariance matrix to vary, but all other parameters were held invariant.

Despite this initial study, the number of parameters or items that can be allowed to vary while still maintaining overall measurement invariance remains unclear. Byrne et al. (1989) suggested that parameter estimates for at least one item needed to be completely invariant for comparisons across groups to be legitimate, while Reise, Widaman and Pugh (1993) argued that at least one-half of the scale items must be invariant for such comparisons to be meaningful. However, such decisions may differ depending on such things as the amount of invariance found as well as the purpose of the study undertaken. For example, no concrete evidence is available that indicates exactly how measurement invariance affects the use of a scale in terms of prediction. The next section addresses this issue.

Effects of Partial Measurement Invariance on Scale Use

The effects of partial ME/I on scale use has been examined in two ways. Either applied studies have investigated partial ME/I in terms of establishing the partial measurement invariance of an instrument (Byrne et al., 1989) or methodological researchers have investigated some of the factors thought to affect partial measurement invariance results (Hutchinson, 2002;

Hutchinson & Young, 2003; Kaplan, 1989; Kaplan & George, 1995). Methodological researchers have conceptualized the detection of partial measurement invariance in two ways, through the examination of the accuracy of selection of persons within groups (Millsap & Kwok, 2004), and through the power of the likelihood ratio test (Hutchinson, 2002; Hutchinson & Young, 2003; Kaplan, 1989; Kaplan & George, 1995).

Relative Accuracy of Selection

Millsap and Kwok (2004) examined the relative accuracy of selection of persons into groups based on scores on an instrument exhibiting partial measurement invariance. They examined four different indices of the accuracy of selection: proportion of persons selected per group, the success ratio, sensitivity (the proportion of true positives divided by the proportion selected) and specificity (the proportion of true negatives divided by the proportion of true negatives plus the proportion of false positives). These authors varied the selection percentage, the number of items, degree of invariance of factor loadings, and size of the sum of unique variances. They found that fewer people were selected into the focal group in instances of noninvariance. Additionally, it appears that greater levels of noninvariance might also be associated with smaller success ratios across both groups. The authors suggest that changes of less than 0.05 between partially and fully invariant models in sensitivity, success ratios, and specificity are probably not meaningful. However, the authors also note that such a determination should be based on the use of the scale and the types of error (false positives/false negatives) that can be tolerated within the context of selection.

Power of the Likelihood Ratio Test

As with measurement invariance, the likelihood ratio test is the most frequently used test of partial ME/I because it is based on the χ^2 and has known distributional properties. Power is

operationally defined as the proportion of true rejections of the likelihood ratio test across the number of replications within any given level of the research design. In 1989, Kaplan used a small population study to investigate the effect of specification error on the power of the likelihood ratio test in a partial measurement invariance context (Kaplan, 1989). He examined a six-variable, two-factor model, in which group one had simple structure; that is, each item loaded on one factor; and group two had one complex loading. He found that the size of a misspecified parameter and its relationship to the other parameters in the model affected the power of the likelihood ratio test to detect partial measurement invariance.

Building on the earlier work, Kaplan and George (1995) examined the power of the likelihood ratio test associated with the test of factor mean differences when the assumption of factorial invariance is untenable. In particular, Kaplan and George investigated the size of factor mean differences, sample size difference, and degree of noninvariance. They found that the power of the likelihood ratio test for detecting factor mean differences is relatively robust to partial measurement invariance, but is affected by unequal sample size across groups (Kaplan & George, 1995).

Hutchinson (2002) conducted a Monte Carlo study using different levels of model size, sample size, number of noninvariant items, and size of between-group loading differences to investigate power of the likelihood ratio test to detect factorial noninvariance. She reported that the likelihood ratio test might not be robust to the location of noninvariant items or the amount of noninvariance, which could lead to erroneous decisions about the comparability of constructs across groups. As a result of this work, Hutchinson and Young (2003) examined how the placement of noninvariant items across factors affected power and found that noninvariant items that are dispersed across factors increase the power of the likelihood ratio test.

Prediction

To date, most of the research studies on ME/I have focused on statistical significance, rather than practical significance. For example, measurement scales are often used to predict outcomes for individuals in different groups. However, if these scales are not invariant across groups, the accuracy of predictions may be affected. Byrne et al. (1989) and Kaplan & George (1995) examined factor mean differences across groups, but no studies to date have studied how much absolute invariance is necessary for prediction of group performance to remain valid.

Summary

It is common in education to make judgments about the progress or achievement of different groups of students; it is also common to make predictions for different groups of students. In order for these comparisons to be valid, the different groups must conceptualize the scales and items being used in the same way. It is critical that educational researchers test the assumption of invariance before making judgments. The concept of invariance may be thought of along a continuum from theoretical concerns to the actual testing of measurement and structural level model parameters. The tests designed to determine ME/I within model parameters were developed by Jöreskog (1971) and are still in use today. The likelihood ratio test is most often used to test ME/I because it is based on the χ^2 and has known distributional properties. The concept of partial ME/I allows some model parameters to be noninvariant while others are invariant. The effect of partial ME/I on the predictive validity of measurement scales is still unknown. The next chapter will provide further treatment of ME/I and partial ME/I literature and a rationale for the proposed study.

CHAPTER 2

REVIEW OF LITERATURE

A great deal of research has been conducted on the topic of measurement equivalence/invariance (ME/I) over the last 30 years. Broadly speaking, the concept of ME/I is concerned with testing whether the same items mean the same thing to members of different groups. In other words, the key question centers on the extent to which items measure the same underlying construct in categorically different populations. The aim of this review is to provide a historical perspective on ME/I and partial ME/I. Following the historical perspective; the factors thought to affect partial ME/I will be presented with relevant research. Finally, the purpose and research questions will be presented.

A Historical Perspective of Measurement Invariance

Kaplan (1995) noted that Jöreskog originally developed the methods for conducting multiple-group confirmatory factor analysis in his oft-cited work of 1971. In fact, in the comprehensive review by Vandenberg and Lance (2000), Jöreskog (1971) is the earliest article addressing the practice of evaluating invariance/equivalence in a systematic way. Jöreskog (1971) indicated that it would be prudent to proceed through the tests of multi-group equivalence in a certain order, as any one test depends on the results of the previous test to move forward. He began by testing the hypothesis of equality of covariance matrices. This test compares the observed covariance matrices for two or more groups. If this test is not significant, there is no difference in the covariance matrices of the two groups and no need to proceed further within the multiple group context. This would indicate that the covariance matrices of the groups under

consideration were equal and, essentially, there was no measurement difference between the groups. This is not likely to happen and this test of equality of covariance matrices is seen by some as a formality (see Byrne, 1998; Byrne, et al. 1989).

If the initial hypothesis of equivalent covariance matrices is found to be untenable, Jöreskog (1971) recommended a set of tests to pinpoint model parameter noninvariance across groups. The first hypothesis he suggested is "the hypothesis of number of common factors" (Jöreskog, 1971, p. 419). He suggested that this could be done by conducting separate analyses on each group using the same number of common factors for each group. If the hypothesis of equal numbers of factors for each group is reasonable, the researcher proceeds to a set of hierarchically nested tests that are evaluated by computing a change in the χ^2 statistic ($\Delta\chi^2$) such that each hypothesized model is compared to some less restrictive model. If the $\Delta\chi^2$ is considered significant, the hypothesis that the restrictions hold is affirmed. The first of these is "the hypothesis of an invariant factor pattern" (Jöreskog, 1971, p. 420). The test of invariant factor patterns is a test that determines whether the pattern of factor loadings is equivalent across groups. If this test is not rejected, the equivalence of the factor loadings values is conducted. This test is known as the test of metric invariance (Horn & McArdle, 1992). In LISREL notation this is referred to as the equivalence of the λ (lambda) matrices across groups. Should the hypothesis of equivalent factor loadings be considered reasonable, Jöreskog suggested the next test in the sequence should be a test of uniquenesses, or testing that the measurement error variances associated with the observed variables are equivalent across groups. Should that hypothesis seem reasonable, he suggested the test of equivalent latent factor variances-covariances. As he presented these hypotheses for consideration, he indicated that the hypothesis prior to the current

one being tested is embedded in the current hypothesis (nested models). In other words, each hypothesis builds on what has been tested previously.

In 1992, Horn and McArdle introduced two new terms to the ME/I lexicon that are still in use (Vandenberg, 2002; Vandenberg & Lance, 2000): configural and metric invariance. Horn and McArdle (1992) took Jöreskog's (1971) notion of equal numbers of common factors and called this a test of configural invariance; they also called it a weak test of measurement invariance. Additionally, they called the test of equivalent factor loadings metric invariance and indicated this is a strong test of measurement invariance.

Meredith (1993) provided a mathematical treatment of the principles of weak, strong and strict measurement invariance. In particular, he showed that if the level of the underlying construct was taken into account, item variances across groups should be the same. This he called weak measurement invariance. Strong measurement invariance was defined as invariant factor loadings and item intercepts. Strict measurement invariance was the same as strong measurement invariance with the addition of the equivalence of measurement error variances. Strict and strong factorial invariance require analysis of both item intercepts and factor means.

Steenkamp and Baumgartner (1998) tried to clarify the terminology of ME/I. They acknowledged the different forms of measurement invariance that appear in the literature and give very clear definitions of them. Configural invariance is indicated by a common factor model across groups. Metric invariance is indicated by common factor loading values across groups. They also introduced scalar invariance, which they defined as the test of equality of measurement intercepts. If a measure can be considered scalar invariant, the differences in observed variable means across groups are due to differences in the means of the underlying constructs. They also introduced the concept of partial configural invariance, that is, noninvariant

common factor structure between groups. They suggested that if partial metric invariance is tenable, further partially invariant parameters could be tested as well. They also proposed that one could make meaningful factor mean comparisons as long as at least one item besides the marker item is invariant. A marker item is an item factor loading set to one in order to identify the covariance model and give the factor a scale. However, they also suggest a majority of factor loadings and intercepts should be invariant for the sake of the reliability of the estimated latent means.

