2011-01-01

Assessing Measurement Invariance in the Presence of Testlets

Luis Andres Alvarado
University of Texas at El Paso, laalvarado@miners.utep.edu

ASSESSING MEASUREMENT INVARIANCE IN THE PRESENCE OF TESTLETS

LUIS A. ALVARADO

Department of Mathematical Sciences

APPROVED:

Amy Wagler, Ph.D.

Naijun Sha, Ph.D

Osvaldo Morera, Ph.D

Benjamin C. Flores , Ph.D.
Acting Dean of the Graduate School

to my

MOTHER and FATHER

with love

ASSESSING MEASUREMENT INVARIANCE IN THE PRESENCE OF TESTLETS

by

LUIS A. ALVARADO

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2011

Acknowledgments

I would like to express my deep-felt gratitude to my advisor, Dr. Amy Wagler of the Mathematical Sciences Department at The University of Texas at El Paso, for her advice, encouragement, and enduring patience. None of this would of been possible without her constant support. For always advising me to keep pushing me forward and showing me that with hard work and dedication anything can be accomplished.

I would also like to thank all my professors and faculty who have constantly challenged me and kept me on my toes throughout my graduate career. It was through their hard work and dedication that has provided me the means to complete my degree and prepare for a career as a Statistician. I would also like to thank Dr. Moschopoulos and Dr. Staiswalis for initially encouraging and inspiring me to apply to the M.S. in Statistics program. Had it not been for their passion and dedication to their field I would never of realized my own passion for Statistics.

To all of my peers; Robert, Melissa, Angelica, Priyangi, Samson, Indika, Sameera, and Saul thank you for helping me get through these past two years. You all have become like family to me and I will always remember the our late night study sessions that got us through all of our exams.

Additionally, I would like to thank my parents for always inspiring me to better myself. Had it not been for my father spending countless hours helping me with my Calculus homework or my mothers enduring love and support when I needed it most. Most importantly I would like to thank my brother for always keeping me on my toes and for helping me realize my true potential.

Finally, I would like to thank my wonderful wife Carla. I know I have put you through a lot this past year but I could not of done any of this without you. All the nights you stayed up to help me study and work on my thesis are memories I will never forget. I love you and I want you to know that I dedicate all of my accomplishments to you.

Abstract

Dealing with measurement invariance has been an issue of concern in confirmatory factor analysis for many years. It is important to establish measurement invariance across groups so that instruments may be validly used in multiple groups for comparison of the mean or summative scores. Throughout the years, many studies have considered testing for measurement invariance in factor models. However, there have been no studies that assess measurement invariance when so-called testlets should be modeled in the factor analytic model. Testlets add nuisance covariation to the model which can interfere when trying to detect measurement invariance. In the past, models have been developed to compensate for any sort of added covariation within a model, such as the correlated error model, CT-C(M-1) model and random intercept factor model. However, can such models help detect measurement invariance in the presence of testlets? Additionally, which testlet model is most useful for detecting the true level of measurement invariance? Simulations help determine when in fact it is possible to compensate for this added testlet-based covariation and determine which method works best for various measurement invariance tests and scenarios. Generally, it is found that in some scenarios none of the models correctly identify the level of measurement invariance and, otherwise, the correlated error model is least prone to type I error.

Table of Contents

Appendix

Chapter 1

Introduction

1.1 Factor Analysis

Factor analysis is a method used to account for the variation and co-variation among observed variables and a set of unobserved latent variables. By analyzing the nature of the relationships between these observed variables, we can determine the number of factors (latent variables) that partially explain their covariation. These factors are unobserved variables that account for a large proportion of the variation in the observed variables. The theory behind factor analysis is that the variation and covariation occurring among observed variables (items) is partially due to the variations that occur within the unobserved variable (factors, latent variables). In a factor analytic model we also routinely observe the nature of the factors, how well the hypothesized factors explain the data, and how much random error (unique variance) is present for each observed variable (item). In order explain such relationships between the latent and observed variables one must first understand how to go about modeling the data. In factor analysis (FA), using the Common Factor Model (CF-model) is one appropriate method used to model such data. The CF-model assesses the degree of the relationship between the observed variable(s) as explained by the latent variables. The CF-model with k common factors is depicted as follows:

$$\boldsymbol{x}_i = \boldsymbol{\tau}_i + \boldsymbol{\Lambda} * F + \boldsymbol{\varepsilon}_i \qquad i = 1, ..., N; \tag{1.1}$$

where

$$var(\boldsymbol{\varepsilon}) = \boldsymbol{\theta} \tag{1.2}$$

with θ diagonal, x is a $nx1$ column vector of scores of person "i" on "n" measured variables, τ_i is a vector of intercept terms, Λ is a nxk matrix of factor loadings and k is the number of latent variables. F is a $kx1$ random matrix of the factor scores and ε is a random error matrix (Johnson & Wichern, 2007). From the CF-model, we will be able to assess the covariance and the mean structures of the data.

1.1.1 Mean and Covariance Structure

The mean and the covariance structure from the model in (1.1) are useful for describing the data modeled using a CF-model. The mean and covariance structures express the population means and variances/covariances of the variables as a function of their parameters (Johnson & Wichern, 2007).The covariance structure of a CF-model is given by:

$$\boldsymbol{\Sigma}_{xx} = \boldsymbol{\Lambda}_x \phi \boldsymbol{\Lambda}_x^{'} + \boldsymbol{\psi}_\delta \tag{1.3}$$

Where $\boldsymbol{\Sigma}_{xx}$ is a (n x n) population covariance matrix of indicator variances and covariances, $\boldsymbol{\Lambda}$ is an (n x k) matrix of factor loadings, $\boldsymbol{\phi}$ is a (k x k) matrix of latent variables (factor) and, $\boldsymbol{\psi}_\delta$ is an (n x n) matrix of uniquenesses on the diagonal.

Now with regard to the mean structure, the CF-model based population mean of the observed variables is a function of its parameters (Meredith, 1993).

$$\boldsymbol{\mu}_x = \boldsymbol{\tau}_x + \boldsymbol{\Lambda}_x * F \tag{1.4}$$

where μ_x is the model based indicator mean, $\boldsymbol{\tau}_x$ are the indicator intercepts, $\boldsymbol{\Lambda}_x$ are the factor loadings, and F is the random vector of factor scores.

1.2 Exploratory vs. Confirmatory Factor Analysis

Factor analysis is customarily designed to examine the covariance structure of a specific set of variables and to help explain the relationships among these variables. However, there

are two forms of factor analysis which are very distinct from one another, exploratory and confirmatory factor analysis. Both exploratory and confirmatory factor analysis utilize the common factor model.

Exploratory factor analysis (EFA) is the most widely used form of factor analysis used today. One reason why EFA is prevalent is because it is highly data driven. No presumptions or specifications are made on the model. In particular, no latent variables or factor loadings are presumed to be set at a certain value or have a specific structure. The main purpose behind EFA is to explore or to investigate how many factors influence a set of items and to explore what the relationship exists between the common (latent) factors and observed variables.

Unlike EFA, Confirmatory Factor Analysis (CFA) requires a strong foundation on the specification of the factor model. With EFA we made no assumptions on the model itself however, with CFA certain assumptions are made about the factor loadings and the latent variables. This is all based on prior research studies that verify a particular latent structure for a set of items. It seeks to determine the number of factors and value of the factor loadings required to confirm what is expected based on previously established theory. This is done by setting the required amount of factors loading structure prior to the analysis. Compared to EFA, CFA requires a much larger sample size in order to perform the inferential statistics needed to decide if the model fits the data and the degree of fit (Brown, 2006).

Despite the restrictions and sample size requirements, CFA has a many advantages over EFA. These include the ability to analyze structural relationships and the ability to relax certain restrictions placed on the parameters within the CFA model. By relaxing these restrictions allows CFA practitioners to examine the equivalence of all measurement parameters within the factor model and to determine if it holds across several populations. This can help to establish whether a measure will hold across several populations and thus establish the equivalence of measurement across groups. However, before discussing measurement equivalence in detail, the methods of assessing fit in CFA models are reviewed.

1.2.1 CFA Fit Measures

This most frequently used fit measure in CFA is $\boldsymbol{\chi}^2$. Utilizing maximum likelihood estimation (MLE),

$$\boldsymbol{\chi}^2 = -2ln\boldsymbol{\Lambda} \tag{1.5}$$

where

$$\boldsymbol{\Lambda} = ln|S_N|^{-N/2} - ln|\boldsymbol{\Sigma}_{xx}|^{-N/2} + N[tr[S_N * \boldsymbol{\Sigma}_{xx}^{-1}] - p] \tag{1.6}$$

$$\boldsymbol{\Sigma}_{xx} = \boldsymbol{\Lambda}_x \boldsymbol{\phi} \boldsymbol{\Lambda}_x^{'} + \boldsymbol{\psi}_\delta$$

where, N is the sample size, p is the number of observed variables, S_N is the sample covariance matrix and $\hat{\Sigma}$ = Fitted model covariance matrix (Brown, 2006). This value is compared to a χ^2 critical value with degree of freedom ν - $\nu_o = \frac{1}{2}[(N-k)^2 - N - k]$. This is considered a "rigorous" test since the sample size must be large to maintain the specified error rate. Thus, less rigorous measures of fit are also available such as the root mean square error of approximation (RMSEA) or the comparative fit index (CFI). These fit indices will be discussed in detail in a later section.

1.2.2 CFA Example

The CFA is commonly used to determine how well a pre-determined factor model fits the data. When applied to psychological and educational data, this analysis is often called "testing construct validity". In both fields it is common to measure such constructs as "emotional intelligence","Personality", "Anxiety", etc by creating measurement tools such as surveys. These measurement tools, which consist of a set of statements or questions that relate to the construct at hand are used to predict the pre-determined factors. However, how do we know if in fact the items are measuring what they were intended to measure? Here is where CFA and construct validity come into play.

Recall that when CFA is utilized the investigator approaches the analysis with prior knowledge of what it is that is being measured and what measurement tools (such as a questionnaire) that have been used in the past. Using this prior knowledge, it is possible to validate the use of a prior measurement tool for a unique population.

The following example assesses the fit of a two-factor model. Each factor is believed to load onto the observed variables. In this case factor 1 measures Auditory Memory and factor 2 measures Visual Memory. The model is broken down in the following:

The model which we will be looking at is a 2 factor model with a sample size of 200 participants and with 6 observable variables. Each observable variable will load onto a specific factor such that X1: Logical Memory, X2: Verbal Paired Association, and X3: Word List will load onto Factor 1: Auditory Memory. X4: Faces, X5: Family Pictures, and X6: Visual Reproduction will load onto Factor 2: Visual Memory.

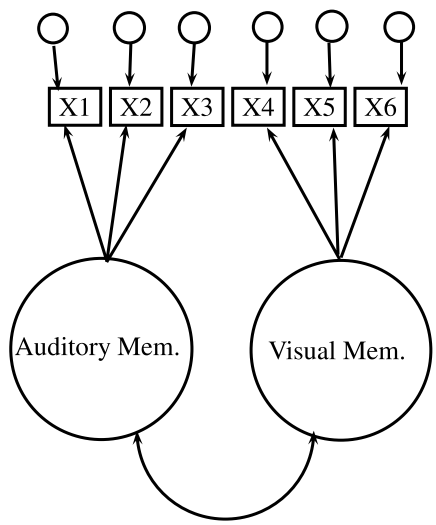

Table 1.1: Model Fit

	χ^2	df	P-value	RMSEA	CFI
Congeneric Solution	4.88	8	0.181	0	1.00

Looking at the Table 1.1, it is apparent that the CFA model holds. It is possible to come to this conclusion by analyzing the p-value associated with the chi-square test, the Root Mean Square Error of Approximation (The closer to 0 the better), and the Comparative Fit Index (the closer to 1 the better).

Chapter 2

Measurement Invariance in CFA

2.1 Measurement Invariance

Measurement Invariance (MI) can be defined in a variety of ways. In the context of CFA, MI tests evaluate whether the common factor model differs among target populations and is often called factorial invariance. Aspects of the factor model to be tested may include the loading structure, uniquenesses and intercept. To better understand this concept allow us to consider an example which shows how certain assumptions of MI can lead to non-meaningful results. Let us assume that we have a measurement scale (e.g. these scales can measure personality traits, stress levels, depression levels, etc...) and within this scale we intend to measure a specific trait. Usually such measurement scales are made up of a variety of questions called items. Initially the principle investigator would assume that the scale measures the same construct(s) (factors) across all groups within the population. At times, the items (observed variables) within such a scale have a certain degree of bias towards specific populations (age, gender, ethnicity, language, etc...) (Ellis, Aguirre-Urreta, Sun, & Marakas, 2008). In other words, can the nature of the item affect the participants response depending on the group (population) he/she is in. If so, how can we identify if the difference in the score are due to the differences in the model or in the population differences?