Jöreskog (1971) provided a series of nested tests to examine ME/I across various structural parameters. Since then several authors have expounded on his ideas and made further terminology suggestions. Partial measurement invariance takes Jöreskog's (1971) ideas further in that it allows one to consider measures with some noninvariant model parameters. Below is a historical perspective of partial measurement invariance.

A Historical Perspective of Partial Measurement Invariance

Within partial measurement invariance some model parameters (i.e., factor loadings, factor covariances, error variances) are considered to be noninvariant while others are invariant, thereby allowing one to use a scale in which there may be some differences in measurement parameters between the groups, while still considering the overall comparison to be meaningful. As noted by Byrne et al. (1989) and Steenkamp and Baumgartner (1998), it is possible to conceptualize noninvariance in multiple model parameters. This review will focus on literature examining partial metric invariance, which is noninvariance in the item factor loadings.

Byrne et al. (1989) tested for ME/I in factor covariances and mean structures in multidimensional self-concept data across low and high academically tracked high school students. These data included responses on the Self Description Questionnaire III, the Affective

Perception Inventory, the Self-Esteem Scale, and the Self-Concept of Ability Scale. Byrne et al. hypothesized a simple structure confirmatory factor analysis (CFA) model in which self-concept responses could be explained by four factors: general self-concept, academic self-concept, English self-concept, and mathematics self-concept. She and her colleagues also hypothesized that the four self-concept factors were correlated and the measurement error variances were uncorrelated. To measure model fit they used the χ^2 statistic, the goodness-of fit index, the root mean square residual, the normed fit index and the non-normed fit index as well as their collective expertise and knowledge of the substantive area. They found that the hypothesized model did not fit properly and therefore resorted to a series of exploratory models, based upon their substantive knowledge as well as the statistical analysis. Based on this information, they identified noninvariance in two parameter estimates in the factor loading matrix and in three parameter estimates in the factor variance/covariance matrix. Specifically, they found that the low track group required one factor loading and three error variance terms to be free, while the high track group required a different factor loading and the same three error variance terms to be free. They admitted that the statistical analysis was less than optimal, but they based their conclusion on several considerations:

- All of the initially hypothesized factor loadings, error variances and factor variances and covariances were statistically significant across groups;
- The two free factor loadings were substantial for both groups;
- Coefficient of determination values were extremely high in both groups relating the assessments to the latent self-concept construct;
- The three free error variances did not appear to adversely affect other model parameters and seemed to represent nonrandom method error.

Kaplan (1989) studied the power of the likelihood ratio test in partial measurement invariance within the context of a population study. In particular, he was interested in the effect a specification error would have on the power of the likelihood ratio test. He defined specification error as "the incorrect restriction of a parameter (e.g., a cross loading) in one group. That is, it is assumed that one group possesses a more complex factor structure which is unknown to the investigator" (Kaplan, 1989 p. 580). The model studied was a six-item, two-factor model. In group one the model was set to simple structure, while in group two there was one complex factor loading. The values for the complex loading were varied by 0.1, 0.2, 0.3 and 0.4. The other item factor loadings for group two were 0.5, 0.6 and 0.8. The item factor loadings for group one were all set to 0.7. He found that power was highest when the specification error was 0.4 and the free loadings were 0.8. He also found that the larger the correlation between the misspecified parameter and the remaining free parameters, the higher the power of the likelihood ratio test to detect partial measurement invariance.

In 1995, Kaplan and George "examined the power associated with the test of factor mean differences when the assumption of factorial invariance is violated" (Kaplan & George, 1995; p. 101) within the context of a population study. The authors wanted to know whether the test of latent mean differences was still valid if metric invariance was untenable. They varied model size, size of loadings and sample size in both noninvariant and partially noninvariant models. The levels of model size were a six-item, two-factor model and a 12-item, two-factor model. The size of loadings varied by group. For group one the size of the item factor loadings varied between 0.5, 0.6, and 0.7, and for group two all of the item factor loadings were set to 0.7. Finally, sample size was varied across groups. The total sample size was set to 1000. In one condition sample size was equal across groups while in the remaining two conditions the

difference in sample size between groups was 500 and 800. Power to detect latent mean differences was assessed by the Wald test. They found that increased levels of noninvariance decreased power of the Wald test and also that unequal sample sizes between groups decreased the power of the Wald test to detect latent mean differences in multiple group confirmatory factor analysis.

Hutchinson (2002) studied the power of the likelihood ratio test to detect metric invariance. She conducted a Monte Carlo study using different levels of model size, sample size, level of noninvariance, and magnitude of between-group loading differences. She examined a six-item, one-factor model and a 12-item, two-factor model. The groups had sample sizes of either 500 or 200 per group. In the complete noninvariance case, group one factor loadings ranged between 0.6 and 0.8 while group two factor loadings differed from those of group one by 0.1 to 0.4. In the partial noninvariance case, group one factor loadings ranged from 0.6 to 0.8 while in group two one-half of the factor loadings were between 0.1 and 0.4 lower than the corresponding group one factor loadings. She found that power to detect complete metric noninvariance was very low. Power was higher in the one factor model in the partial noninvariance case than the two factor model. She also found that the likelihood ratio test tended to "incorrectly suggest noninvariance at subsequent steps in the invariance testing process" (Hutchinson, 2002 p. 10). As a result of this she conducted post hoc tests on the two factor model to determine whether the pattern of noninvariance was the cause. She found by evenly distributing the noninvariant factor loadings across the two factors (as opposed to having all noninvariant loadings on one factor) that power did indeed increase. She therefore summarized that the likelihood ratio test might not be robust to the location of noninvariant items or the

amount of noninvariance, leading to erroneous decisions about the comparability of constructs across groups.

In a follow-up study, Hutchinson and Young (2003) examined power of the likelihood ratio test with larger models, varying group sample sizes and different proportions of noninvariant items to determine whether dispersing the noninvariance across factors would improve power. Again, they found that power was low in the complete noninvariance conditions and improved considerably in the partial invariance conditions. They found larger models (3 or 4 factor models) exhibited higher power than smaller models (one factor model). They also found that factor loading differences seemed to have a pronounced effect on power, with power being very high as the factor loading difference was larger. They also found that when they dispersed noninvariant items across factors power was much higher under most conditions of partial noninvariance in larger models.

Meade and Lautenschlager (2004) examined the sensitivity of invariance tests for establishing ME/I when it was known not to exist. They conducted a Monte Carlo study using different levels of sample size, model size, and number of items noninvariant. They included three levels of sample size, 150, 500 and 1000. The levels of model size were a six-item, one-factor model and a 12-item, one-factor model. The number of items noninvariant was varied between two, four, and in the case of the 12-item model, eight items were invariant. They found that "detection rates for both the overall omnibus test and specific tests of factor loading differences displayed a ceiling effect that…makes comparisons of the efficacy of different conditions for large sample sizes impossible" (Meade & Lautenschlager, 2004, p. 66). As a result, they focused their discussion on the conditions with a sample size of 150. They found that when more items were simulated to differ across groups the nested model χ^2 test was better able

to detect a lack of invariance. They also found that a mixed pattern of factor loadings (some simulated higher, some simulated to be lower than those in group one) in group two had higher detection rates than when factor loadings were uniformly lower for group two than group one.

Millsap and Kwok (2004) examined another aspect of partial invariance. They were interested in the effect partial measurement invariance had on the accuracy of selection. Specifically, they wanted to know the consequences of using a measure for selection purposes when partial metric invariance was present in that measure. They suggested four quantities that would aid the decision-making process, proportion selected into each group, success ratio, sensitivity and specificity. These four quantities are made up of the proportion of true positive, false negative, true negative and false negative. Proportion selected into each group is the proportion of true positives plus the proportion of false positives. Success ratio is defined as the proportion of true positives divided by the proportion selected. Sensitivity is defined as the proportion of true positives divided by the proportion of true positive plus the proportion of false negatives. Finally, specificity is defined as the proportion of true negatives divided by the proportion of true negatives plus the proportion of false positives. They manipulated aspects of the selection process and the factor structure. They used two selection percentages corresponding to the 75^{th}-percentile and the 90^{th}-percentile, four levels of the number of items in the observed measure (4, 8, 12 or 16 items), degree of metric invariance as defined by the percentage of loadings invariant (100%, 75%, 50%, 25% or 0%), and high versus low communality levels. They found that the higher communality level led to higher proportions of individuals selected in the reference group, higher success ratios, sensitivity, and specificity for both groups. They also found noninvariance led to fewer focal group members being selected which leads to lower sensitivity and a corresponding increase in sensitivity in the reference group. They suggest that

across levels of noninvariance, changes in success ratios, sensitivity and specificity less than 0.05 were unlikely to be meaningful. However, the authors also note that such a determination should be based on the use of the scale and the types of error (false postives/false negatives) that can be tolerated within the context of selection.

This section has presented an overview of the history of research on partial measurement invariance, the notion that some model parameters can be considered noninvariant while others are invariant. Byrne et al. (1989) are credited with the initial implementation of the concept of partial measurement invariance. Kaplan and other researchers translated this initial work into empirical research within the confirmatory factor analytic context. Finally, recent advances in the field have also examined the extent to which partial ME/I affects group selection into referent or focal groups. In sum, the factors that emerge as most important in the partial measurement invariance literature are sample size, model size, and level of noninvariance. The next section will examine these factors in more detail.

Factors that Affect Measurement Invariance

Sample Size

It is widely known that one must have adequate sample size in order to make stable judgments about model fit in confirmatory factor analysis. Sample size has been an independent variable in several studies of ME/I (Hutchinson & Young, 2003; Hutchinson, 2002; Kaplan & George, 1995; Meade & Lautenschlager, 2004). Meade and Lautenschlager (2004) used sample sizes ranging from 150 to 1000. They noted a ceiling effect (nearly perfect detection rate) in detecting differences of group item factor loadings with sample sizes of 500 and 1000. Hutchinson (2003) used sample sizes of 200 and 500 per group. She found an increase in power in the larger sample sizes, as one might expect. This suggests that in order to examine practical

differences in various study conditions it is necessary to use sample sizes of less than 1000 in order to avoid a potential ceiling effect where differences are always detected.