For example, when a scale is used for more than one population, it must be established that the same latent structure is measured for both populations. If the latent structure differs across the populations, then any observed differences in the scores may be due to the inadequacy of the CFA model for that data rather than true differences between the

populations. Therefore, MI helps to detect such issues by creating a variety of models each of which has a specific restriction of equality on one of its parameters (factor loading, intercept, and uniquenesses). We will be discussing model specification in detail in the following sections.

2.2 Restriction and Model Specification

Measurement invariance is assessed by setting constraints on structure of the CFA model (eq. 1.1) utilized for two or more populations. Nested models are obtained from the full model by fixing or eliminating certain parameters. Fixing or eliminating parameters in the CFA model allows the investigator to theorize that either the same latent structure, loadings, intercept, or uniquenesses apply across populations. In the case of testing to see for which level of measurement invariance holds, a comparison of a variety of nested models must be made in order to determine the most appropriate model. Theoretically there are a variety of nested models that can be created in order to decide which model best fits the data. It is necessary to realize that all statistical models are theoretically considered to be incorrect, invariance models are no exception. Therefore it is important to choose the most appropriate model and not necessarily the correct model.

Model specification plays an important role in deciding which level of invariance might hold. The question is how each model differs from the other and how to decide what parameters to fix or eliminate? The following sections describe the restrictions placed on the CF-model for testing MI, beginning with the so-called baseline model. This model is considered to have the same pattern of factor loadings but allow different loading values, uniquenesses and intercepts. We can now take this information and restrict the baseline model and restrict certain parameters such as the factor loadings, intercepts, and uniquenesses from the covariance and mean structure models. It is by restricting these models that we will be able to assess different levels of measurement invariance. These different levels of MI are tested within the framework of nested models hypothesis tests. In the next

few sections we will discuss each level of MI and the nested model hypothesis in greater detail.

2.2.1 Configural Invariance

MI testing always begins with testing the baseline model, or configural invariance model, for appropriate fit. The configural model assumes pattern of fixed and free factor loadings to be equal across groups with no equality restraints. Therefore, by not restricting the covariance or mean structure model with equality constraints we can test for configural invariance. If it is not possible to demonstrate configural invariance, this indicates that the loading structures differ across groups. Therefore, further tesing of measurement invariance would not be necessary. In fact, finding evidence of configural invariance can be considered a prerequisite for further testing in measurement invariance. The covariance and the mean structures of the configural model are:

$$\mathbf{\Sigma}_{xx}^{(z)} = \mathbf{\Lambda}_x^{(z)} \phi^{(z)} \mathbf{\Lambda}_x^{(z)} + \boldsymbol{\psi}_\delta^{(z)} \tag{2.1}$$

$$\boldsymbol{\mu}^{(z)} = \boldsymbol{\tau}_x^{(z)} + \mathbf{\Lambda}_x^{(z)} F^{(z)} \tag{2.2}$$

Notice that in equations 2.1 and 2.2 there is a z superscript placed within the model. The z is what allows the loading, intercepts, and uniquenesses to vary freely across groups. The only restriction placed on the configural model is that the loading pattern is assumed identical. However, this is not to be interpreted as the loadings being the same, only their pattern.

2.2.2 Weak Invariance

There are few differences between configural invariance and weak invariance. In fact, if we look at the model below, it seems almost identical to the configural invariance model. However, there are some subtle differences that make the weak invariance model different

from that of the configural invariance model. If you notice the (z) superscript has been removed from the factor loadings in both the covariance and mean structure models. The reason it is removed is so that the factor loading are proportionally equal across groups. This ensures an equal relationship between the latent factor and the indicator items in the model (Dimitrov, 2010). Note that there are occasions when such parameters $\boldsymbol{\tau}_x^{(z)}$ and $F^{(z)}$ can not all be estimated without further constraints on the model. Therefore, it is possible to omit the mean structure model from the weak invariance test and work only with the covariance structure model.

$$\boldsymbol{\Sigma}_{xx}^{(z)} = \boldsymbol{\Lambda}_x \phi^{(z)} \boldsymbol{\Lambda}_x + \boldsymbol{\psi}_\delta^{(z)} \tag{2.3}$$

$$\boldsymbol{\mu}^{(z)} = \boldsymbol{\tau}_x^{(z)} + \boldsymbol{\Lambda}_x F^{(z)} \tag{2.4}$$

However, if weak invariance is established it is possible to compare the mean score across groups. If weak invariance is not established any difference in the mean score between groups may be due to the difference in the loadings. Note: The z superscript has been removed from the factor loadings in both the covariance and mean structure models.

2.2.3 Strong Invariance

Strong invariance increases the group equivalence to include equal loadings as well as equal intercepts in both the covariance and mean structure models. In order to set this equality restriction we must remove the superscript z from $\boldsymbol{\Lambda}_x^{(z)}$ and from $\boldsymbol{\tau}_x^{(z)}$ (intercept). The resulting model will be as follows:

$$\boldsymbol{\Sigma}_{xx}^{(z)} = \boldsymbol{\Lambda}_x \phi^{(z)} \boldsymbol{\Lambda}_x + \boldsymbol{\psi}_\delta^{(z)} \tag{2.5}$$

$$\boldsymbol{\mu}^{(z)} = \boldsymbol{\tau}_x + \boldsymbol{\Lambda}_x F^{(z)} \tag{2.6}$$

Note that once strong invariance is established, observed differences in the mean score would be due to differences in the latent dimension.

2.2.4 Strict Invariance

The strict level of MI adds an equality constraint on the error structures of the CF-models. The restrictions from the previous levels of invariance (weak & strong) carry over onto strict invariance as well. Therefore, within strict invariance the factor loadings, intercepts, and the uniquenesses (error variance) is assumed equal across groups. In the past, researchers have found that testing for this high level of invariance is unnecessary. Some hold the view that the model is far too restricted at this level and no "real" data attains this level of invariance. So it was presumed that it would only be necessary to test for weak and strong invariance (Wu, Li, & Zumbo, 2007). However, arguments have been made stating that testing for strict invariance is a necessary condition for a fair and impartial comparison (Meredith, 1993). Thus, this level is considered relevant for this study. The resulting model is as follows:

$$\mathbf{\Sigma}_{xx}^{(z)} = \mathbf{\Lambda}_x \phi^{(z)} \mathbf{\Lambda}_x + \boldsymbol{\psi}_\delta \tag{2.7}$$

$$\boldsymbol{\mu}^{(z)} = \boldsymbol{\tau}_x + \mathbf{\Lambda}_x F^{(z)} \tag{2.8}$$

2.3 Fit Statistics

As discussed in previous sections, it has become clear that it is not particularly appropriate to analyze a single model, but in fact, it is more appropriate to analyze multiple nested models. Acquiring these nested models will allow us to compare them to other nested models. The chi-square test statistic given in section 1.2.1 analyzes the difference between the observed covariance matrix and the model covariance matrix. The chi-square test statistic provides a foundation that our tests for MI are based on. Before exploring the details of the MI test the null and alternative hypothesis of the MI tests are reviewed. In all testing, the null hypothesis is that the nested model fits the data well. The alternate hypothesis states the complement.

Ho: Level of invariance in the nested model holds

H1: Level of invariance in the nested model does not hold

Once the chi-square statistic has been derived it is easy to derive the chi-square difference test. The chi-square difference test also follows a chi-square distribution with the degrees of freedom of the difference between the competing models degrees of freedom (Bain & Engelhardt, 2000).

$$\boldsymbol{\chi}^2_{diff} = -2(ln(\boldsymbol{\Lambda}_{null}) - ln(\boldsymbol{\Lambda}_{alternate})) \tag{2.9}$$

Above we can see how the chi-square difference test comes from the likelihood ratio test however, to simplify the purpose behind the test let us look at a simpler form of the test.

$$\boldsymbol{\chi}^2_{diff} = \boldsymbol{\chi}^2_N - \boldsymbol{\chi}^2_F \tag{2.10}$$

In this form we can see that the "N" subscript denotes the nested model which has more degrees of freedom yet it is the model which is more restricted than the full model. Next, the "F" subscript denotes the full model test. This chi-square statistic depicts the model with the highest level of restrictions. In the case of MI we can view the full model as configural, weak, or strong invariance. It all depends on what comparison is desired. For example we could compare configural(F) to weak (N) or we could compare weak (F) to strong(N).

When we compare these two models, we are assessing the lack of fit due to the equality constraints placed on the model. Therefore a small p value (p $\leq$ .05) indicated that the model constraints significantly reduce the fit of the model across groups. Conversely, a large p-value (p $\geq$ .05) provides evidence that the equality constraints hold across the populations and therefore we can not reject our null hypothesis. By not rejecting the null hypothesis we are able to assume that the nested model holds and that the level of invariance holds.

Goodness of Fit Indexes

In the past, the usefulness of the chi-square difference test has been questioned (Brannick, 1995). It is clear how chi-square is affected by sample size and/or the model complexity. In particular the chi-square test is overly conservative. In other words it rejects the null hypothesis too often which causes rejections of reasonable models (Wu et al., 2007). It would be irresponsible to solely base model decisions on one specific statistic. Thus, a variety of fit indexes with cut-off criteria have been developed in order to provide a solution to this problem (Wu et al., 2007). Such fit indexes include the Root Mean Square Error of Approximation (RMSEA), Comparative Fit Index (CFI), and the Bayesian Information Criterion (BIC) (Kenny, 2003).

The RMSEA is not directly affected by the models complexity. Using the RMSEA, is considered a good fit when the RMSEA is less than .06 (Chen, Curran, Bollen, Kirby, & Paxton, 2008). If the model shows a value greater than .10 than it would be considered a poor fit. The RMSEA formula is as follows:

$$RMSEA = \sqrt{((\chi^2/df) - 1)/(N - 1)} \tag{2.11}$$

Where df is the degrees of freedom and N is the sample size. Due note that when the chi-square is less than the degrees of freedom the RMSEA is set to zero. For MI testing, the differences of the competing model RMSEA's are observed. The accepted rule of thumb is that a difference of 0.01 between RMSEA's indicates a significant lack of fit.

The CFI is another fit index useful for assessing model fit. When the CFI shows a result greater than .95 or .90 then the model is a good fit however, if the CFI shows a result less than 0.9 than there is a problem with the model (Wu, 2007). The CFI is as follows:

$$CFI = 1 - \frac{(\chi^2 - df(Alt.model))}{(\chi^2 - df(Nullmodel))} \tag{2.12}$$

when testing for MI a difference in the CFI mode than 0.01 indicates a significant reduction in fit.

Finally let us review the Bayesian Information Criterion (BIC). The BIC is yet another criterion used for model selection. Much like the other criterion's the BIC corrects for a large sample size by penalizing the model in a way to regulate it. When you try to increase the likelihood by adding parameters and therefore adding to your sample size it may result in over fitting of the model. The BIC corrects for such problem. There is no specific cut off point for the BIC because it is mainly used as a comparison between the two models. The BIC is as follows:

$$BIC = \chi^2 + [p(p-1)/2 - df] \ln(N) \tag{2.13}$$

where p is the number of free parameters and the $\ln(N)$ is the natural logarithm of the sample size (Kenny, 2003).

Chapter 3

Testlets in CFA

3.1 Testlets

In CFA models, items may exhibit explainable covariation not related to the latent factors of interest. These clusters of items are commonly referred to as testlets. Testlets often originate from a common stimulus and are sometimes formed intentionally. There are many reasons why researchers would group items into testlets. One reason is to reduce the amount of covariation exhibited due to the complexity of the items not directly related to the latent construct(s). For example, another reason testlets are used is to allow a participant to utilize his time efficiently by working off a common paragraph. These groupings of items are commonly used in standardized tests due to the length of the exams. Specifically in standardized tests, items are commonly grouped within the passage in which they belong. Participants are often asked to read a set of passages and then answer several related questions pertaining to the passage. By grouping these questions to a specific passage, it allows for a quicker response from the participant and eliminates any confusion between the passages that would occur if the questions had not been grouped in content specific testlets (DeMars, 2006). Testlets also appear in psychological and educational scales when items have a common theme or wording. However, problems due arise when grouping the items into testlets.

Items within testlets tend to create covariation among items not directly attributable to the latent factor (DeMars, 2006). Since testlets are grouped as context specific items they can create a secondary factor possibly not related to the primary factor(s). This generally tends to happen due to the participants prior knowledge, specific skills, or a higher level of

interest relating to the topic in question. Researchers often refer to this secondary factor as a nuisance factor due to the fact that they have no desire in scoring this factor. For illustration, an example of data with a subset of items belonging to a testlet is presented.