Kaplan and George (1995) examined unequal sample sizes across the two groups. They found power to be much lower when the samples sizes were extremely unequal (e.g., n_1=100, n_2=900). The use of unequal sample sizes across groups is interesting because in many practical cases the focal group is smaller than the reference group. For example, there is typically a great group size discrepancy in the case of comparisons of persons with and without learning disabilities.

Model Size/Complexity

ME/I simulation studies to date have included models of similar size. Meade and Lautenschlager (2004) used two models, each of which had one factor with six and twelve items, respectively. Hutchinson and Young (2003) used two models, one of which had one factor and six items, the other of which had two factors and twelve items. Kaplan and George (1995) also used two models, which had two factors with six and twelve items, respectively. However, nearly all of the results presented by these researchers centered on independent variables other than model size and complexity, though Hutchinson and Young (2003) did recommend future research on models of various sizes.

Number of Loadings Variant/Invariant

Researchers have systematically varied the amount of noninvariance in studies of ME/I in two ways; by examining the magnitude of the factor loading differences (Hutchinson & Young, 2003; Kaplan, 1989; Kaplan & George, 1995) and by varying the number of factor loadings invariant (Hutchinson & Young, 2003; Kaplan & George, 1995; Meade & Lautenschlager, 2004). Researchers have investigated both partial and complete noninvariance scenarios with

mixed results (Hutchinson & Young, 2003; Kaplan & George, 1995; Meade & Lautenschlager, 2004). Meade and Lautenschlager (2004) found higher power for situations in which more loadings differed between groups. Hutchinson and Young (2003) found low power for the complete noninvariance case regardless of level of factor loading difference across groups. These researchers found somewhat higher power in the partial measurement invariance case, particularly with a larger magnitude of factor loading differences. Kaplan and George (1995) found that power differences had more to do with the inequality of the group mean differences than the level of noninvariance.

Summary

Sample size, model size and levels of invariance are factors that affect measurement invariance. As mentioned earlier, there have been researchers who have suggested various acceptable levels of noninvariance, but most of the research studies on partial ME/I have focused on statistical significance, rather than practical significance. If measurement scales are not invariant across groups, the accuracy of group predictions may be affected. Byrne et al. (1989) and Kaplan and George (1995) examined factor mean differences across groups, but it is still unclear how much absolute invariance is necessary for prediction of group performance to remain valid.

Purpose

In most partial ME/I studies the interest has been in establishing the power of the likelihood ratio test to detect partial measurement invariance. The effects of partial measurement invariance on an instrument used for prediction and/or group comparisons is inconclusive and currently the ME/I literature offers little guidance in this area. The purpose of this study is therefore to investigate the amount of partial measurement noninvariance that can be tolerated

while still allowing for comparable predictions across groups. Specifically, varying degrees of size of factor loadings, model size, sample size, amount of partial measurement invariance, factor loading differences across groups, and differing levels of predicative influence across groups will be examined from the standpoint of their effects on power, accuracy in prediction, and parameter estimate bias.

Examining the effect of partial measurement noninvariance on predictive influence will contribute to a better understanding of the practical consequences of measurement differences across groups. Once the consequences are better understood, measurement professionals will be more prepared to determine tolerable levels of partial noninvariance for different situations.

Research Questions

A Monte Carlo simulation will be designed to answer the following research questions:

1. How will group sample size in partial metric invariance influence accuracy of predictive influence, power and model fit?
2. How will model size in partial metric invariance influence accuracy of predictive influence, power and model fit?
3. How will factor loading size in partial metric invariance influence accuracy of predictive influence, power and model fit?
4. How will the number of invariant factor loadings influence accuracy of predictive influence, power and model fit?
5. How will the magnitude of loading differences across groups influence accuracy of predictive influence, power and model fit?

6. How will equal versus unequal levels of predictive influence across groups in partial measurement invariance influence accuracy of predictive influence, power and model fit?

Predictive influence is operationally defined as the regression coefficient (path) between a latent exogenous variable and an observed endogenous variable. To date, the effect of partial measurement invariance on prediction has not been examined.

Sample size is a frequently studied factor of interest in partial measurement invariance. It is widely known that researchers must have adequate sample size in order to make accurate judgments about model fit in confirmatory factor analysis, however what is adequate will depend on factors such as model complexity and potentially the amount of partial measurement invariance encountered in a scale. The work of Meade and Lautenschlager (2004) suggests that methodological researchers also need to be careful of having too large a sample size. It is possible that having too large a sample size while trying to 'detect' partial measurement invariance in a simulation study might lead to perfect power and detection rates thereby masking the effects of other independent variables. In addition, unequal group sample sizes (Kaplan & George, 1995) will likely reduce the power to detect partial measurement invariance. Therefore, it is expected that larger but equivalent group sample sizes will provide more accurate estimates of prediction than unequal group sample sizes.

ME/I simulation studies to date have included models of similar size. Nearly all of the results presented by previous researchers have centered on independent variables other than model size and complexity.

Based in part on the findings of Kaplan (1989) it is expected that larger factor loadings will be provide more accurate estimates of prediction than smaller factor loadings. However,

Meade and Lautenschlager (2004) found that a mixed pattern of factor loadings (some simulated higher, some simulated to be lower than those in group one) in group two resulted in higher detection rates than when factor loadings were uniformly lower for group two than group one.

Kaplan and George (1995) found that increased levels of loading noninvariance decreased power of the Wald test to detect differences in latent means. However, Meade and Lautenschlager (2004) found that when more items were simulated to differ across groups the nested model χ^2 test was better able to detect a lack of invariance in factor loadings. Hutchinson and Young (2003) found that power to detect complete metric noninvariance was very low, which tends to agree with the findings of Kaplan and George (1995). However, it is expected that more noninvariance will lead to greater bias in the estimate of predictive influence.

CHAPTER 3

METHODS AND PROCEDURES

The purpose of this study was to investigate the amount of partial measurement noninvariance that can be tolerated while still allowing for comparable predictions across groups. Specifically, a Monte Carlo simulation study varying size of factor loadings, model size, sample size, amount of partial measurement invariance, factor loading differences across groups, and differing levels of predictive influence across groups was conducted to examine the effects of these variables on power and accuracy in prediction. Predictive influence was operationally defined as the magnitude of the regression coefficient (path) between a latent exogenous variable to an observed endogenous variable. A structural equation model with a single latent exogenous factor with a varying number of indicators with a path to an observed endogenous variable was the basic model under investigation (Figure 1).

Independent Variables

The six independent variables are: size of the factor loadings (3 levels); model size (3 levels); number of items non-invariant (3 levels); sample size (3 levels); size of factor loading difference across groups (3 levels) and different levels of prediction across groups (2 levels) for a total of 486 cells of interest. Each cell represents the intersection of the six independent variables, for a completely crossed design. There were 250 replications per cell and all data were generated as multivariate normal. In order to make comparisons between invariant models and noninvariant models, 54 additional baseline cells were created to correspond to the independent variables that did not create the partial metric invariance. These independent variables were: size

of factor loading, model size and differing levels of prediction. For these cells all items were simulated to be noninvariant across groups.

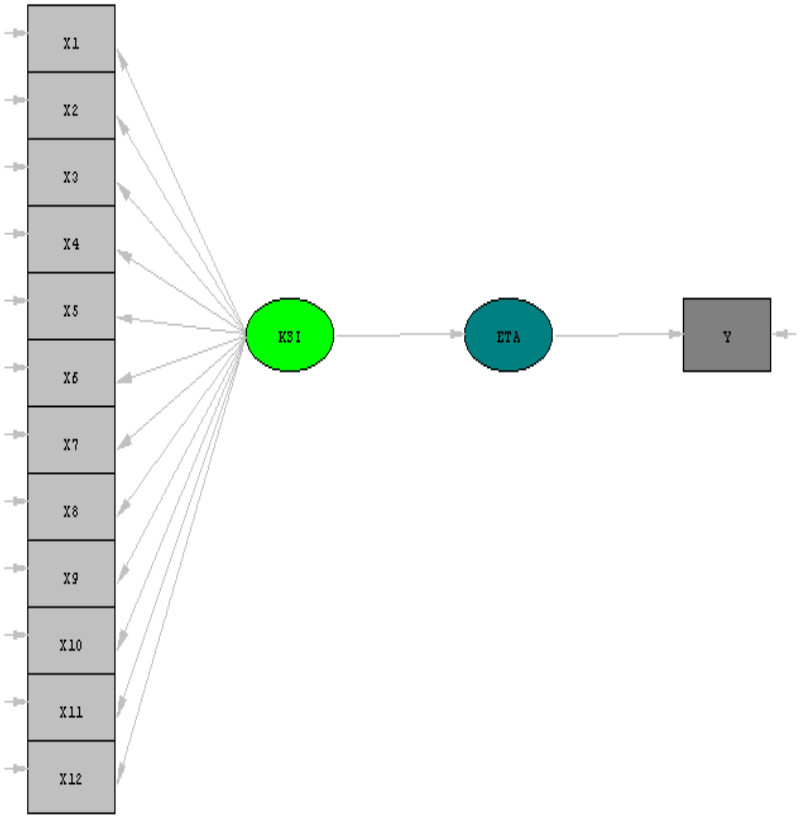

Figure 1. Example of Model Used in Study (MS2).

The three levels of size of factor loading are defined as SFL_1, SFL_2, and SFL_3. SFL_1 included loadings for both groups that were between 0.3 and 0.7. SFL_2 included loadings for both groups that were between 0.5 and 0.9. SFL_3 included a mixture of loadings between 0.3 and 0.9 across both groups. These factor loadings are somewhat similar to those used by Hutchinson and Young (2003), Kaplan and George (1995), and Kaplan (1989).

The three levels of model size are defined as MS_1, MS_2, and MS_3. In MS_1 there was one six-item exogenous factor, MS_2 had one twelve-item exogenous factor and MS_3 had one 18-item exogenous factor. MS_1 has been used in a previous study (Hutchinson & Young, 2003) without the regression path. MS_2 and MS_3 are included to further examine the problematic nature of larger models (Cheung & Rensvold, 2002). As noted above, each model had a concomitant outcome regression path.

The three levels of invariance are defined as II_1, II_2, and II_3. II_1 had one item per model set to be non-invariant, II_2 had one-third of the items set to be non-invariant, and II_3 had two-thirds of the items non-invariant. For example, in the cells that examined MS_2 and II_2, there were four noninvariant items as four is one-third of twelve, the number of items in that particular model.

The three levels of sample size are defined as SS_1, SS_2, and SS_3. SS_1 had 200 per group, SS_2 had 450 per group, and SS_3 had 600 in the referent group and 300 in the focal group. These levels are in keeping with previous research (Cheung & Rensvold, 2002; Hutchinson & Young, 2003) and should provide stable parameter estimates, while maintaining practicality.

The three levels of factor loading difference are defined as FLD_1, FLD_2, and FLD_3. FLD_1 had 0.2 factor loading differences between noninvariant items across the groups, FLD_2 had 0.3

factor loading differences, and FLD_3 had 0.4 factor loading differences across groups on noninvariant items.

The two levels of predictive influence across groups are defined as GAM_1 and GAM_2. In GAM_1, the level of predictive influence across groups was equivalent, that is, the regression path was equal in the population across groups (0.5 for both groups). In the GAM_2 level, the level of predictive influence across groups was not equivalent; the regression path was different by 0.2 in the population across groups, such that the gamma population value for group one was 0.5 and the gamma population value for group two was 0.3.

Dependent Variables

Values of goodness of fit indices, power of the likelihood ratio test and bias in the estimates of the regression path of interest were examined as dependent variables. The likelihood ratio test was calculated as the difference between two likelihood ratio values in two nested models; one in which parameters are constrained, one in which those same parameters are allowed to be freely estimated. Power was operationally defined as the proportion of true rejections of the likelihood ratio test across the number of replications within any given level of the research design. Bias was operationally defined as the average deviation of a sample value from the concomitant population value divided by the population value (Bandalos, in press).

Data Generation

The data for the study was generated within LISREL 8.72 and PRELIS 2.72 (Jöreskog & Sörbom, 1996a; 1996b; 1996c). LISREL was used to generate population covariance matrices for the groups, from which the sample covariance matrices were generated by PRELIS. The data were analyzed using LISREL within a multiple-group CFA framework. The goodness of fit

indices and parameter estimates for each cell from each invariance analysis were stored for further analyses in SPSS 11.5.

In LISREL the invariance tests were performed in the following order:

1. The tenability of equality of the factor structure of the generated samples.
2. The tenability of invariant factor loadings across groups.
3. The tenability of invariant latent factor variance.
4. The tenability of invariant measurement error variances across groups.
5. The tenability of the invariance of the path value between the latent exogenous variable and the latent endogenous variable.

Data Analysis

A completely crossed 3x3x3x3x3x2 design was utilized in this study. Power of the likelihood ratio test and bias were first examined with descriptive statistics. Goodness of fit indices, and bias were also analyzed using separate analyses of variance. To date, partial measurement invariance has been analyzed descriptively with the examination of power of the likelihood ratio test. As a result, the variables that affect partial measurement invariance are relatively well known so the next step is to examine these variables using inferential statistics. The use of analysis of variance indicates a certain level of theoretical development in the factors (Keppel, 1991) that affect partial measurement invariance as evidenced also by the research presented in Chapter 2.

CHAPTER 4

RESULTS

The purpose of this study was to investigate the amount of partial measurement noninvariance that can be tolerated while still allowing for equally accurate predictions across groups. In particular, the research questions were: (1) how will group sample size in partial metric invariance influence accuracy of predictive influence, power and model fit, (2) how will model size in partial metric invariance influence accuracy of predictive influence, power and model fit, (3) how will factor loading size in partial metric invariance influence accuracy of predictive influence, power and model fit, (4) how will the number of invariant factor loadings influence accuracy of predictive influence, power and model fit, (5) how will the magnitude of loading differences across groups influence accuracy of predictive influence, power and model fit, (6) how will equal versus unequal levels of predictive influence across groups in partial measurement invariance influence accuracy of predictive influence, power and model fit.

A Monte Carlo simulation examined varying degrees of size of factor loadings, model size, sample size, amount of partial measurement invariance, factor loading differences across groups, and differing levels of predictive influence across groups from the standpoint of their effects on power and accuracy in prediction (Table 2). Predictive influence was operationally defined as the regression coefficient (path) between a latent exogenous variable to an observed endogenous variable. There were 250 replications per cell and all data were generated as multivariate normal.

The results section will be divided into two parts: baseline condition and partial metric invariance. The baseline condition section provides descriptive statistics on additional cells created to provide information on the values of the dependent variables under invariant conditions across groups. These tables are based on 250 replications of the test of metric invariance.

Table 2

Design of Study

Independent Variables						
Level	SFL	MS	II	SS	FLD	GAM
0			No difference		No difference	
1	Between 0.7 and 0.3 mean SFL_1=0.55	1 factor – 6 item	1 item noninvariant	200/group	0.2 difference	Equal gamma in population (0.5 in both groups)
2	Between 0.9 and 0.5 mean SFL_2=0.76	1 factor – 12 item	1/3 items noninvariant	450/group	0.3 difference	Unequal gamma in population (0.5 in group1, 0.3 in group 2
3	Between 0.9 and 0.3 mean SFL_3=0.65	1 factor – 18 items	2/3 items noninvariant	600/g_1 300/g_2	0.4 difference	

The partial metric invariance section will provide the results relevant to the research questions. Further, to save space in the main document, tables of descriptive statistics for the dependent variables are provided in Appendix A. These tables are also based on 250 replications of the test of metric invariance.

Baseline Condition

The baseline condition examined the dependent variables under completely invariant conditions. Basic descriptive statistics collapsed across all invariant independent variables are

presented for the tests of factor structure (configural invariance) and factor loadings (metric invariance) on the following dependent variables: bias in the group 1 gamma parameter estimate, bias in the group 2 gamma parameter estimate, root mean square error of approximation (RMSEA), normed fit index (NFI), non-normed fit index (NNFI), comparative fit index (CFI), minimum fit function χ^2 statistic (χ^2), the likelihood ratio ($\Delta\chi^2$) between equality of factor structure and equality of factor loading, and the difference between CFI (ΔCFI) in configural invariance and metric invariance (Table 3a and 3b). Basic descriptive statistics for each independent variable are also included (Tables 4-7).

As can be seen in Tables 3a and 3b, there is not a great deal of difference between values of the parameter estimates and fit indices from the tests of factor structure and of factor loadings in the baseline condition. Power estimates of the likelihood ratio test show between two and eight percent of cell replications reject the null hypothesis of no significant difference in the χ^2 across nested models, indicating the relative Type I error rate one might expect in noninvariant models.

Across size of factor loading, model size, sample size and differing levels of prediction (Tables 4-7), the bias estimate for the gamma parameter estimates for group one and group two indicate either no bias or a slight underestimation of this parameter. RMSEA, NFI, NNFI, CFI and ΔCFI all appear to be stable across levels of size of factor loading, model size, sample size and differing levels of predictive influence for the baseline condition. As would be expected, the values of the likelihood ratio ($\Delta\chi^2$) test presented in Table 5 increase as model size increases. In sum, when conditions are such that invariance holds, bias and fit indices give favorable results.

Table 3a

Test of Equality of Factor Structure Descriptive Statistics Collapsed Across All Factors for Baseline Cells (n_{cells}=54)

Factor Structure	Statistics		
Dependent variables*	Mean	Median	SD
biasg1g	-0.01	-0.01	0.09
biasg2g	-0.01	-0.01	0.12
RMSEA	0.01	0.00	0.01
NFI	0.99	0.99	0.01
NNFI	1.00	1.00	0.00
CFI	1.00	1.00	0.00
χ^2	157.53	131.70	118.44

Note: * biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, χ^2 degrees of freedom were 28, 130, 304 depending on model size.

Table 3b

Test of Equality of Factor Loadings Descriptive Statistics Collapsed Across All Factors for Baseline Cells (n_{cells}=54)

Factor Loading	Statistics		
Dependent variables*	Mean	Median	SD
biasg1g	-0.01	-0.01	0.09
biasg2g	-0.01	-0.01	0.11
RMSEA	0.01	0.00	0.01
NFI	0.99	0.99	0.01
NNFI	1.00	1.00	0.00
CFI	1.00	1.00	0.00
χ^2	168.68	142.81	123.53
$\Delta\chi^2$	11.15	10.00	8.62
ΔCFI	0.00	0.00	0.00

Note: * biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, $\Delta\chi^2$ = likelihood ratio χ^2 between equality of factor structure and equality of factor loading, χ^2 degrees of freedom were 33, 141, or 321 depending on model size, $\Delta\chi^2$ degrees of freedom were 5, 11 or 17 depending on model size, ΔCFI= difference between equality of factor structure and equality of factor loading cfi values.

Table 4

Descriptive Statistics by Levels of Factor Loading Size for Baseline Cells (n_{cells}=54)

	SFL Level								
	Low			High			Mixed		
Dependent variables*	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
biasg1g	-0.05	-0.04	0.10	0.00	0.00	0.10	0.00	0.00	0.07
biasg2g	-0.05	-0.04	0.13	0.00	0.00	0.10	0.00	0.00	0.10
RMSEA	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01
NFI	0.99	0.99	0.01	0.99	0.99	0.00	0.99	0.99	0.00
NNFI	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
CFI	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
χ^2	168.38	143.19	123.34	168.76	142.83	123.65	168.91	142.16	123.63
$\Delta\chi 2$	11.04	10.02	6.85	11.16	10.11	6.97	11.26	9.86	11.28
ΔCFI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: * biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, $\Delta\chi^2$ = likelihood ratio χ^2 between equality of factor structure and equality of factor loading, χ^2 degrees of freedom were 33, 141, or 321 depending on model size, $\Delta\chi^2$ degrees of freedom were 5, 11 or 17 depending on model size, ΔCFI = difference between equality of factor structure and equality of factor loading cfi values.