3.1.1 Testlet Example

Allow us to look at the following example involving the model development of the General Social Anxiety Scale (Brown, 2006). Figure 3.1 represents a six item test where it believed that each item loads onto a single factor. The six items and factor are as follows:

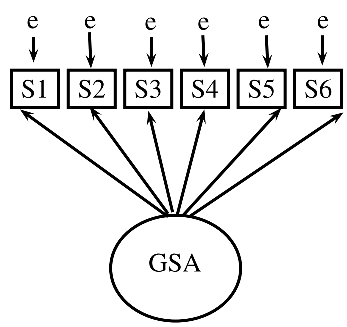

Figure 3.1: Single Factor

- S1: making eye contact
- S2: maintaining a conversation
- S3: meeting strangers
- S4: speaking on the telephone
- S5: giving a speech to a large group
- S6: Introducing yourself to large groups
- Single Factor: General Social Anxiety (GSA)

In this example, the researcher is interested in examining the latent dimensionality of GSA. In other words, this study intends to validate that the scale measures GSA through these six items all of which measure anxiety in social situations. The researcher finds that although all six items correspond to GSA items S5 and S6 tend to lean more towards anxiety in public speaking situations. Due to the fact that items S6 and S7 have a much stronger correlation with public speaking anxiety (PSA) rather than GSA it causes a lot of added covariation within the model. This added covariation to model leads the researcher to find a poor fit for the model (i.e. the model is not a good tool for predicting GSA). However, the researcher finds that by modeling the testlet with items S5 and S6, the researcher could avert any problems with poor model fit (Brown, 2006). Figure 3.2 represents the two factor solution assuming items S5 and S6 are formed into a testlet.

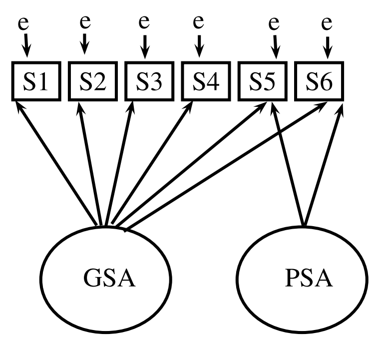

Figure 3.2: Two Factor Model (factor covariance is fixed to zero)

3.2 Correcting Issues with Testlets

As discussed earlier, grouping items into testlets can cause several issues when analyzing the data. Models that accommodate the presence of testlets play an important role in properly measuring test performance, anxiety levels, or even intelligence levels. Bias due to the presence of testlets can obscure the true latent structure of scales and it is often difficult to compensate for such bias. Modeling the testlets allows the researcher to minimize this problem. In order to compensate for this added covariation we may utilize different model

structures. Each model structure has shown a significant improvement to the model when testlets are present.

3.2.1 Correlated Traits Correlated Methods Minus One Model

The correlated traits-correlated methods minus one model (CT-C(M-1)) was developed to address the case when additional covariance between items exist that is independent of the latent structure. In the CT-C(M-1) approach, all the items relate to a factor while a specific set of items relate to a nuisance factor. In the CT-C(M-1) model, the factor which relates to all the items does not correlate with the nuisance factor (orthogonal).When applying the CT-C(M-1) model to data we must consider the testlets as the nuisance factor that is not correlated with the primary factor.

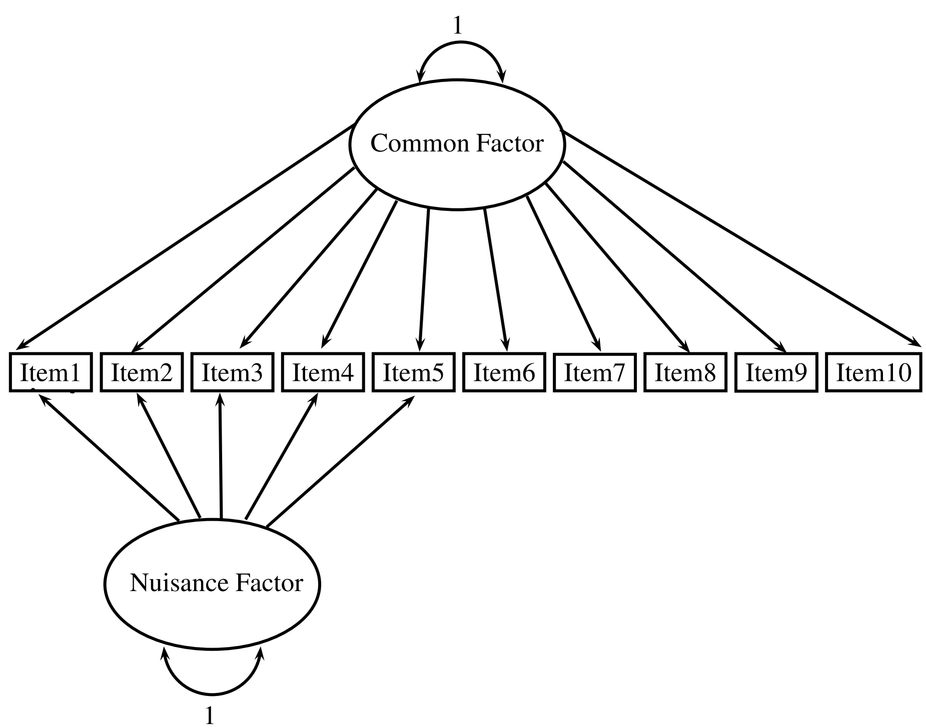

Figure 3.3: CT-C(M-1) Model

3.2.2 Correlated Error Model

It is common in most factor models for the researcher to not correlate the errors within the model. This assumes that all the items load on a factor due to its latent dimension. The uncorrelated error approach appears to work well when dealing with no testlet structure. In the case of having testlets within the model, it is then acceptable to use a correlated error model to compensate for the added covariation. By correlating the measurement errors within the testlet, covariance between the items within the testlet is explained independent of the primary factor. Note that only the testlets will contain the correlated errors and not the items outside the testlets.

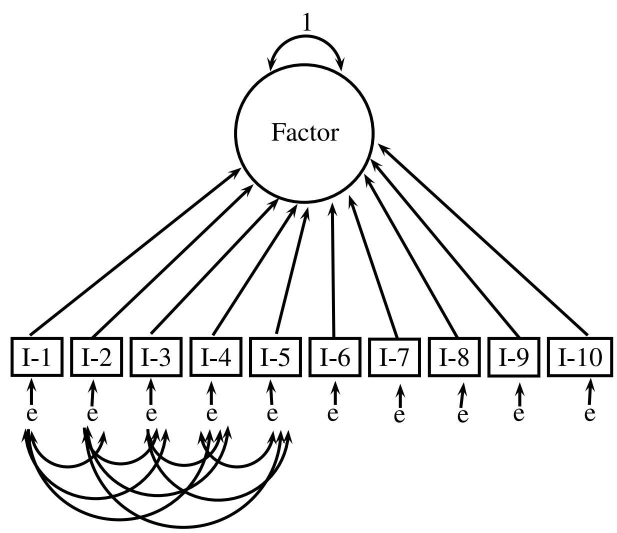

Figure 3.4: Correlated Error Model

3.2.3 Random Intercepts model

The random intercept model is commonly used in CFA as a possible solution for modeling testlets. The random intercept model is used to account for this issue. It does so by allowing the intercepts in the factor model to no longer be fixed but to vary randomly across subjects. In other words it allows the intercept to vary from respondent to respondent (Olivares &

Coffman, 2006). The following factor model demonstrates how the intercept is interpreted in the random intercept model. A random intercept common factor model is

$$\boldsymbol{x}_i = \boldsymbol{\tau}_i + \boldsymbol{\Lambda}_i * F + \boldsymbol{\varepsilon}_i \qquad i = 1, ..., N \tag{3.1}$$

$$\boldsymbol{\tau}_i = \boldsymbol{\mu} + 1 * \boldsymbol{\zeta}_i \tag{3.2}$$

Where $\boldsymbol{\mu}$ (A x 1) and $\boldsymbol{\zeta}$ is common to all items and varies randomly from item to item, "N" is the number of subjects and "1" is a vector while $\boldsymbol{\zeta}_i$ is allowed to vary from respondent to respondents and is common to all items. $\boldsymbol{\zeta}$ is not directly estimated in the analysis rather, the variance $\boldsymbol{\phi}$ is estimated. Like the CF-model there are certain assumptions that need to hold within the random intercepts model:

- The mean of the random intercept is zero and the variance is $\boldsymbol{\phi}$.
- The intercept $\boldsymbol{\zeta}$ is uncorrelated with the error items.
- The intercept $\boldsymbol{\zeta}$ is also uncorrelated with the common factor.

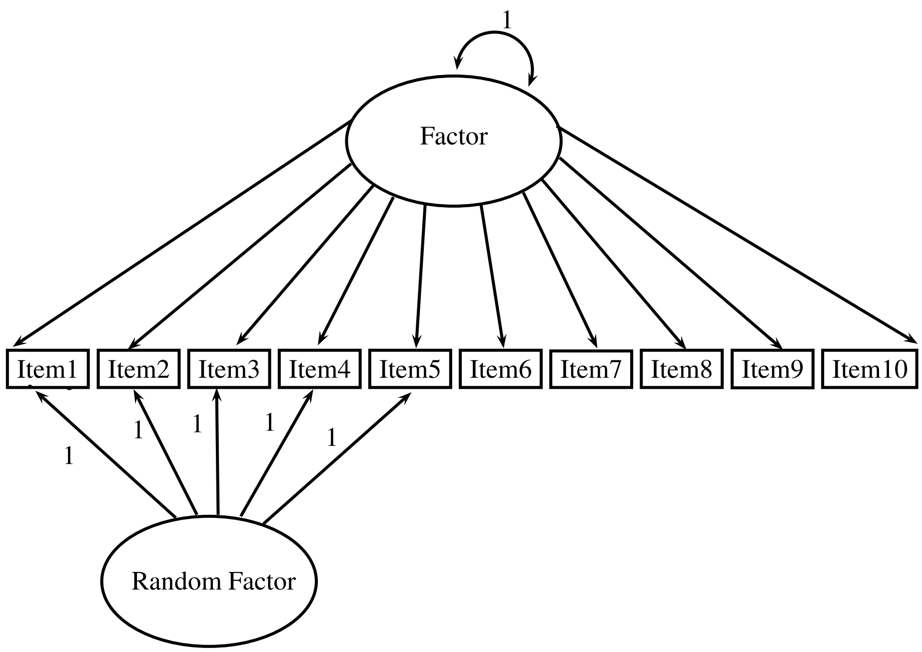

Figure 3.5: Random Intercept

3.3 Testing for Measurement Invariance in the Presence of Testlets

In research, when dealing with multiple populations, there is reason to believe that one or more of the populations may have data consistent with an underlying factorial structure but exhibiting covariation attributable to a testlet structure. Much of the time, the testlet structure is ignored. However, by doing so, it causes many issues when trying to detect MI. If there is a problem in detecting MI due to a testlet structure in the data, then a scale may be improperly assumed to not function across populations.

All of the models which have been discussed above have been proposed as a solution for detecting MI in the presence of a testlet structure. However the task is to find out which of the proposed solutions best detect MI when it exists. To answer this question a number of simulations will be run. Each of these cases model a specific level of MI while also simulating a testlet structure.

Research Question: When testing for MI in the presence of testlets which model is most appropriate? Five scenarios are examined for testing MI when testlets are present in one or both populations.

- Case 1 (Strong MI): Two populations with identical loading structure, but only one exhibits extraneous covariation due to testlets.
- Case 2 (Strict MI): Two populations with identical loading structure, but both both populations exhibit extraneous covariation due to testlets.
- Case 3 (Configural MI): Two populations with differing loading structure and additional testlet-based covariation in both populations.
- Case 4 (Weak MI): Two populations with identical loading structure but varying intercepts. Both populations have testlets.

Chapter 4

Simulation Results

4.1 Introduction

When testlets are present it could affect the ability to detect measurement invariance in CFA. When such an issue occurs, the researcher may lack the ability to detect vital measurement differences that could completely change the outcome of the study. Therefore it is necessary to run a set of simulations that would help determine how testlets affect a CF-model and how it plays a role in detecting measurement invariance.

4.2 Simulation Setup

In order to simulate the different types of cases, the simulations are based on the common factor model described in chapter 1 (see eq. 1.1). From this response model the appropriate data for each population is generated and the covariance and mean structure model to test for the presence of MI is extracted. Based on the response model, (eq. 1.1) it is evident that four different parameters ($\boldsymbol{\tau}$, $\boldsymbol{\Lambda}$, $\mathbf{F}$, and $\boldsymbol{\varepsilon}$) are to be generated.