Table 5

Descriptive Statistics by Levels of Model Size for Baseline Cells (n_{cells}=54)

	MS Level								
	6 items			12 items			18 items		
Dependent variables*	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
biasg1g	0.00	0.00	0.08	-0.05	-0.04	0.10	0.00	0.00	0.07
biasg2g	0.00	0.00	0.11	-0.05	-0.04	0.12	0.00	0.00	0.11
RMSEA	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01
NFI	0.99	0.99	0.00	0.99	0.99	0.01	0.99	0.99	0.01
NNFI	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
CFI	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
χ^2(df)	33.27 (33)	32.51	8.27	143.54 (141)	142.81	16.99	329.24 (321)	328.04	25.87
$\Delta\chi^2$(df)	4.95 (5)	4.29	3.16	11.07 (11)	10.46	4.70	17.44 (17)	16.61	10.61
ΔCFI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: * biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, df = degrees of freedom, $\Delta\chi^2$ = likelihood ratio χ^2 between equality of factor structure and equality of factor loading, ΔCFI = difference between equality of factor structure and equality of factor loading cfi values.

Table 6

Descriptive Statistics by Levels of Sample Size for Baseline Cells (n_{cells}=54)

	SS Level								
	n_1=n_2=200			n_1=n_2=450			n_1=600; n_2=300		
Dependent variables*	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
biasg1g	-0.01	-0.01	0.11	-0.01	-0.01	0.08	-0.02	-0.01	0.07
biasg2g	-0.01	-0.01	0.13	-0.01	-0.01	0.10	-0.01	-0.01	0.11
RMSEA	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.00	0.01
NFI	0.98	0.99	0.00	0.99	0.99	0.00	0.99	0.99	0.00
NNFI	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
CFI	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
χ^2	170.51	144.15	125.37	167.47	141.74	122.53	168.06	142.68	122.68
$\Delta\chi^2$	11.07	10.00	6.88	11.03	10.15	6.81	11.36	9.83	11.35
ΔCFI	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: * biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, $\Delta\chi^2$ = likelihood ratio χ^2 between equality of factor structure and equality of factor loading, χ^2 degrees of freedom were 33, 141, or 321 depending on model size, $\Delta\chi^2$ degrees of freedom were 5, 11 or 17 depending on model size, ΔCFI= difference between equality of factor structure and equality of factor loading cfi values.

Table 7

Descriptive Statistics by Levels of GAM for Baseline Cells (n_{cells}=54)

	GAM Level					
	G1=G2=0.5			G1=0.5; G2=0.3		
Dependent variables*	Mean	Median	SD	Mean	Median	SD
biasg1g	-0.01	-0.01	0.09	-0.02	-0.01	0.09
biasg2g	-0.01	-0.01	0.09	-0.01	-0.01	0.13
RMSEA	0.01	0.00	0.01	0.01	0.00	0.01
NFI	0.99	0.99	0.01	0.99	0.99	0.01
NNFI	1.00	1.00	0.00	1.00	1.00	0.00
CFI	1.00	1.00	0.00	1.00	1.00	0.00
χ^2	168.90	142.87	123.79	168.46	142.62	123.28
$\Delta\chi^2$	11.23	9.83	10.10	11.07	10.13	6.81
ΔCFI	0.00	0.00	0.00	0.00	0.00	0.00

Note: * biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, $\Delta\chi^2$ = likelihood ratio χ^2 between equality of factor structure and equality of factor loading, χ^2 degrees of freedom were 33, 141, or 321 depending on model size, $\Delta\chi^2$ degrees of freedom were 5, 11 or 17 depending on model size, ΔCFI = difference between equality of factor structure and equality of factor loading cfi values.

Partial Metric Invariance

Goodness of fit indices, power of the likelihood ratio test and bias between the sample estimate and population values of the regression path of interest and parameter estimates were examined as dependent variables. The likelihood ratio test was calculated as the difference between the χ^2 values of the test of configural invariance and test of metric invariance. Power was operationally defined as the proportion of true rejections of the likelihood ratio test across the number of replications within any given level of the research design. Bias was operationally defined as the average deviation of a sample value from the concomitant population value divided by the population value (Bandalos, in press). Power of the likelihood ratio test was examined with descriptive statistics and presented by independent variable and significant interactions of the goodness of fit indices (Tables 14-17). Tables of descriptive statistics for the dependent variables are provided in Appendix A.

Separate six-way analyses of variance were run on the goodness of fit indices RMSEA, NFI, NNFI, CFI, ΔCFI, χ^2, and bias estimates for group 1 gamma (biasg1g) and group 2 gamma (biasg2g). The independent variables in the model were size of factor loading (SFL), model size (MS), proportion of items noninvariant (II), sample size (SS) factor loading difference (FLD) and differing levels of predictive influence in the population (GAM).

The large number of cells and of replications per cell required the examination of practical significance in the form of partial η^2 rather than significance tests. Effects greater than 0.14 (Cohen, 1988) were further analyzed. Results are presented by dependent variable (Table 8). The largest effect was the proportion of noninvariant items on RMSEA, the smallest effect of practical significance was the interaction of factor loading size and proportion of noninvariant items on bias of the group 2 gamma parameter.

Table 8

Partial η^2 Values by Dependent Variable

Dependent Variables	Effect	Partial η^2
RMSEA	II	0.51
	FLD	0.36
NFI	MSxII	0.17
	FLD	0.23
NNFI	MSxIIxFLD	0.24
CFI	MSxIIxFLD	0.21
χ^2	MS	0.46
	II	0.19
ΔCFI	MSxIIxFLD	0.23
bias of group 1 gamma	--	--
bias of group 2 gamma	SFLxII	0.15

Goodness of Fit Indices

RMSEA appears to be most affected by proportion of items noninvariant and factor loading difference, but not by the interaction of these two independent variables. The RMSEA estimated marginal means and their 95% confidence intervals are located in Table 9. According to Byrne (1998) RMSEA should be less than 0.05 for the model to be considered optimal, while Hu and Bentler (1999) suggest a cut off value of 0.06. In this study, when the proportion of noninvariant items was largest (2/3 items noninvariant) and when the factor loading difference was largest (0.4 difference) the value of RMSEA rose to above 0.08; however, even with 1/3 of items noninvariant, values of RMSEA still suggest reasonable model fit in metric invariance.

The analysis of NFI produced a MS x II interaction as well as a FLD main effect. Values of NFI greater than 0.95 would indicate acceptable model fit (Hu & Bentler, 1998, 1999). As can be seen in Figure 2, the graph of NFI estimated marginal means by MS and II shows that when the model is small and the proportion of noninvariance is high, NFI is much lower than when the proportion of noninvariance is less or the model is larger. In the case of a small model with a large proportion of noninvariant items, NFI falls below the recommended 0.95. In all other cases

across model size and proportion of noninvariant items, NFI suggests acceptable model fit in metric invariance. The NFI estimated marginal means for FLD suggest the larger the factor loading difference, the lower NFI (Table 10), but NFI still remains above 0.95.

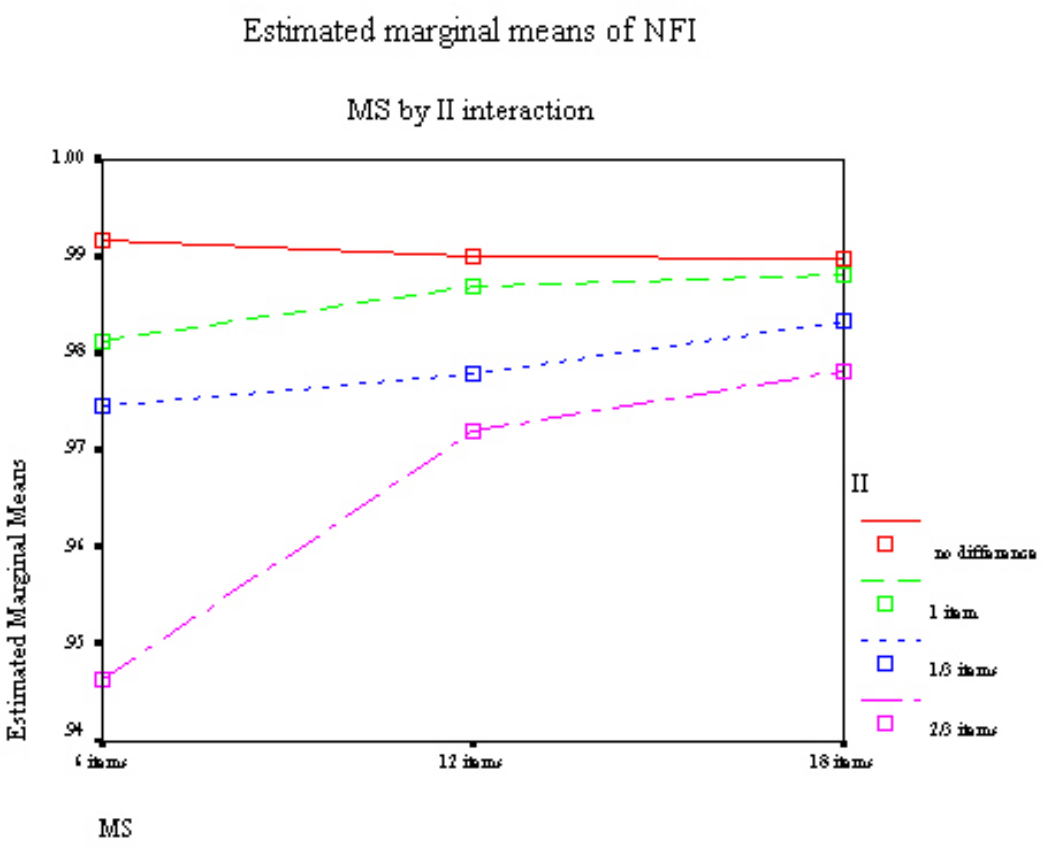

Figure 2. NFI Estimated Marginal Means by MS and II

Table 9

RMSEA Estimated Marginal Means for Proportion of Noninvariant Items and Factor Loading Difference

Independent Variable	Estimated Marginal Mean	95% Confidence Interval
II0	0.008	0.008-0.009
II1	0.033	0.033-0.033
II2	0.061	0.061-0.062
II3	0.091	0.091-0.091
FLD0	0.008	0.008-0.009
FLD1	0.040	0.040-0.040
FLD2	0.062	0.062-0.062
FLD3	0.084	0.083-0.084

Table 10

NFI Estimated Marginal Means for Factor Loading Difference

Independent Variable	Estimated Marginal Mean	95% Confidence Interval
FLD0	0.990	0.990-0.991
FLD1	0.985	0.985-0.985
FLD2	0.978	0.977-0.978
FLD3	0.967	0.967-0.967

Both NNFI and CFI have skewed distributions overall, so the following results need to be examined with caution. In both cases there was a MS x II x FLD interaction. The magnitudes of the estimated means for NNFI and CFI are similar; therefore only CFI estimated marginal means will be reported (Table 11). Acceptable values for CFI are those greater than 0.95 (Hu & Bentler 1998, 1999). In this study, CFI values were worst when the model size was small, two-thirds of the items were noninvariant and the factor loading difference between groups was 0.4. The trend of the interaction seems to be such that in small models, the proportion of noninvariant items and factor loading differences has a greater effect than in larger models.

Table 11

CFI Estimated Marginal Means for Model Size, Proportion of Noninvariant Items and Factor Loading Difference

Dependent Variables			Estimated Marginal Mean	95% Confidence Interval
Model Size	Proportion Items Noninvariant	Factor Loading Difference		
6 items	No difference	No difference	0.999	0.999-0.999
	1 item	0.2 difference	0.996	0.996-0.996
		0.3 difference	0.991	0.991-0.991
		0.4 difference	0.983	0.983-0.983
	1/3 items	0.2 difference	0.994	0.994-0.994
		0.3 difference	0.985	0.985-0.985
		0.4 difference	0.972	0.972-0.972
	2/3 items	0.2 difference	0.982	0.982-0.982
		0.3 difference	0.959	0.958-0.959
		0.4 difference	0.924	0.924-0.924
12 items	No difference	No difference	0.999	0.999-1.000
	1 item	0.2 difference	0.999	0.998-0.999
		0.3 difference	0.997	0.997-0.997
		0.4 difference	0.995	0.995-0.995
	1/3 items	0.2 difference	0.995	0.995-0.995
		0.3 difference	0.988	0.988-0.989
		0.4 difference	0.979	0.979-0.980
	2/3 items	0.2 difference	0.992	0.992-0.992
		0.3 difference	0.983	0.983-0.983
		0.4 difference	0.970	0.970-0.970
18 items	No difference	No difference	1.000	0.999-1.000
	1 item	0.2 difference	0.999	0.999-0.999
		0.3 difference	0.998	0.998-0.999
		0.4 difference	0.997	0.997-0.998
	1/3 items	0.2 difference	0.997	0.996-0.997
		0.3 difference	0.993	0.993-0.993
		0.4 difference	0.988	0.987-0.988
	2/3 items	0.2 difference	0.995	0.995-0.995
		0.3 difference	0.989	0.988-0.989
		0.4 difference	0.980	0.980-0.980

Though by definition χ^2 does not have a normal distribution, when sampled within a cell it does exhibit normal distribution properties. Allowing for this fact, χ^2 appears to be most affected by model size and proportion of noninvariance as main effects. It is not surprising that χ^2 should be affected by model size as the degrees of freedom in the equation increase with more parameters to estimate (see Table 12 for estimated marginal means).

Table 12

χ2 Estimated Marginal Means for Model Size and Proportion Noninvariant Items

Independent variable	Estimated marginal mean	95% confidence interval
MS1	118.51	116.61-120.41
MS2	310.14	308.24-312.04
MS3	578.77	576.86-580.67
II0	168.68	164.44-172.92
II1	213.69	211.24-216.13
II2	356.14	353.70-358.59
II3	493.30	490.85-495.75

Note: MS_1 df=33, MS_2 df=141, MS_3 df=321.

The difference between values of CFI (ΔCFI) in the tests of configural invariance and metric invariance was also examined following the work of Cheung and Rensvold (2002). They examined 20 change in goodness of fit indices and suggest a cut off value for ΔCFI of –0.01, below which "the null hypothesis of invariance should not be rejected" (Cheung & Rensvold, 2002, p. 251 As with CFI, ΔCFI values also exhibited a MS x II x FLD interaction (Table 13). In this study, the ΔCFI is largest when the factor loading difference is 0.4, across all model sizes and particularly when two-thirds of the items were noninvariant.

Power of the Likelihood Ratio Test

Power of the likelihood ratio test to detect metric noninvariance was high. Following the practically significant results of the goodness of fit indices, power estimates of the likelihood ratio test are provided (Tables 14-17). In this study, the power of the likelihood ratio test to detect metric noninvariance increased as the amount of noninvariance increased. Noninvariance was less likely to be detected when the amount of noninvariance was small.

Table 13.

ΔCFI Estimated Marginal Means for Model Size, Proportion of Items Noninvariant and Factor Loading Difference

Dependent Variables			Estimated Marginal Mean	95% Confidence Interval
Model Size	Proportion Items Noninvariant	Factor Loading Difference		
6 items	No difference	No difference	0.000	0.000-0.000
	1 item	0.2 difference	-0.003	-0.003-(-0.003)
		0.3 difference	-0.008	-0.008-(-0.008)
		0.4 difference	-0.016	-0.016-(-0.015)
	1/3 items	0.2 difference	-0.005	-0.006-(-0.005)
		0.3 difference	-0.014	-0.014-(-0.014)
		0.4 difference	-0.027	-0.027-(-0.027)
	2/3 items	0.2 difference	-0.017	-0.017-(-0.017)
		0.3 difference	-0.041	-0.041-(-0.040)
		0.4 difference	-0.075	-0.075-(-0.075)
12 items	No difference	No difference	0.000	0.000-0.000
	1 item	0.2 difference	-0.001	-0.001-(-0.001)
		0.3 difference	-0.002	-0.002-(-0.002)
		0.4 difference	-0.004	-0.005-(-0.004)
	1/3 items	0.2 difference	-0.005	-0.005-(-0.004)
		0.3 difference	-0.011	-0.011-(-0.011)
		0.4 difference	-0.020	-0.020-(-0.020)
	2/3 items	0.2 difference	-0.007	-0.007-(-0.007)
		0.3 difference	-0.017	-0.017-(-0.016)
		0.4 difference	-0.030	-0.030-(-0.030)
18 items	No difference	No difference	0.000	0.000-0.000
	1 item	0.2 difference	0.000	-0.001-0.000
		0.3 difference	-0.001	-0.001-(-0.001)
		0.4 difference	-0.002	-0.002-(-0.002)
	1/3 items	0.2 difference	-0.003	-0.003-(-0.003)
		0.3 difference	-0.007	-0.007-(-0.007)
		0.4 difference	-0.012	-0.012-(-0.012)
	2/3 items	0.2 difference	-0.005	-0.005-(-0.005)
		0.3 difference	-0.011	-0.011-(-0.011)
		0.4 difference	-0.019	-0.019-(-0.019)

Table 14

Power Estimates of the Likelihood Ratio Test by Independent Variable

	Independent Variable					
Level	SFL	MS	II	SS	FLD	GAM
1	0.97	0.98	0.93	0.94	0.93	0.98
2	0.98	0.98	0.99	1.00	0.99	0.98
3	0.97	0.97	1.00	0.99	1.00	

Table 15

Power Estimates of the Likelihood Ratio Test across Size of Factor Loading and Proportion of Items Noninvariant

	Proportion of Items Noninvariant		
Size of Factor Loading	1 item	1/3 items	2/3 items
Small	0.93	0.99	1.00
Large	0.94	1.00	1.00
Mixture	0.93	0.99	1.00

Table 16

Power Estimates of the Likelihood Ratio Test across Model Size and Proportion of Items Noninvariant

	Proportion of Items Noninvariant		
Model Size	1 item	1/3 items	2/3 items
6 items	0.95	0.98	1.00
12 items	0.93	1.00	1.00
18 items	0.92	1.00	1.00

Table 17

Power Estimates of the Likelihood Ratio Test across Model Size, Proportion of Items Noninvariant and Magnitude of Factor Loading Differences

	Proportion of Items Noninvariant								
	1 item			1/3 items			2/3 items		
	Factor Loading Difference								
Model Size	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
6 items	0.86	0.99	1.00	0.95	0.99	1.00	1.00	1.00	1.00
12 items	0.80	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00
18 items	0.77	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00

Parameter Estimates and Bias

There were no results of practical significance (partial η^2>0.14) in bias for the group one parameter estimate. The only practically significant result in group two gamma bias was a SFLxII interaction effect (Table 18). Bias (in the form of proportion of a gamma parameter estimate) was greater for smaller factor loadings, and roughly similar for large factor loadings and a mixture of factor loadings.

Table 18

Group 2 Bias Estimated Marginal Means for Size of Factor Loading and Proportion of Noninvariant Items

Dependent Variable		Effect	95% Confidence Interval
SFL1	II0	-0.046	-0.050-(-0.042)
	II1	-0.016	-0.019-(-0.014)
	II2	0.048	0.046-0.051
	II3	0.254	0.252-0.257
	Mean	0.06	
SFL2	II0	0.000	-0.004-0.004
	II1	0.022	0.019-0.024
	II2	0.029	0.027-0.031
	II3	0.014	0.011-0.016
	Mean	0.02	
SFL3	II0	0.004	0.000-0.008
	II1	0.030	0.028-0.033
	II2	0.062	0.060-0.064
	II3	0.017	0.015-0.020
	Mean	0.03	

Summary

The descriptive statistics for the baseline condition provide informative results. There is a Type 1 error rate of between 0.02 and 0.08 in the invariant conditions. There was a slight

underestimation of bias in the invariant conditions, but generally bias and fit indices gave favorable results.

The results for partial measurement invariance suggest that model size, level of noninvariance and factor loading differences affect goodness of fit indices while the size of the factor loading has a slight affect on bias.

CHAPTER 5

DISCUSSION

The purpose of this study was to investigate the amount of partial metric invariance that can be tolerated while still allowing for equally accurate predictions across groups. By varying size of factor loadings, model size, sample size, amount of partial measurement invariance, factor loading differences and differing predictive influence across groups the following research questions could be answered: how will group sample size influence prediction, how will model size influence prediction, how will factor loading size influence prediction, how will the number of invariant factor loadings and level of factor loading differences influence prediction, and how will differing amounts of predictive influence be affected by partial measurement invariance.