These simulations will run in the "R" programming environment and using the lavaan package within R (Rosseel, 2011). This package is used for latent variable analysis and has a built in function for CFA and structural equation modeling which will be needed in order to create the different models proposed in chapter 3.

Prior to deciding how to generate the different parameters within the response model the basic structure of the model should be determined. In order to have a valid CF-model, no less than three items should load on an individual factor (McDonald, 1999). Since there

are two populations and at times either one or both populations will contain a testlet based structure, it is prudent to have at least k=10 items loading onto a factor. As for sample size, the general rule in CFA and in most forms of analysis is that the larger the sample size the more accurate the simulation results. Therefore a sample size of n=500 is sufficient for the simulations. Sample sizes of 250 and 100 are also run for comparison. Note that two response models are generated in order to compare the measurement differences between two populations.

Most researchers consider a weak factor loading to be between 0.4 and 0.6; however, in actual applied research, a weak factor loading size to be between 0.2 and 0.3. In a Monte Carlo simulation study regarding loading sizes in CFA, it was found that 0.5 is considered a moderate leading size while 0.7 would be considered a high loading size (Ximenez, 2006). In accordance with (Meade & Lautenschlager, 2004) and keeping a "moderate" and "high" level of correlation, these simulations utilize 0.6 and 0.8 for the loadings. The factor score and random error of the model are generated to be standard normal. To ensure, independence of the population scores, two iid N(0,1) score vectors are utilized in all but the last scenario of the simulations. In all models, except case IV, the intercept is set to zero.

In order to add covariation due to testlets or both of the populations, we adjust the uniqueness or error of the factor analysis model are adjusted so that,

$$\boldsymbol{\varepsilon}^* = \boldsymbol{\Lambda}_b * F + \boldsymbol{\varepsilon} \tag{4.1}$$

where $\boldsymbol{\Lambda}_b$ = (0.4,0.4,0.4,0.4,0.4,0,0,0,0,0) and $\mathbf{F} \sim$ iid N(0,1).

In the following sections, the simulation results are discussed for each of the four scenarios described in section 3.3.

4.3 Results

In chapter 3, a set of models were discussed that compensate for the added covariation in a model when a testlet structure is present. The models proposed were the CT-C(M-1)

model, correlated error model, and the random intercept model. The simulations will also evaluate, the single factor model (naive model). The single factor model will not correct for the added covariation from the testlet structure, but yields insight into the effect of testlets in MI when not modeled appropriately.

In order to test MI in the presence of testlets, the simulations will be evaluated at four different forms of invariance (configural, weak, strong, and strict). The fit measures RMSEA, CFI, p-value and BIC will be used to test for model fit. The difference between the following models will also be used to determine fit.

- Configural v. Weak
- Weak v. Strong
- Strong v. Strict

This is done in order to detect the level of invariance holds for that specific set of results.

4.3.1 Case 1: Strong Measurement Invariance

In case 1 there are two populations with identical factor structures. However, the testlet are only added to one of the populations. Therefore, five of the items will contain the added testlet covariation while the others do not. These populations exhibit a strong level of MI since the factor loadings and intercept are equivalent across populations. The following results are for the case when N=500, at k=10 items and $\boldsymbol{\lambda}$=0.6.

Table 4.1: Measurement Invariance Results

Naive Model (λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.058	0.00	0.973	21375.94	193.637	70
Weak	.059	0.00	0.968	21341.69	221.55	79
Strong	.0547	0.00	0.9708	21418.17	222.04	88
Strict	.0524	0.00	0.9701	21362.26	235.219	98

Table 4.2: Measurement Invariance Results

Correlated Error Model (λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0098	0.487	0.9987	21375.11	51.01	50
Weak	0.0045	0.675	0.999	21313.47	51.5	59
Strong*	0.00167	0.819	0.999	21389.81	51.89	68
Strict	0.0241	0.101	0.994	21372.72	103.88	98

Table 4.3: Measurement Invariance Results

CT-C(M-1) Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0078	0.547	0.9988	21303.15	58.37	60
Weak	.0045	0.676	0.99931	21219.47	64.5	73
Strong	.00215	0.7966	0.999682	21302.37	64.51	81
Strict*	.00075	0.8903	0.999889	21234.14	65.35	91

Table 4.4: Measurement Invariance Results

Random Intercepts Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0087	0.5098	0.998	21257.37	68.11	68
Weak	0.0044	0.667	0.999265	21195.76	68.67	77
Strong	0.0022	0.7883	0.999632	21278.66	68.67	85
Strict*	0.0009	0.8851	0.999851	21210.11	69.203	95

Given Table 4.1, it is evident that no level of MI holds when using the naive model. Thus, the naive model does not correctly conclude the strong level of MI. In contrast, the correlated error model on average correctly identifies the strong level of MI. The CT-C(M-1) and random intercept models on average yield the result that the strict level of MI holds.

Table 4.5: case 1,n=100

	CFI	p-value	RMSEA
Strict	0.005	0.00	0.00
Strong	0.999	1.00	1.00

Table 4.6: case 1,n=500

	CFI	p-value	RMSEA
Strict	0.067	0.09	0.00
Strong	0.933	0.910	1.00

Table 4.5 and 4.6 describe the number of times a certain model was chosen when utilizing the correlated error model. Initially it was hypothesized that case 1 would show a strong

level of MI. Therefore, 4.5 and 4.6 show the proportion of times the results indicated either strict or strong invariance. Note that the chi-square test correctly picks the strong level of invariance 91 % of the time when N=500. For this case, most of the differences observed in the RMSEA were greater than 0.01 but less than 0.02.

4.3.2 Case 2: Strict Measurement Invariance

Case 2 scenario consists of two independent populations. However, now two independent testlet structures are incorporated into both populations. Those populations exhibit strict MI since they are equivalent with respect to the intercept, loading and error structure.

Table 4.7: Measurement Invariance Results

Naive Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0834	0.00	0.9477	21891.34	315.9	70
Weak	0.077	0.00	0.9493	21830.88	317.7	79
Strong	0.071972	0.00	0.951	21907.59	318.4242	88
Strict	0.0669	0.00	0.9528	21840.15	320.06	98

Table 4.8: Measurement Invariance Results

Correlated Error Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0098	0.49	0.9987	21770.52	50.9845	50
Weak	0.0045	0.67	0.9994	21709.06	51.69	59
Strong	0.0018	0.82	0.9997	21785.57	52.22	68
Strict*	0.00075	0.90	0.9999	21718.98	54.699	78

Table 4.9: Measurement Invariance Results

CT-C(M-1) Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0086	0.52	0.99872	21697.9	59.7495	60
Weak	0.0037	0.71	0.99947	21610.68	35.3285	73
Strong	0.001578	0.83	0.99978	21693.59	62.34	81
Strict*	0.000494	0.91	0.999926	21625.04	62.87935	91

Table 4.10: Measurement Invariance Results

Random Intercepts Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.008584	0.51	0.99866	21653.52	67.96	68
Weak	0.004361	0.67	0.99931	21592.06	68.67	77
Strong	0.0021	0.79	0.9996	21674.95	68.67	85
Strict*	0.0008	0.89	0.9998	21606.38	69.17	95

Given Table 4.7, it is clear that no level of MI holds using the naive model. In contrast, Tables 4.8 - 4.10 show that the correlated error, CT-C(M-1), and random intercept models on average correctly identify the strict level of invariance in MI.

Table 4.11: case 2(n=100)

	CFI	p-value	RMSEA
Strict	0.99	1.00	1.00
Strong	0.01	0.00	0.00

Table 4.12: case 2(n=500)

	CFI	p-value	RMSEA
Strict	1.00	1.00	1.00
Strong	0.00	0.00	0.00

Tables 4.11 and 4.12 describe the number of times a certain model was chosen when utilizing the correlated error model. Initially it was hypothesized that case 2 would show a strict level of MI. Therefore, Tables 4.11 and 4.12 show the amount of times the results indicated either strict or strong invariance. Each table shows the RMSEA, CFI, and p value results at N=100 and N=500. Note that the strict level was appropriately identified in almost every case when utilizing the correlated error model.

4.3.3 Case 3: Configural Measurement Invariance

In case 3, the testlet structure is added to both populations. The new latent structure will be as follows: $\boldsymbol{\Lambda}$= ($\boldsymbol{\lambda}_{i1}$,$\boldsymbol{\lambda}_{i2}$, **0.3**, **0.3**, $\boldsymbol{\lambda}_{\boldsymbol{i5}}$, $\boldsymbol{\lambda}_{\boldsymbol{i6}}$, **0.3**, **0.3**, $\boldsymbol{\lambda}_{i9}$, $\boldsymbol{\lambda}_{i10}$). As a result, these populations should exhibit MI only to the configural level.

As in the previous scenario, not all items load onto the factor with the same loading size. However, it is important to determine if in fact changing the loading size to 0.3 make a difference the analysis.

Table 4.13: Measurement Invariance Results

Naive Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.008292	0.51534	0.99847	20817.86	70.02	70
Weak	0.02867	0.06203	0.99137	20800.78	115.11	79
Strong	0.0228	0.1301	0.9934	20876.82	115.167	88
Strict	0.017032	0.2339	0.9952	20808.78	116.204	98

Table 4.14: Measurement Invariance Results

Correlated Error Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural*	0.00997	0.4877	0.9985	20941.6	51.037	50
Weak	0.03064	0.06348	0.9926	20918.27	89.87	59
Strong	0.023102	0.148	0.99463	20994.31	89.93	68
Strict	0.01769	0.242272	0.999587	20928.98	93.67	78

Table 4.15: Measurement Invariance Results

CT-C(M-1) Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural*	0.00734	0.56854	0.9988	20871.13	57.413	60
Weak	0.02601	0.10134	0.9931	20825.41	101.49	73
Strong	0.02011	0.18813	0.9948	20908.33	101.52	81
Strict	0.0141	0.31616	0.9964	20840.6	102.87	91

Table 4.16: Measurement Invariance Results

Random Intercepts Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural*	0.00879	0.5096	0.9984	20824.9	68.15	68
Weak	0.02643	0.0883	0.99264	20802.23	107.6	77
Strong	0.02086	0.1653	0.9944	20885.14	107.668	85
Strict	0.01511	0.283	0.9961	20817.37	108.97	95

Given Table 4.13, it is evident that again no level of MI holds for the naive model. The correlated error, CT-C(M-1), and random intercept model show that the configural level of MI holds for Tables 4.14 - 4.16.

Table 4.17: case 3,n=100

	CFI	p-value	RMSEA
Weak	0.15	0.27	0.00
Configural	0.85	0.73	1.00

Table 4.18: case 3,n=500

	CFI	p-value	RMSEA
Weak	0.083	0.008	0.00
Configural	0.917	0.992	1.00

Tables 4.17 and 4.18 describe the number of times a certain model was chosen when utilizing the correlated error model. Initially it was hypothesized that case 3 would show a weak level of MI. Therefore, Tables 4.17 and 4.18 show the proportion of times the results indicated either weak or configural invariance. Each table shows the RMSEA, CFI, and p value results at N=100 and N=500.

4.3.4 Case 4: Weak Measurement Invariance

In the previous cases the intercept parameter was always assumed to be zero (therefore equal across both populations). This scenario investigates what changes when the intercept is varied across populations. In order to simulate a change in the intercept with practical significance, τ is set to 0.8. The tau intercept is then added to one of the response models. Consequently, these populations should exhibit weak invariance.

Table 4.19: Measurement Invariance Results

Naive Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0834	0.00	0.948	21890.31	315.87	70
Weak	0.0783	0.00	0.9481	21835.93	323.665	79
Strong	0.078443	0.00	0.94218	21949.01	360.7551	88
Strict	0.0732	0.00	0.943842	21881.93	362.75	98

Table 4.20: Measurement Invariance Results

Correlated Error Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	.00989	0.4864	0.998763	21769.05	50.99	50
Weak	0.0085	0.514039	0.998843	21714.63	58.74362	59
Strong	0.0053	0.6198	0.999245	21794.65	62.77249	68
Strict*	0.00266	0.748059	0.999626	21728.36	65.56	78

Table 4.21: Measurement Invariance Results

CT-C(M-1) (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0088	0.51374	0.998716	21701.99	59.96	60
Weak	0.0065	0.581	0.9989	21621.64	69.42	73
Strong	0.0043	0.6656	0.9993	21707.83	72.72	81
Strict*	0.0017	0.8034	0.99973	21639.51	73.47	91

Table 4.22: Measurement Invariance Results

Random Intercepts Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0086	0.51237	0.9986	21653.22	67.96	68
Weak	0.0074	0.542	0.998	21598.47	75.385	77
Strong	0.0051	0.629	0.99912	21684.51	78.53	85
Strict*	0.00232	0.769	0.9996	21616.19	79.28	95

Given Table 4.19, it is clear that no level of MI holds in the naive model. Consequently, the correlated error, CT-C(M-1), and random intercept model yield the result that the strict level of MI holds when $\mathbf{\Lambda}$=0.6.