There was a steady increase in the RMSEA value (and therefore decline in fit) the larger the amount of noninvariance and factor loading difference. Under the conditions studied, RMSEA was optimal when the proportion of items noninvariant was no larger than 1/3 and the factor loading difference across groups was no larger than 0.3 across the noninvariant items. In the conditions studied, the values of NFI indicate acceptable fit even in the worst case of a large proportion of noninvariance coupled with a small model or the greatest amount of factor loading difference. The MS by II interaction shows that when the model is small and the proportion of noninvariance is high, NFI is much lower than when the proportion of noninvariance is less or the model is larger. A similar result was found in CFI in that the smaller model with a high proportion of noninvariance and larger factor loading difference yielded a lower CFI than other model sizes/proportions of noninvariance/factor loading size. The analysis of ΔCFI produced

similar results to those of CFI. Cheung and Rensvold (2002) in their examination of goodness of fit indices under the null hypothesis of invariance found average ΔCFI (metric invariance – configural invariance) values of -0.0001, (first percentile value= -0.0085), while the values in this study were much larger, indicating noninvariance. Independent of model size, the χ^2 fit statistic increases with increasing amounts of noninvariance, indicating robustness for differentiating poor fit not seen in NFI, NNFI or CFI.

Bias in the group two gamma parameter estimate was large (0.254) and tended to overestimate the group two gamma parameter estimate in conditions where the factor loading size was small and the proportion of noninvariance was large. Bias for the group two gamma parameter estimate was roughly similar for large factor loadings and a mixture of factor loadings across levels of noninvariance.

As noted previously, sample size is a frequently studied factor in partial measurement invariance. Meade and Lautenschlager (2004) found 1000 per group to yield excessive power as their rates of detection were near 100 percent. Kaplan and George (1995) found that unequal sample size was likely to reduce power to detect partial measurement invariance. In the current study, it was expected that larger but equivalent group sample seizes would provide more stable estimates of prediction than unequal sample sizes. This does not appear to be the case, as sample size did not prove to be a factor of practical significance in any of the six-way ANOVAs. That is, the sample sizes chosen did not provide sufficiently different results as to be meaningful.

ME/I simulation studies to date have used models of similar size, but nearly all of the results presented focus on other aspects of the research design. It was expected that smaller models would provide more accurate estimates of prediction than larger models. Model size seemed to play an important role in the goodness of fit indices – while interacting with number

of items invariant (NFI) and factor loading differences (NNFI, CFI and ΔCFI), the smallest model seemed to incur the worst fit. Power was somewhat diminished when there were larger models with only one item noninvariant. Therefore, model size appears to be an important consideration in partial measurement invariance.

As was expected, the levels of factor loading size that included larger factor loadings resulted in more accurate estimates of gamma than the level with only smaller factor loadings. These results are similar to those of Kaplan (1989) and Meade and Lautenschlager (2004).

It was expected that smaller proportions of noninvariance would provide more accurate estimates of prediction than larger proportions of noninvariance. Both the proportion of items noninvariant and the factor loading differences across groups seem to act in concert with model size. The larger the proportion of items noninvariant and the larger the factor loading difference across groups, the worse the fit as measured by all goodness of fit indices examined. As well, bias of the gamma parameter estimate was somewhat affected by smaller factor loadings and larger proportions of noninvariance in that these conditions yielded larger bias estimates. Power was slightly diminished when factor loadings were smaller and the proportion of noninvariant items was small.

Limitations

The results of this study provide further information about how partial metric invariance affects the decisions one might make regarding comparisons across groups in simulated conditions. However, as a simulation study these results may not reflect real data conditions and may not be generalizable to all model sizes/complexities, other sizes of factor loadings, or other levels of predictive influence.

Suggestions for Future Research

There are an infinite number of combinations of factor loading size that might be utilized and it is important to look more fully at how a mixture (both larger and smaller loadings across groups) of factor loadings might affect accuracy of prediction across models with more than one latent factor. Further comparisons of noninvariance and predictive influence in one and two (or more) factor models are necessary to understand more fully these effects. Sample size did not have the expected effect on prediction, estimated parameter estimates, or bias. It is possible that the sample sizes were not sufficiently different from each other that such a difference in the dependent variables could be noted. Larger unequal sample sizes have been studied by others but the focus was on power of the likelihood ratio test (Hutchinson & Young, 2003, Kaplan & George, 1995). Therefore, in further research examining accuracy of prediction, the unequal sample size level should have a greater difference between the two groups.

Further analysis also needs to be done to examine the change in goodness of fit indices put forward by Cheung and Rensvold (2002) in other contexts. These authors examined 20 goodness of fit indices under the null hypothesis of invariance and recommended reporting ΔCFI in results of invariance tests because they found it was not significantly correlated with CFI in configural invariance, and it was not affected by model complexity. However, the correlations between CFI and ΔCFI in this study were somewhat larger than those found by Cheung & Rensvold, therefore, further analysis is needed to examine these different results.

Finally, future research should be focused on mean and covariance structure analysis (MACS). MACS is an extension of the traditional multiple group covariance analysis in that the comparison of intercepts and latent means across groups is an integral part of the methodological process. Equality of item intercepts indicates that those with equal amounts of the factor have

equivalent scores on the indicators. If this is not the case, comparisons of latent means are specious because they may reflect differential response patterns on the indicators rather than true differences in mean factor levels. Therefore, before one can interpret latent mean differences across groups one must examine whether item intercepts are equivalent. Whether item intercepts and/or latent means are allowed to be freely estimated or are constrained to be equal across groups is dependent upon whether or not latent mean differences are theoretically meaningful. When using MACS a researcher can examine latent mean differences across groups, if appropriate, then instead of moving back to an observed variable analysis like analysis of variance, carry out latent pairwise comparisons across different groups, and analyze repeated measures over time and across groups (Ployhart and Oswald, 2004). This is advantageous for various reasons including the fact that measurement error is accounted for throughout the analysis of interest, that MACS can be used to specifically model various violations in the assumptions found in analysis of variance (e.g. homogeneity of variance) and has estimation methods that are robust to violations of normality (Yuan, Bentler, and Zhang, 2005).

Recommendations for Applied Researchers

The interaction between model size, proportion of items noninvariant and factor loading differences plays a large role in partial measurement invariance. Byrne et al. (1989) suggested that only one item needed to be invariant and Reise et al. (1993) suggested that at least half of the items needed to be invariant for group comparisons to be meaningful. Contrary to these recommendations it appears from the results of this study that the proportion of noninvarniant items should be no more than one-third with a sufficiently large model. The applied researcher should exercise care when interpreting invariance results in small models with small factor

loadings coupled with large factor loading differences across groups as any comparison across groups may be inaccurate.

REFERENCES

Bandalos, D. L. (in press). Use of simulation studies in SEM research. In G. R. Hancock & R. O. Mueller (Eds.), *Methods in education and the behavioral sciences: Issues, research, and teaching.* Greenwich, CT: Information Age Publishing.

Byrne, B. M. (1998). S*tructural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. New Jersey: Erlbaum.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research, 18*(3), 117-144.

Hu, L. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424-453.

Hu, L & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*(2), 131-152.

Hutchinson, S. R. (2002, October). *Power of the likelihood ratio test under factor loading noninvariance in confirmatory factor analysis models.* Paper presented at the annual meeting of the Northern Rocky Mountain Educational Research Association, Estes Park, CO.

Hutchinson, S. R., & Young, J. D. (2003, April). *Power of the likelihood ratio test under various conditions of factor loading noninvariance in confirmatory factor analysis models*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*(4), 409-426.

Jöreskog, K., & Sörbom, D. (1996a). *LISREL 8: User's reference guide*. Chicago: SSI.

Jöreskog, K., & Sörbom, D. (1996b). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago: SSI.

Jöreskog, K., & Sörbom, D. (1996c). PRELIS 2: User's reference guide. Chicago: SSI.

Kaplan, D. (1989). Power of the likelihood ratio test in multiple group confirmatory factor analysis under partial measurement invariance. *Educational and Psychological Measurement, 49*, 579-586.

Kaplan, D., & George R. (1995). A Study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling, 2*, 101-118.

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. New Jersey: Prentice-Hall, Inc.

Meade, A. W., & Lautenschlager, G. L. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling 11*(1), 60-72.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543.

Millsap, R. E., & Kwok, O-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93-115.

Muthén, B. & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46*(4), 407-419.

Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: integrating correlational and experimental approaches. *Organizational Research Methods, 7*(1), 27-65.

Poortinga, Y.H. (1989). Equivalence of cross cultural data: An overview of basic issues. *International Journal of Psychology, 24*, 737-756.

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

Steenkamp, J-B. E. M. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *The Journal of Consumer Research, 25*(1), 78-90.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158.

Vandenberg, R. J., & Lance, C. L. (2000). The review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69.

Yuan, K-H., Bentler, P. M., & Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis: The univariate case and its multivariate implication. *Sociological Methods and Research, 34*(2), 240-258.

APPENDIX A

DESCRIPTIVE STATISTIC FOR ALL INDEPENDENT VARIABLES

IN THE NONINVARIANCE PART OF THE STUDY

The following tables (Tables A1.1-A7) include descriptive statistics for all of the independent variables in the noninvariance part of the study. These tables are based on the test of factor loadings (metric invariance).

Table A1.1

Descriptive Statistics for Test of Equality of Factor Structure for Partial Measurement Invariance Cells (n_{cells}=486)

Factor Structure	Statistics		
Dependent variables*	Mean	Median	SD
biasg1g	0.00	0.00	0.08
biasg2g	0.00	0.00	0.12
RMSEA	0.01	0.00	0.01
NFI	0.99	0.99	0.00
NNFI	1.00	1.00	0.00
CFI	1.00	1.00	0.00
χ^2	157.67	131.76	118.60

Note: * biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit indexχ^2 degrees of freedom were 28, 130, or 304 depending on model size.