Table 4.23: case 4,n=100

	CFI	p-value	RMSEA
Weak	0.999	0.999	1.00
Configural	0.001	0.001	0.00

Table 4.24: case 4,n=500

	CFI	p-value	RMSEA
Weak	1.00	0.999	1.00
Configural	0.00	0.001	0.00

Tables 4.23 and 4.24 describe the number of times a certain model was chosen based off the nested model hypothesis test. Initially it was hypothesized that case 4 would show a weak level of MI. Therefore, Tables 4.23 and 4.24 show the amount of times the results indicated either weak or configural invariance. Each table shows the RMSEA, CFI, and p value results at N=100 and N=500.

Chapter 5

Discussion

When testing nested models one must take several factors into consideration when analyzing the data. To begin with, one must be able to evaluate a nested model based off a hypothesis test. Recall the hypothesis discussed in chapter 2:

Ho: Level of invariance in the nested model holds

H1: Level of invariance in the nested model does not hold

Accepting or not accepting the hypothesis based off the p-values of the difference between the models will help determine which model fits best. Evaluation of RMSEA, CFI, BIC, and χ^2 must also be taken into consideration when trying to interpret the results. The end goal is to choose the most parsimonious model. In other words, when trying to decide which level of MI fits best it is unrealistic to expect one model to show the "best fit" across all the fit statistics. Therefore, one must determine the best model fit based on the overall results and deciding which is the most parsimonious model (McDonald, 1999).

5.1 Case 1: Strong Measurement Invariance

Given Tables A.1-A.8 in Appendix A it is clear that in the naive model, all fit statistics indicated a poor fitting model for $\boldsymbol{\lambda} = 0.6$ or 0.8. Thus, no level of invariance holds when using the naive model. This is expected given the fact that the naive model was never intended to correct for the added covariation due to the testlet structure. In contrast, you can clearly see that the correlated error model on average identifies the strong level of MI. In Tables A3 and A4 you can clearly see a low RMSEA value (less than .06) and high CFI value (greater than .98) (Chen et al., 2008). All fit statistics but the BIC value appear to indicate that the correlated error model identifies with the strong level of MI for both $\boldsymbol{\lambda} = 0.6$ and 0.8. Note that in both cases of lambda, the CFI value at the strong level of MI was not indicative as being the best fit. However, given all the other fit statistics and comparing the difference with the better CFI value we can still determine that the differences are minimal and that there is overwhelming evidence that strong MI holds.

The CT-C(M-1) and random intercept model both indicate that strict level of invariance holds for both cases of lambda ($\boldsymbol{\lambda} = 0.6$ or 0.8). Again this result comes from identifying the low RMSEA value and the high p and CFI values. All of which indicate a good model fit for the strict level of MI. Each table shows the RMSEA, CFI, and p value results at n=100 and n=500. In summary, the correlated error model best picks the correct level of MI for this scenario.

5.2 Case 2: Strict Measurement Invariance

As in case 1, the fit statistics for the naive model indicate a poor model fit for both cases of lambda (see Table A.9 and A.10). Therefore, no level of invariance holds when using the naive model. By attempting to correct for the added covariation to the model you can clearly see an improvement in the fit statistics throughout the simulation. In the correlated error, CT-C(M-1), and random intercept model it is immediately clear that the data fits much better when it is modeled using those three approaches (see Table A.11 - A.16). However, to determine what level of MI each model holds it is necessary to analyze the fit statistics more closely. After looking at the RMSEA, CFI, p-value, and $\boldsymbol{\chi}^2$ it is evident that in the correlated error, CT-C(M-1), and random intercept model strict MI holds for both cases of lambda ($\boldsymbol{\lambda} = 0.6$ or 0.8). This is determined by seeing that the RMSEA value is lowest at the strict level while the p-value and CFI values are larger at the strict level of MI. However, as in case 1 the BIC level is not at its smallest within the strict level of MI. It appears to be smallest at the weak level of MI for both cases of lambda in all three models. Even though the BIC value does not agree with the other fit statistics it is still not enough evidence to conclude weak. Therefore, strict MI would still be concluded such that the factor loadings, intercept, and uniquenesses are equivalent across both populations. Finally in summary, all three testlet models correctly identify the level of MI in this case.

5.3 Case 3: Configural Measurement Invariance

As in the previous cases, the fit statistics for the naive model indicates a poor model fit for both cases of lambda (see Table A.25 and A.26). Yet again, no level of MI holds when using the naive model. Recall that in case 3 the loading structure was varied so that items 3, 4, 7, and 8 varied their loading value. Specifically, they would load onto the factor at $\boldsymbol{\lambda} = 0.3$ and testlets are still found in both populations. This significantly affected the results of the simulation. In all the cases (excluding the naive model) the configural level

of MI was found in both cases of lambda ($\boldsymbol{\lambda} = 0.6$ or 0.8). Again, this was determined by analyzing the fit statistics and choosing the most parsimonious model. The RMSEA level showed a significantly better fit than the other levels of MI, as did the p-value statistic. The CFI and BIC level were not the best in comparison with the other levels of MI. However; they still appeared to be a good foot for the data at the configural level of MI. Since the configural level of MI is considered the "best" fit therefore, no further analysis is required. In summary, all competing models appear to identify a change in fit when moving from the configural to weak level of measurement invariance.

5.4 Case 4: Weak Measurement Invariance

As in the previous cases, the fit statistics for the naive model indicate a poor model fit for both cases of lambda (see Table A.33 and A.34). No level of MI holds when using the naive model. In case 4 we are still dealing with two populations each of which, has a testlet structure. What makes case 4 different is that an intercept is added to the model. Consequently, by adding this intercept, the populations should exhibit weak invariance. However, after looking at the data it appears that the strict level of MI holds across both populations. As before, it was possible to determine this level of invariance by analyzing the fit statistics. The RMSEA, CFI, and p-value showed a good for the strict level of MI (see Tables A.35 - A.40). The BIC value again appears not to be the best fit at the strict level of MI. However, it is not enough to affect the strict level of MI that was chosen.

The failure of any model to correctly identify the weak level of MI may be explainable by investigating the covariance structure of the data consistent with a model with correlated errors. Recall the CF-model

$$x = \tau + \lambda F + \varepsilon. \tag{5.1}$$

When the covariance of ε is allowed to be a non zero on the off diagonal, it is possible to replace the typical model with eq. 4.2 where the testlet structure was added to the response

model. Now the typical factor model can be replaced with

$$\boldsymbol{x} = \boldsymbol{\tau} + \boldsymbol{\lambda} F + \boldsymbol{\lambda}_2 F_2 + \boldsymbol{\varepsilon}^* \tag{5.2}$$

where F_2 is a k_2 x 1 vector of additional attributes with mean 0 and variance 1, $\boldsymbol{\lambda}_2$ is a n x k_2 matrix of discriminabilities and the $\boldsymbol{\varepsilon}^*$ are iid item errors with var($\boldsymbol{\varepsilon}^*$) = $\boldsymbol{\theta}^*$. Observe that the

$$cov(F, F_2) = \begin{bmatrix} 1 & 0 \\ 0 & \psi_2 \end{bmatrix}$$

with ψ_2 a k_2 x k_2 matrix. This implies that the model has a dimension greater than k because of the additional variation contributed by F_2 and var($\boldsymbol{\varepsilon}$) = $\boldsymbol{\lambda}_2\ \boldsymbol{\theta}^*\ \boldsymbol{\lambda}' + \boldsymbol{\theta}^*$.

The correlated error, CT-C(M-1) and random intercept models all appropriately account for this added heterogeneity in the model as described. Apparently, in the simulations, the intercept which gave a change in location with "practical" significance, is too small in comparison to the added change in location due to the "additional" attributes. Future simulations could investigate greater changes in the intercept in order to determine the size with "statistical" significance.

5.5 M.I. Example

A great example of being able to detect MI in the presence of testlets is found in (Brown, 2006) which discusses a psychological study that involves evaluating the generalizability of the DSM-IV criteria in the diagnosis of Major Depression Disorder (MDD) between sexes. Each patient was rated on the severity of nine symptoms (items) related to MDD. These symptoms are described as follows:

- MDD1: depressed mood
- MDD2: loss of interest
- MDD3: weight/appetite change

- MDD4: sleep disturbance
- MDD5: psycho motor agitation/retardation
- MDD6: fatigue/loss of energy
- MDD7: feeling of worthlessness/guilt
- MDD8: concentration difficulties
- MDD9: thoughts of death/suicide

The study consisted of a sample size of 750 participants; 375 being male and 375 female. The study was modeled as a uni-dimensional model of MDD in accordance with its DSM-IV conceptualization. The first two items were also allowed to correlate with one another. So in a sense, they are considered as a testlet structure. Thus, a correlated error model is utilized for the analysis of this data.

The problem is that there is a possibility of salient sex differences in the expression of mood disorder. In other words, there is a belief that there is a difference as to how males and females express their mood when it comes to certain issues. For example, weight/appetite change may be more strongly related to depression in women than in men. Therefore running this prior to compensating for MI could lead to inaccurate results.

Table 5.1: DMS-IV Major Depressive Disorder in Men and Women

Level of Invariance	Chi-Square	df	pvalue	CFI	RMSEA	BIC
Configural	98.911	52.000	0.000	0.963	0.049	27665.358
Weak	102.839	60.000	0.000	0.966	0.044	27616.326
Strong	115.309	68.000	0.000	0.963	0.043	27694.997
Strict	125.021	77.000	0.000	0.962	0.041	27645.128

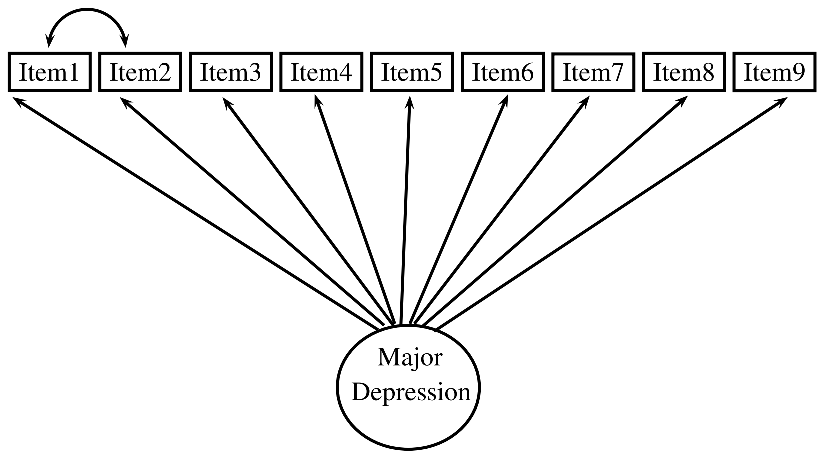

Figure 5.1: Uni-dimensional Model: MDD

Table 5.1 allows us to see how the model reacts at different levels of MI. It is clear that the overall fit statistics for the uni-dimensional solution show a good model fit throughout. The results indicate that the strict level of MI can be considered as the "best" model fit for the data. Therefore, the factor loadings, intercept, and uniquenesses are free to vary across both groups (men & women).

5.6 Conclusion

The main purpose behind this thesis was to assess one main issue. How can a researcher detect measurement invariance in the presence of testlets and which testlet model is most useful for detecting the true level of invariance? Three different models were proposed as a solution. The correlated error model, correlated traits correlated method minus one model (CT-C(M-1)), and the random intercept model. All three models appeared to show a great improvement in the fit of the model for the different cases that were proposed. However, even though they showed an improvement in model fit the CT-C(M-1) model appeared to have singularity issues when running the simulations. About 60 % of the time the simulation would not converge which would result in inconclusive results for the CT-C(M-1) model. Therefore, the CT-C(M-1) model could not be considered as an appropriate

solution for detecting MI in the presence of testlets.

Case 4 also appeared to have a lot of problems when trying to detect the appropriate level of MI for every proposed solution. Of all the cases it was the only one where the appropriate level of MI (weak) could not be detected. Taking all of this into consideration and by analyzing the results for the best model fit and model comparison, it is clear that the correlated error model is the most appropriate solution when trying to detect MI in the presence of testlets. For cases 1-3 the correlated error model detects the appropriate level of MI a majority of the time. Tables 5.1 - 5.6 clearly show through the p-value, CFI, and RMSEA fit indices that the correlated error model chose the appropriate level of MI a majority of the time. In some cases it choose the appropriate level 100 % of the time.

In conclusion, to help determine if in fact the correlated error is the best solution for detecting MI in the presence of testlets it would be necessary to test it with different sets of data which could represent a variety of different cases. Like in the simulation study, the data needs to be analyzed using the different proposed models to determine if in fact the correlated error model is the more appropriate solution. Another proposal for future research would be to run this simulation using different cases. By creating a variety of possible cases one could simulate possible solutions for different sets of data in applied research. This could benifit future research temendously by giving them guide as to how to go about treating certing models when testlets are present. e solution.

Références

Bain, L., & Engelhardt, M. (2000). *Introduction to probability and mathematical statistics.* Duxbury Press.

Brannick, M. (1995). Applying covariance structure modeling. *Journal of Organizational Behavior*, *16*, 201-213.

Brown, T. (2006). *Confirmatory factor analysis for applied research.* Guilford Publications, Inc.

Chen, F., Curran, P., Bollen, K., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in rmsea test statistic in structural equation models. *Sociological Methods and Research*, *36*, 462-494.

DeMars, C. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*, 145-168.

Dimitrov, D. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*, 121-149.

Ellis, M., Aguirre-Urreta, M., Sun, W., & Marakas, G. (2008). *Establishing the need for measurement invariance in information systems research: A step by step example using technology acceptance research.*

Johnson, R., & Wichern, D. (2007). *Applied multivariate statistical analysis.* Pearson Education.

Kenny, D. (2003). *Structural equation modeling.* Disponible sur `http://davidakenny.net/cm/causalm.htm`

McDonald, R. (1999). *Test theory.* Lawrence Erlbaum Associates.

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

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance.

Psychometrika, *58*, 525-543.

Olivares, A., & Coffman, D. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344-362.

Rosseel, Y. (2011). *Lavaan: An r package for structural equation modeling and mode version 0.4-7.*

Wu, A., Li, Z., & Zumbo, B. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with timss data. *Practical Assessment and Evaluation*, *12*, 1-26.

Ximenez, C. (2006). A monte carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Structural Equation Modeling*, *13*, 587-614.

Appendix A

Tables

A.1 Model Fit-Case-1-n=500, Λ=0.6 and 0.8

Table A.1: Measurement Invariance Results

Naive Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.058	0.00	0.973	21375.94	193.637	70
Weak	.059	0.00	0.968	21341.69	221.55	79
Strong	.0547	0.00	0.9708	21418.17	222.04	88
Strict	.0524	0.00	0.9701	21362.26	235.219	98

Table A.2: Measurement Invariance Results

Naive Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
configural	0.0617	0.00	0.981	21899.46	205.31	70
weak	0.0596	0.00	0.980	21853.42	221.44	79
strong	0.0547	0.00	0.981	21929.91	221.94	88
strict	0.0526	0.00	0.9807	21875.23	236.34	98

Table A.3: Measurement Invariance Results

Correlated Error Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0098	0.487	0.9987	21375.11	51.01	50
Weak	0.0045	0.675	0.999	21313.47	51.5	59
Strong	0.00167	0.819	0.999	21389.81	51.89	68
Strict	0.0241	0.101	0.994	21372.72	103.88	78

Table A.4: Measurement Invariance Results

Correlated Error Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
configural	0.0094	0.5034	0.9992	21889.11	50.36	50
weak	0.0043	0.6874	0.9996	21827.38	50.81	59
strong	0.0016	0.8259	0.9998	21903.76	51.2	68
strict	0.0225	0.1258	0.9967	21884.54	101.06	78

Table A.5: Measurement Invariance Results

CT-C(M-1) Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0078	0.547	0.9988	21303.15	58.37	60
Weak	.0045	0.676	0.99931	21219.47	64.5	73
Strong	.00215	0.7966	0.999682	21302.37	64.51	81
Strict	.00075	0.8903	0.999889	21234.14	65.35	91

Table A.6: Measurement Invariance Results

CT-C(M-1) Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
configural	0.0082	0.5316	0.9992	21835.65	59.23	60
weak	0.0045	0.6661	0.9995	21751.80	65.19	73
strong	0.0021	0.7905	0.9997	21834.71	65.2	81
strict	0.0008	0.887	0.9999	21766.48	66.05	91

Table A.7: Measurement Invariance Results

Random Intercepts Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0087	0.5098	0.998	21257.37	68.11	68
Weak	0.0044	0.667	0.999265	21195.76	68.67	77
Strong	0.0022	0.7883	0.999632	21278.66	68.67	85
Strict	0.0009	0.8851	0.999851	21210.11	69.203	95

Table A.8: Measurement Invariance Results

Random Intercepts Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
configural	0.0095	0.4775	0.9989	21794.92	69.62	68
weak	0.0050	0.6411	0.9994	21733.26	70.12	77
strong	0.0024	0.7681	0.9997	21816.15	70.15	85
strict	0.0009	0.8739	0.9998	21747.62	70.67	95

A.2 Model Fit-Case-2-n=500, Λ=0.6

Table A.9: Measurement Invariance Results

Naive Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0834	0.00	0.9477	21891.34	315.9	70
Weak	0.077	0.00	0.9493	21830.88	317.7	79
Strong	0.071972	0.00	0.951	21907.59	318.4242	88
Strict	0.0669	0.00	0.9528	21840.15	320.06	98

Table A.10: Measurement Invariance Results

Naive Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
configural	0.0874	0.00	0.9625	22417.31	339.57	70
weak	0.0810	0.00	0.9636	22356.53	340.97	79
strong	0.0755	0.00	0.9647	22433.412	341.86	88
strict	0.0704	0.00	0.9658	22366.15	343.68	98

Table A.11: Measurement Invariance Results

Correlated Error Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0098	0.49	0.9987	21770.52	50.9845	50
Weak	0.0045	0.67	0.9994	21709.06	51.69	59
Strong	0.0018	0.82	0.9997	21785.57	52.22	68
Strict	0.00075	0.90	0.9999	21718.98	54.699	78

Table A.12: Measurement Invariance Results

Correlated Error Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0095	0.5023	0.9992	22273.98	50.42	50
Weak	0.0044	0.6840	0.9996	22212.46	51.06	59
Strong	0.0016	0.8209	0.9998	22289.03	51.65	68
Strict	0.0006	0.8992	0.99994	22222.53	54.23	78

Table A.13: Measurement Invariance Results

CT-C(M-1) Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0086	0.52	0.99872	21697.9	59.7495	60
Weak	0.0037	0.71	0.99947	21610.68	35.3285	73
Strong	0.001578	0.83	0.99978	21693.59	62.34	81
Strict	0.000494	0.91	0.999926	21625.04	62.87935	91

Table A.14: Measurement Invariance Results

CT-C(M-1) Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0091	0.4987	0.9991	22221.76	60.67	60
Weak	0.0037	0.704671854765123	0.9996	22134.38	63.08	73
Strong	0.0016	0.821000405175815	0.9998	22217.29	63.10	81
Strict	0.0005	0.910041890528976	0.9999	22148.76	63.65	91

Table A.15: Measurement Invariance Results

Random Intercepts Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.008584	0.51	0.99866	21653.52	67.96	68
Weak	0.004361	0.67	0.99931	21592.06	68.67	77
Strong	0.0021	0.79	0.9996	21674.95	68.67	85
Strict	0.0008	0.89	0.9998	21606.38	69.17	95

Table A.16: Measurement Invariance Results

Random Intercept Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0095	0.4760	0.999	22178.77	69.62	68
Weak	0.005	0.638511112950461	0.9994	22117.29	70.301	77
Strong	0.0024	0.76779931640004	0.999747882632472	22200.18	70.307	85
Strict	0.0009	0.8747	0.9998	22131.65	70.85	95

A.3 Model Fit-Case-3 n=500, Λ=0.6

Table A.17: Measurement Invariance Results

Naive Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.008292	0.51534	0.99847	20817.86	70.02	70
Weak	.02867	0.06203	0.99137	20800.78	115.11	79
Strong	0.0228	0.1301	0.9934	20876.82	115.167	88
Strict	0.017032	0.2339	0.9952	20808.78	116.204	98

Table A.18: Measurement Invariance Results

Naive Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0087	0.5017	0.9989	21287.62	70.52	70
Weak	0.0317	0.03693	0.9931	21277.02	122.0	79
Strong	0.0261	0.08239	0.9945	21353.06	122.15	88
Strict	0.0204	0.1621	0.9958	21284.69	122.86	98

Table A.19: Measurement Invariance Results

Correlated Error Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.00997	0.4877	0.9985	20941.6	51.037	50
Weak	0.03064	0.06348	0.9926	20918.27	89.87	59
Strong	0.023102	0.148	0.99463	20994.31	89.93	68
Strict	0.01769	0.242272	0.999587	20928.98	93.67	78

Table A.20: Measurement Invariance Results

Correlated Error Model (Λ=.8,n=500)

Configural	0.0094	0.5053	0.9991	21409.97	50.33	50
Weak	0.0349	0.0285	0.9939	21394.98	97.50	59
Strong	0.0276	0.0816	0.9952	21471.02	97.56	68
Strict	0.0213	0.1649	0.9964	21404.38	100	78

Table A.21: Measurement Invariance Results

CT-C(M-1) Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.007344	0.5685	0.9988	20871.13	57.413	60
Weak	0.026	0.101346	0.9931	20825.41	101.49	73
Strong	0.02011	0.18813	0.9948	20908.33	101.52	81
Strict	0.014	0.3162	0.9964	20840.6	102.87	91

Table A.22: Measurement Invariance Results

CT-C(M-1) Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0076	0.5586	0.9991	21353.04	58.01	60
Weak	0.02949	0.0610	0.9943	21313.53	108.30	73
Strong	0.0239	0.1214	0.9955	21396.48	108.36	81
Strict	0.0178	0.2242	0.9968	21328.39	109.34	91

Table A.23: Measurement Invariance Results

Random Intercepts Model(Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.00879	0.5096	0.9984	20824.9	68.15	68
Weak	0.0264	0.0883	0.99264	20802.23	107.6	77
Strong	0.0208	0.1653	0.9944	20885.14	107.668	85
Strict	0.0151	0.283	0.9961	20817.37	108.97	95

Table A.24: Measurement Invariance Results

Random Intercept Model (Λ=.8,n=500)

Configural	0.0097	0.4731	0.9988	21314.64	69.80	68
Weak	0.0310	0.04234	0.9935	21300.01	117.35	77
Strong	0.0258	0.0897	0.9948	21382.92	117.36	85
Strict	0.01978	0.17785	0.9961	21314.71	118.23	95

A.4 Model Fit-Case-4-n=500, Λ=0.6

Table A.25: Measurement Invariance Results

Naive Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0834	0.00	0.948	21890.31	315.87	70
Weak	0.0783	0.00	0.9481	21835.93	323.665	79
Strong	0.078443	0.00	0.94218	21949.01	360.7551	88
Strict	0.0732	0.00	0.943842	21881.93	362.75	98

Table A.26: Measurement Invariance Results

Naive Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0874	0.00	0.9625	22418.02	339.73	70
Weak	0.0822	0.00	0.9626	22364.24	348.12	79
Strong	0.0787	0.00	0.961	22455.11	363.01	88
Strict	0.0735	0.00	0.9629	22388.16	365.13	98

Table A.27: Measurement Invariance Results

Correlated Error Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	.0098	0.4864	0.998	21769.05	50.99	50
Weak	0.0085	0.514	0.998	21714.63	58.74	59
Strong	0.0053	0.619	0.999	21794.65	62.77	68
Strict	0.00266	0.748	0.999	21728.36	65.56	78

Table A.28: Measurement Invariance Results

Correlated Error Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0092	0.5077	0.9992	22273.77	50.24	50
Weak	0.0082	0.5228	0.9992	22219.84	58.48	59
Strong	0.0047	0.6546	0.9995	22298.47	61.12	68
Strict	0.0021	0.7778	0.9997	22232.08	63.81	78

Table A.29: Measurement Invariance Results

CT-C(M-1) Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0088	0.51374	0.998716	21701.99	59.96	60
Weak	0.0065	0.581	0.9989	21621.64	69.42	73
Strong	0.0043	0.6656	0.9993	21707.83	72.72	81
Strict	0.0017	0.8034	0.99973	21639.51	73.47	91

Table A.30: Measurement Invariance Results

CT-C(M-1) Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.00926	0.4966	0.9991	22219.55	60.82	60
Weak	0.0070	0.5511	0.9992	22139.86	70.93	73
Strong	0.0042	0.6654	0.9995	22224.69	72.87	81
Strict	0.0017	0.8047	0.9998	22156.33	73.58	91

Table A.31: Measurement Invariance Results

Random Intercepts Model (Λ=.6,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0086	0.51237	0.9986	21653.22	67.96	68
Weak	0.0074	0.542	0.998	21598.47	75.385	77
Strong	0.0051	0.629	0.99912	21684.51	78.53	85
Strict	0.00232	0.769	0.9996	21616.19	79.28	95