Table A1.2

Descriptive Statistics for Test of Equality of Factor Loadings for Partial Measurement Invariance Cells (n_{cells}=486)

Factor Loading	Statistics		
Dependent variables*	Mean	Median	SD
biasg1g	-0.03	-0.02	0.08
biasg2g	0.05	0.03	0.16
RMSEA	0.06	0.06	0.04
NFI	0.98	0.98	0.02
NNFI	0.98	0.99	0.02
CFI	0.99	0.99	0.02
χ^2	354.38	294.40	285.52
$\Delta\chi^2$	196.71	111.17	222.03
ΔCFI	-0.01	-0.01	0.02

Note: * biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, $\Delta\chi^2$= likelihood ratio χ^2 between equality of factor structure and equality of factor loading, χ^2 degrees of freedom were 33, 141, or 321 depending on model size, $\Delta\chi^2$ degrees of freedom were 5, 11 or 17 depending on model size, ΔCFI = difference between equality of factor structure and equality of factor loading cfi values.

Table A2

Descriptive Statistics by Level of Size of Factor Loading for Partial Measurement Invariance Cells (n_{cells}=486)

	SFL Level								
	1 (0.7-0.3)			2 (0.9-0.5)			3 (0.9-0.3)		
Dependent variables*	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
biasg1g	-0.05	-0.05	0.09	-0.01	-0.01	0.07	-0.02	-0.02	0.07
biasg2g	0.10	0.05	0.21	0.02	0.02	0.11	0.04	0.03	0.12
RMSEA	0.06	0.05	0.03	0.07	0.06	0.04	0.06	0.06	0.04
NFI	0.97	0.98	0.02	0.98	0.99	0.01	0.98	0.98	0.02
NNFI	0.98	0.99	0.02	0.99	0.99	0.02	0.98	0.99	0.02
CFI	0.99	0.99	0.02	0.99	0.99	0.01	0.98	0.99	0.02
χ^2	306.16	261.78	207.52	381.35	318.89	315.77	375.61	316.00	313.51
$\Delta\chi^2$	148.36	101.42	134.37	223.68	119.95	252.76	218.08	114.61	249.85
ΔCFI	-0.01	-0.01	0.02	-0.01	-0.01	0.04	-0.01	-0.01	0.02

Note: * biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, $\Delta\chi^2$ = likelihood ratio χ^2 between equality of factor structure and equality of factor loading, χ^2 degrees of freedom were 33, 141, or 321 depending on model size, $\Delta\chi^2$ degrees of freedom were 5, 11 or 17 depending on model size, ΔCFI = difference between equality of factor structure and equality of factor loading cfi values.

Table A3

Descriptive Statistics by Level of Model Size for Partial Measurement Invariance Cells

(n_{cells}=486)

	MS Level								
	1 (6 items)			2 (12 items)			3 (18 items)		
Dependent variables*	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
biasg1g	-0.03	-0.03	0.08	-0.03	-0.03	0.09	-0.02	-0.02	0.08
biasg2g	0.06	0.05	0.14	0.04	0.02	0.17	0.05	0.03	0.16
RMSEA	0.08	0.08	0.04	0.06	0.05	0.03	0.05	0.04	0.03
NFI	0.97	0.98	0.02	0.98	0.98	0.01	0.98	0.99	0.01
NNFI	0.97	0.98	0.03	0.99	0.99	0.01	0.99	0.99	0.01
CFI	0.98	0.98	0.02	0.99	0.99	0.01	0.99	0.99	0.01
χ^2(df)	127.98 (33)	98.83	91.98	328.65 (141)	259.94	186.82	606.49 (321)	499.27	292.79
$\Delta\chi^2$(df)	99.59 (5)	70.27	91.61	196.19 (11)	124.36	186.52	294.35 (17)	183.00	292.82
ΔCFI	-0.02	-0.02	0.02	-0.01	-0.01	0.01	-0.01	0.00	0.01

Note: biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, df=degrees of freedom, $\Delta\chi^2$ = likelihood ratio χ^2 between equality of factor structure and equality of factor loading, ΔCFI = difference between equality of factor structure and equality of factor loading cfi values.

Table A4

Descriptive Statistics by Level of Items Invariant for Partial Measurement Invariance Cells

(n_{cells}=486)

	II Level								
	1 (1 item)			2 (1/3 items)			3 (2/3 items)		
Dependent variables*	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
biasg1g	-0.03	-0.03	0.09	-0.02	-0.02	0.08	-0.02	-0.02	0.08
biasg2g	0.01	0.01	0.12	0.05	0.04	0.13	0.10	0.05	0.21
RMSEA	0.03	0.03	0.02	0.06	0.06	0.02	0.09	0.08	0.04
NFI	0.99	0.99	0.01	0.98	0.98	0.01	0.97	0.97	0.02
NNFI	0.99	1.00	0.01	0.99	0.99	0.01	0.97	0.98	0.03
CFI	1.00	1.00	0.01	0.99	0.99	0.01	0.97	0.98	0.02
χ^2	213.69	184.52	129.87	356.14	316.33	252.71	493.30	399.28	353.18
$\Delta\chi^2$	56.02	48.96	31.95	198.37	146.48	163.90	335.73	241.86	284.39
ΔCFI	0.00	0.00	0.01	-0.01	-0.01	0.01	-0.02	-0.02	0.02

Note: * biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, $\Delta\chi^2$ = likelihood ratio χ^2 between equality of factor structure and equality of factor loading, χ^2 degrees of freedom were 33, 141, or 321 depending on model size, $\Delta\chi^2$ degrees of freedom were 5, 11 or 17 depending on model size, ΔCFI = difference between equality of factor structure and equality of factor loading cfi values.

Table A5

Descriptive Statistics by Level of Sample Size for Partial Measurement Invariance Cells

(n_{cells}=486)

	SS Level								
	1 (n_1=n_2=200)			2 (n_1=n_2=450)			3 (n_1=600, n_2=300)		
Dependent variables*	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
biasg1g	-0.03	-0.03	0.10	-0.03	-0.02	0.07	-0.02	-0.02	0.06
biasg2g	0.04	0.02	0.17	0.04	0.02	0.14	0.07	0.05	0.17
RMSEA	0.06	0.06	0.04	0.06	0.06	0.04	0.06	0.05	0.04
NFI	0.97	0.97	0.02	0.98	0.98	0.02	0.98	0.99	0.01
NNFI	0.98	0.99	0.02	0.98	0.99	0.02	0.99	0.99	0.02
CFI	0.98	0.99	0.02	0.99	0.99	0.02	0.99	0.99	0.01
χ^2	278.04	225.74	201.00	410.39	346.57	330.12	374.70	322.37	292.93
$\Delta\chi^2$	118.21	71.52	120.52	253.93	143.27	267.78	217.99	124.39	227.52
ΔCFI	-0.01	-0.01	0.02	-0.01	-0.01	0.02	-0.01	-0.01	0.01

Note: * biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, $\Delta\chi^2$ = likelihood ratio χ^2 between equality of factor structure and equality of factor loading, χ^2 degrees of freedom were 33, 141, or 321 depending on model size, $\Delta\chi^2$ degrees of freedom were 5, 11 or 17 depending on model size, ΔCFI = difference between equality of factor structure and equality of factor loading cfi values.

Table A6

Descriptive Statistics by Level of Factor Loading Difference for Partial Measurement Invariance Cells (n_{cells}=486)

	FLD Level								
	1 (0.2 difference)			2 (0.3 difference)			3 (0.4 difference)		
Dependent variables*	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
biasg1g	-0.02	-0.02	0.08	-0.03	-0.02	0.08	-0.03	-0.03	0.08
biasg2g	0.02	0.02	0.12	0.05	0.03	0.15	0.08	0.04	0.19
RMSEA	0.04	0.04	0.02	0.06	0.06	0.03	0.08	0.08	0.04
NFI	0.99	0.99	0.01	0.98	0.98	0.01	0.97	0.97	0.02
NNFI	0.99	0.99	0.01	0.98	0.99	0.02	0.97	0.98	0.03
CFI	0.99	1.00	0.01	0.99	0.99	0.01	0.98	0.98	0.02
χ^2	249.58	211.46	172.04	345.14	299.61	249.01	468.41	387.01	359.01
$\Delta\chi^2$	91.95	60.82	82.92	187.41	116.83	175.06	310.77	190.72	293.77
ΔCFI	-0.01	0.00	0.01	-0.01	-0.01	0.01	-0.02	-0.02	0.02

Note: biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, $\Delta\chi^2$ = likelihood ratio χ^2 between equality of factor structure and equality of factor loading, χ^2 degrees of freedom were 33, 141, or 321 depending on model size, $\Delta\chi^2$ degrees of freedom were 5, 11 or 17 depending on model size, ΔCFI = difference between equality of factor structure and equality of factor loading cfi values.

Table A7

Descriptive Statistics by Level of GAM for Partial Measurement Invariance Cells (n_{cells}=486)

	GAM Level					
	1 (G_1=G_2=0.5)			2 (G_1=0.5, G_2=0.3)		
Dependent variables*	Mean	Median	SD	Mean	Median	SD
biasg1g	-0.03	-0.02	0.08	-0.03	-0.02	0.08
biasg2g	0.05	0.03	0.14	0.05	0.03	0.18
RMSEA	0.06	0.06	0.04	0.06	0.06	0.04
NFI	0.98	0.98	0.02	0.98	0.98	0.02
NNFI	0.98	0.99	0.02	0.98	0.99	0.02
CFI	0.99	0.99	0.02	0.99	0.99	0.02
χ^2	354.60	294.19	285.67	354.15	294.86	285.36
$\Delta\chi^2$	196.88	111.30	222.29	196.54	111.03	221.76
ΔCFI	-0.01	-0.01	0.02	-0.01	-0.01	0.02

Note: biasg1g = bias of group 1 gamma parameter, biasg2g = bias for group 2 gamma parameter, RMSEA = root mean square error of approximation, NFI = normed fit index, NNFI = non-normed fit index, CFI = comparative fit index, $\Delta\chi^2$ = likelihood ratio χ^2 between equality of factor structure and equality of factor loading, χ^2 degrees of freedom were 33, 141, or 321 depending on model size, $\Delta\chi^2$ degrees of freedom were 5, 11 or 17 depending on model size, ΔCFI = difference between equality of factor structure and equality of factor loading cfi values.