Table A.32: Measurement Invariance Results

Random Intercept Model (Λ=.8,n=500)

	RMSEA	P-value	CFI	BIC	χ^2	df
Configural	0.0093	0.4768	0.99902	22176.72	69.53	68
Weak	0.0085	0.4932	0.99904	22122.72	77.7	77
Strong	0.0054	0.6087	0.9994	22207.50	79.59	85
Strict	0.0023	0.7579	0.9997	22139.13	80.29	95

Appendix B

Model Comparisons

B.1 Model Difference Tests-Case-1-n=500, λ=0.6

Table B.1: Model Difference Tests

Naive Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak (1v2)	27.91	9	0.0145	0.00411	0.0064
Strong (3v2)	0.4911	9	0.9977	-0.00185	0.0048
Strict(4v3)	13.17	10	0.2505	0.00068	0.00229

Table B.2: Model Difference Tests

Correlated Error Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	0.5349	9	0.9989	-0.00069	-0.01428
Strong(3v2)	0.3525	9	0.9992	-0.00035	0.00287
Strict(4v3)	51.98	10	1.651e-05	0.00545	-0.0224

Table B.3: Model Difference Tests

CT-C(M-1) Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	6.12	13	0.8870	-0.0004	0.00698
Strong(3v2)	0.0096	8	0.9999	-0.00037	0.0023
Strict(4v3)	0.8434	10	0.9995	-0.0002	0.0014

Table B.4: Model Difference Tests

Random Intercept Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	0.5601	9	0.9987	-0.0006	0.00781
Strong(3v2)	0.00	8	1	-0.0003	0.0021
Strict(4v3)	0.5304	10	0.9999	-0.0002	0.0013

B.2 Model Difference Tests-Case-2-n=500, λ=0.6

Table B.5: Model Difference Tests

Naive Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	1.7095	9	0.9797	-0.0015	0.01651
Strong(3v2)	0.7229	9	0.9943	-0.0017	0.00537
Strict(4v3)	1.632	10	0.9954	-0.0017	0.0051

Table B.6: Model Difference Tests

Correlated Error Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	0.7123	9	0.9974	-0.0006	0.00909
Strong(3v2)	0.5219	9	0.997	-0.0003	0.00268
Strict(4v3)	2.481	10	0.9726	-0.00013	0.0012

Table B.7: Model Difference Tests

CT-C(M-1) Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	0.7123	9	0.997	-0.00067	0.0091
Strong(3v2)	0.5219	9	0.997	-0.00033	0.0026
Strict(4v3)	2.48	10	0.9726	-0.00013	0.0012

Table B.8: Model Difference Tests

Random Intercept Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	0.7049	9	0.9974	-0.00064	0.00772
Strong(3v2)	0.00	8	1	-0.0003	0.0022
Strict(4v3)	0.5053	10	0.9999	-0.0002	0.00124

B.3 Model Difference Tests-Case-3-n=500, λ=0.6

Table B.9: Model Difference Tests

Naive Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	45.08	9	0.0009	0.007	-0.0087
Strong(3v2)	0.0555	9	0.9999	-0.002	0.0058
Strict(4v3)	1.0368	10	0.9995	-0.0018	0.0057

Table B.10: Model Difference Tests

Correlated Error Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	38.83	9	0.0025	0.0059	-0.0077
Strong(3v2)	0.05362	9	0.999	-0.002	0.0075
Strict(4v3)	3.745	10	0.9412	-0.0012	0.0054

Table B.11: Model Difference Tests

CT-C(M-1) Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	44.07	13	0.0054	0.0056	-0.00674
Strong(3v2)	0.0282	8	0.9993	-0.00172	0.0058
Strict(4v3)	1.347	10	0.9948	-0.0016	0.006

Table B.12: Model Difference Tests

Random Intercept Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	39.51	9	0.0013	0.0057	-0.0063
Strong(3v2)	0.01	8	0.999	-0.0017	0.0055
Strict(4v3)	1.309	10	0.9988	-0.0016	0.0057

B.4 Model Difference Tests-Case-4-n=500, λ=0.6

Table B.13: Model Difference Tests

Naive Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	7.795	9	0.589	-0.0002	0.0102
Strong(3v2)	37.09	9	0.00339	0.0059	0.00
Strict(4v3)	1.999	10	0.9912	-0.0017	0.00522

Table B.14: Model Difference Tests

Correlated Error Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	7.753	9	0.5867	0.00	0.0072
Strong(3v2)	4.02	9	0.8758	-0.0004	0.0032
Strict(4v3)	2.792	10	0.9659	-0.00038	0.0026

Table B.15: Model Difference Tests

CT-C(M-1) Model (λ=0.6)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	9.45	13	0.714	-0.0002	0.0071
Strong(3v2)	3.29	8	0.883	-0.0003	0.00218
Strict(4v3)	0.751	10	0.9997	-0.00039	0.0026

Table B.16: Model Difference Tests

Random Intercept Model (λ=0.6, N=500)

	$\Delta\chi^2$	Δ df	Δ P.value	Δ CFI	Δ RMSEA
Weak(1v2)	7.42	9	0.611	0.00	0.0062
Strong(3v2)	3.14	8	0.894	-0.00038	0.0023
Strict(4v3)	0.752	10	0.9997	-0.00046	0.0027

Appendix C

Chi-square difference test power calculations

C.1 CT-C(M-1) model λ=0.6

Table C.1: Case 1-n=100

	CFI	p-value	RMSEA
Strict	0.97	1.00	1.00
Strong	0.02	0.00	0.00

Table C.2: Case 1-n=500

	CFI	p-value	RMSEA
Strict	1.00	1.00	1.00
Strong	0.00	0.00	0.00

C.2 Random Intercept Model λ=0.6

Table C.3: Case 2-n=100

	CFI	p-value	RMSEA
Strict	0.98	1.00	1.00
Strong	0.02	0.00	0.00

Table C.4: Case 2-n=500

	CFI	p-value	RMSEA
Strict	1.00	1.00	1.00
Strong	0.00	0.00	0.00

Table C.5: Case 3-n=100

	CFI	p-value	RMSEA
Weak	0.91	0.99	1.00
Configural	0.09	0.01	0.00

Table C.6: Case 3-n=500

	CFI	p-value	RMSEA
Weak	0.00	0.181	0.00
Configural	1.00	0.819	1.00

Table C.7: Case 4-n=100

	CFI	p-value	RMSEA
Weak	1.00	1.00	1.00
Configural	0.00	0.00	0.00

Table C.8: Case 4-n=500

	CFI	p-value	RMSEA
Weak	1.00	0.999	1.00
Configural	0.00	0.001	0.00

	CFI	p-value	RMSEA
Strict	0.97	1.00	1.00
Strong	0.03	0.00	0.00

Table C.11: Case 1-n=500

	CFI	p-value	RMSEA
Strict	0.00	0.00	0.00
Strong	1.00	1.00	1.00

Table C.12: Case 2-n=100

	CFI	p-value	RMSEA
Strict	0.97	1.00	1.00
Strong	0.02	0.00	0.00

Table C.13: Case 2-n=500

	CFI	p-value	RMSEA
Strict	1.00	1.00	1.00
Strong	0.00	0.00	0.00

Table C.14: Case 3-n=100

	CFI	p-value	RMSEA
Weak	0.86	0.97	1.00
Configural	0.14	0.03	0.00

Table C.15: Case 3-n=500

	CFI	p-value	RMSEA
Weak	1.00	0.115	1.00
configural	0.00	0.885	0.00

Table C.16: Case 4-n=100

	CFI	p-value	RMSEA
0	1.00	0.999	1.00
1	0.00	0.001	0.00

Table C.17: Case 4-n=500

	CFI	p-value	RMSEA
0	1.00	0.998	1.00
1	0.00	0.002	0.00

Appendix D

R:Code

```
library(lavaan)
set.seed(2381)
rep=1000
chi.config = matrix(NA,rep,1)
chi.weak = matrix(NA,rep,1)
chi.strong = matrix(NA,rep,1)
chi.strict = matrix(NA,rep,1)
df.1 = matrix(NA,rep,1)
df.2 = matrix(NA,rep,1)
df.3 = matrix(NA,rep,1)
df.4 = matrix(NA,rep,1)
weak.dif = matrix(NA,rep,1)
strong.dif = matrix(NA,rep,1)
strict.dif = matrix(NA,rep,1)
strong2 = matrix(NA,rep,1)
strict2 = matrix(NA,rep,1)
strong2.df=matrix(NA,rep,1)
strict2.df=matrix(NA,rep,1)
df.w = matrix(NA,rep,1)
df.sg = matrix(NA,rep,1)
df.st = matrix(NA,rep,1)
cfi.config = matrix(NA,rep,1)
cfi.weak = matrix(NA,rep,1)
cfi.strong = matrix(NA,rep,1)
cfi.strict = matrix(NA,rep,1)
rmsea.config = matrix(NA,rep,1)
rmsea.weak = matrix(NA,rep,1)
rmsea.strong = matrix(NA,rep,1)
rmsea.strict = matrix(NA,rep,1)
bic.config = matrix(NA,rep,1)
bic.weak = matrix(NA,rep,1)
bic.strong = matrix(NA,rep,1)
```

```
bic.strict = matrix(NA,rep,1)
pvalue.config=matrix(NA,rep,1)
pvalue.weak=matrix(NA,rep,1)
pvalue.strong=matrix(NA,rep,1)
pvalue.strict=matrix(NA,rep,1)
p.weak=matrix(NA,rep,1)
p.strong=matrix(NA,rep,1)
p.strong2=matrix(NA,rep,1)
p.strict=matrix(NA,rep,1)
p.strict2=matrix(NA,rep,1)
final_1=matrix(NA,rep,1)
final_2=matrix(NA,rep,1)
final_3=matrix(NA,rep,1)
final_4=matrix(NA,rep,1)
final_5=matrix(NA,rep,1)
se.c =matrix(NA,rep,1)
se.w=matrix(NA,rep,1)
se.sg=matrix(NA,rep,1)
se.st=matrix(NA,rep,1)

# Measurement Model
Model1 = '
f1 =~ i1 + i2 + i3 + i4 + i5 + i6 + i7 + i8 + i9 + i10

'
Model2 = '
f1 =~ i1 + i2 + i3 + i4 + i5 + i6 + i7 + i8 + i9 + i10

# correlated residuals
i1 ~~ i2
i1 ~~ i3
i1 ~~ i4
i1 ~~ i5
i2 ~~ i3
i2 ~~ i4
i2 ~~ i5
i3 ~~ i4
i3 ~~ i5
i4 ~~ i5

'
```

```
#CT-C(M-1) model
Model3 = '
f1 =~ i1 + i2 + i3 + i4 + i5 + i6 + i7 + i8 + i9 + i10
f2 =~ i1 + i2 + i3 + i4 + i5
#covariances
f1 ~~ 0*f2
'

# Random Intercepts model
Model4 = '
f1 =~ i1 + i2 + i3 + i4 + i5 + i6 + i7 + i8 + i9 + i10
r =~ 1*i1 + 1*i2 + 1*i3 + 1*i4 + 1*i5
#covariances
f1 ~~ 0*r
'

k=10
tau=matrix(rep(0.5,k))
lambda2=matrix(c(rep(lambda.o,2),.3,0.3,rep(lambda.o,k-4)),10,1)

for(n in c(100,250,500))
{
for(lambda.o in c(.6,.8))
{
lambda=matrix(rep(lambda.o,k))

y=matrix(NA,n,10)
x=matrix(NA,n,10)
sig.f=1

block=matrix(c(.4,.4,.4,.4,.4,0,0,0,0,0))
for(mod in 1:4)
{
if (mod==1) {Model=Model1}
if (mod==2) {Model=Model2}
if (mod==3) {Model=Model3}
if (mod==4) {Model=Model4}

for(j in 1:rep)
{

####################Case 1 #####################################
```

```
for (i in 1:n) {
psi=abs(rnorm(10,0,1))
f=matrix(rnorm(1,0,sig.f))
f2=matrix(rnorm(1,0,sig.f))
psi2=block%*%f2+psi
x[i,]=(lambda%*%f+psi2)
y[i,]=(lambda%*%f2+psi)
}
#######################################################################

####################Case 2 ########################################

#for (i in 1:n) {
#psi=abs(rnorm(10,0,1))
#f=matrix(rnorm(1,0,sig.f))
#f2=matrix(rnorm(1,0,sig.f))
#psi2=block%*%f2+psi
#x[i,]=(lambda%*%f+psi2)
#y[i,]=(lambda%*%f2+psi2)
#}
########################################################################

####################Case 3 ########################################

lambda2=matrix(c(rep(lambda.o,2),.3,0.3,rep(lambda.o,k-4)),10,1)
#for (i in 1:n) {
#psi=abs(rnorm(10,0,1))
#psi2=block%*%f2+psi
#f=matrix(rnorm(1,0,sig.f))
#f2=matrix(rnorm(1,0,sig.f))
#x[i,]=(lambda2%*%f+psi2)
#y[i,]=(lambda2%*%f2+psi1)
#}
######################################################################

####################Case 4 ########################################

#for (i in 1:n) {
#psi=abs(rnorm(10,0,1))
#psi1=block%*%f+psi
#f=matrix(rnorm(1,0,sig.f))
```

```
#f2=matrix(rnorm(1,0,sig.f))
#y[i,]=(tau+lambda%*%f+psi1)
#x[i,]=(lambda%*%f2+psi1)
#}

################################################################


colnames(y)=c("i1","i2","i3","i4","i5","i6","i7","i8","i9","i10")
colnames(x)=c("i1","i2","i3","i4","i5","i6","i7","i8","i9","i10")


# Naming the columns and binding the group names to the data
place=c(rep("a",n),rep("b",n))
groups=rbind(x,y)
#groups=matrix(c(x,y),(2*n),10,byrow=FALSE)
colnames(groups)=c("i1","i2","i3","i4","i5","i6","i7","i8","i9","i10")

#groups
place2 = data.frame(place)
data_1=cbind(place2,groups)



#create multiple fit with group contraints



#Configural
fit.config=cfa(Model,data=data_1,group="place",group.equal=NULL)
#summary(fit.config, standardized=TRUE, fit.measures=TRUE)
#get to the slot 'se' and if 0 then record this as a warning.
se.slot.c=mean((fit.config@Fit)@se)
se.c[j]=ifelse(se.slot.c==0,1,0)



#repeat this code for all models below
#weak
fit.weak=cfa(Model,data=data_1,group="place",group.equal="loadings")
#summary(fit.weak, standardized=TRUE, fit.measures=TRUE)
se.slot.w=mean((fit.config@Fit)@se)
```

```
se.w[j]=ifelse(se.slot.w==0,1,0)


#strong
fit.strong=cfa(Model,data=data_1,group="place",group.equal=
c("intercepts","loadings"))
#summary(fit.strong, standardized=TRUE, fit.measures=TRUE)
se.slot.sg=mean((fit.config@Fit)@se)
se.sg[j]=ifelse(se.slot.sg==0,1,0)


#strict
fit.strict=cfa(Model,data=data_1,group="place",group.equal=
c("residuals","intercepts","loadings"))
#summary(fit.strict, standardized=TRUE, fit.measures=TRUE)
se.slot.st=mean((fit.config@Fit)@se)
se.st[j]=ifelse(se.slot.st==0,1,0)


#Extracting fit indices

#inspect(fit.config)
fit.indices1=inspect(fit.config, what = "fit")
chi.config[j]=fit.indices1["chisq"]
df.1[j]=fit.indices1["df"]
cfi.config[j]=fit.indices1["cfi"]
rmsea.config[j]=fit.indices1["rmsea"]
bic.config[j]=fit.indices1["bic"]
pvalue.config[j]=fit.indices1["pvalue"]

#inspect(fit.weak)
fit.indices2=inspect(fit.weak, what = "fit")
chi.weak[j]=fit.indices2["chisq"]
df.2[j]=fit.indices2["df"]
cfi.weak[j]=fit.indices2["cfi"]
rmsea.weak[j]=fit.indices2["rmsea"]
bic.weak[j]=fit.indices2["bic"]
pvalue.weak[j]=fit.indices2["pvalue"]

#inspect(fit.strong)
fit.indices3=inspect(fit.strong, what="fit")
chi.strong[j]=fit.indices3["chisq"]
```

```
df.3[j]=fit.indices3["df"]
cfi.strong[j]=fit.indices3["cfi"]
rmsea.strong[j]=fit.indices3["rmsea"]
bic.strong[j]=fit.indices3["bic"]
pvalue.strong[j]=fit.indices3["pvalue"]

#inspect(fit.strict)
fit.indices4=inspect(fit.strict, what="fit")
chi.strict[j]=fit.indices4["chisq"]
df.4[j]=fit.indices4["df"]
cfi.strict[j]=fit.indices4["cfi"]
rmsea.strict[j]=fit.indices4["rmsea"]
bic.strict[j]=fit.indices4["bic"]
pvalue.strict[j]=fit.indices4["pvalue"]



#chi-square difference

#Weak:
#Model 1 1 vs Model 2
weak.dif[j]=chi.weak[j]-chi.config[j]
df.w[j]=df.2[j]-df.1[j]
p.weak[j]=1-(pchisq(weak.dif[j],df=df.w[j]))

#Strong:
#Model 1 vs Model 3
strong.dif[j]=chi.strong[j]-chi.config[j]
df.sg[j]=df.3[j]-df.1[j]
p.strong[j]=1-(pchisq(strong.dif[j],df=df.sg[j]))

#Model 3 vs Model 2
strong2[j]=chi.strong[j]-chi.weak[j]
strong2.df[j]=df.3[j]-df.2[j]
p.strong2[j]=1-(pchisq(strong2[j],df=strong2.df[j]))

#Strict:
#Model 4 vs Model 1
strict.dif[j]=chi.strict[j]-chi.config[j]
df.st[j]=df.4[j]-df.1[j]
p.strict[j]=1-(pchisq(strict.dif[j],df=df.st[j]))
```

```
#Model 4 vs Model 3
strict2[j]=chi.strict[j]-chi.strong[j]
strict2.df[j]=df.4[j]-df.3[j]
p.strict2[j]=1-(pchisq(strict2[j],df=strict2.df[j]))

# Evaluation of results
final_1[j]=ifelse(p.weak[j]>.05,"weak","config")
#final_2[j]=ifelse(p.strong[j] > .05,"strong","config")
final_3[j]=ifelse(p.strong2[j] > .05,"strong","weak")
#final_4[j]=ifelse(p.strict[j] > .05,"strict","config")
final_5[j]=ifelse(p.strict2[j] > .05,"strict","strong")
}

#Results count
#CvW=ifelse(final_1=="strong",1,0)
#CvSG=ifelse(final_2=="strong",1,0)
#WvSG=ifelse(final_3=="strong",1,0)
#WvST=ifelse(final_4=="strong",1,0)
#SGvST=ifelse(final_5=="strong",1,0)


if (final_1[j]=="config") {conclusion="final_1[j]"}
else if (final_3[j]=="weak") {conclusion="final_3[j]"}
else if (final_5[j]=="strong") {conclusion="final_5[j]"}

#sum(CvW)
#sum(CvSG)
#sum(WvSG)
#sum(WvST)
#sum(SGvST)


# Results table
#Results = data.frame(CvW,CvSG,WvSG,WvST,SGvST)

chisq.stat = data.frame(chi.config,df.1,pvalue.config,
cfi.config,rmsea.config,bic.config,chi.weak,df.2,pvalue.weak,
cfi.weak,rmsea.weak,bic.weak,chi.strong,df.3,
pvalue.strong,cfi.strong,rmsea.strong,bic.strong,chi.strict,df.4,
pvalue.strict,cfi.strict,rmsea.strict,bic.strict)
chisq.diff=data.frame(weak.dif,df.w,p.weak,strong.dif,
df.sg,p.strong,strong2,strong2.df,p.strong2,strict.dif,
```

```
df.st,p.strict,strict2,strict2.df,p.strict2)
final_results=data.frame(final_1,final_3,final_5)
std_errors = data.frame(se.c,se.w,se.sg,se.st)
#cfi.fit=data.frame(cfi.config,cfi.weak,cfi.strong,cfi.strict)
#rmsea.fit=data.frame(rmsea.config,rmsea.weak,rmsea.strong,rmsea.strict)
#bic.fit=data.frame(bic.config,bic.weak,bic.strong,bic.strict)

#mi.results1=measurementinvariance(Model2,data=data_1,group="place",strict=TRUE)


setwd("File Name")
#setwd("F:\\Thesis\\results")
write.table(chisq.stat,paste("chisq.results-mod=",
mod,"n=",n,"k=",k,"lam=",lambda.o,".dat"),
sep=" ",quote=FALSE,row.name=F,col.names=T)
write.table(chisq.diff,paste("chi_diff.results-mod=",
mod,"n=",n,"k=",k,"lam=",lambda.o,".dat"),
sep=" ",quote=FALSE,row.name=F,col.names=T)
#write.table(cfi.fit,
paste("cfi.fit-mod=",mod,"n=",n,"k=",k,"lam=",lambda.o,".dat"),
sep=" ",quote=FALSE,row.name=F,col.names=T)
#write.table(rmsea.fit,
paste("rmsea-mod=",mod,"n=",n,"k=",k,"lam=",lambda.o,".dat"),
sep=" ",quote=FALSE,row.name=F,col.names=T)
#write.table(bic.fit,
paste("bic-mod=",mod,"n=",n,"k=",k,"lam=",lambda.o,".dat"),
sep=" ",quote=FALSE,row.name=F,col.names=T)

#write.table(final_results,
paste("conclusion-mod=",mod,"n=",n,"k=",k,"lam=",lambda.o,".dat"),
 sep=" ",quote=FALSE,row.name=F,col.names=T)
#write.table(std_errors,
paste("SE-mod=",mod,"n=",n,"k=",k,"lam=",lambda.o,".dat"),
sep=" ",quote=FALSE,row.name=F,col.names=T)

}#closes model loop
}#closes lambda loop
}

#chi_diff_06=read.table("chi_diff_06.dat",header=TRUE,sep="")
#chi_diff_08=read.table("chi_diff_08.dat",header=TRUE,sep="")
#chi_stat_06=read.table("chisq.results_06.dat",header=TRUE,sep="")
```

```
#chi_stat_08=read.table("chisq.results_08.dat",header=TRUE,sep="")

############################## Extract Fit Statistics ###############################

x=read.table("chisq.results-mod= 2 n= 500 k= 10 lam= 0.6 .dat",header=TRUE,sep="")
y=read.table("chi_diff.results-mod= 2 n= 500 k= 10 lam= 0.6 .dat",header=TRUE,sep="")
z=read.table("SE-mod= 2 n= 500 k= 10 lam= 0.6 .dat",header=TRUE,sep="")
k=read.table("conclusion-mod= 2 n= 500 k= 10 lam= 0.6 .dat",header=TRUE,sep="")

head(x)
head(y)
head(z)
head(k)

# Model Choice
mod1=ifelse(k$final_1=="config",1,0)
mod2=ifelse(k$final_3=="strong",1,0)
mod3=ifelse(k$final_5=="strict",1,0)
sum(mod1)/1000
sum(mod2)/1000
sum(mod3)/1000

#Model Statistics
mean(x$chi.config)
mean(x$chi.weak)
mean(x$chi.strong)
mean(x$chi.strict)

mean(x$rmsea.config)
mean(x$rmsea.weak)
mean(x$rmsea.strong)
mean(x$rmsea.strict)

mean(x$bic.config)
mean(x$bic.weak)
mean(x$bic.strong)
mean(x$bic.strict)

mean(x$cfi.config)
mean(x$cfi.weak)
mean(x$cfi.strong)
mean(x$cfi.strict)
```

```r
mean(x$pvalue.config)
mean(x$pvalue.weak)
mean(x$pvalue.strong)
mean(x$pvalue.strict)

mean(x$df.1)
mean(x$df.2)
mean(x$df.3)
mean(x$df.4)


# Chi-square difference test
mean(y$weak.dif)
mean(y$strong.dif)
mean(y$strong2)
mean(y$strict.dif)
mean(y$strict2)

mean(y$p.weak)
mean(y$p.strong)
mean(y$p.strong2)
mean(y$p.strict)
mean(y$p.strict2)

mean(y$df.w)
mean(y$df.sg)
mean(y$strong2.df)
mean(y$df.st)
mean(y$strict2.df)

# Comparative fit index
a=mean(x$cfi.config)
b=mean(x$cfi.weak)
c=mean(x$cfi.strong)
d=mean(x$cfi.strict)
c1=a-b
c2=a-c
c3=b-c
c4=a-d
c5=c-d
```

```
#  RMSEA
e=mean(x$rmsea.config)
f=mean(x$rmsea.weak)
g=mean(x$rmsea.strong)
h=mean(x$rmsea.strict)
r1=e-h
r2=e-g
r3=f-g
r4=e-h
r5=g-h

#Chi-statistics
chi=matrix(c(mean(x$chi.config),mean(x$chi.weak),
mean(x$chi.strong),mean(x$chi.strict)),4,1)
P.VALUE=matrix(c(mean(x$pvalue.config),
mean(x$pvalue.weak),mean(x$pvalue.strong),mean(x$pvalue.strict)),4,1)
RMSEA=matrix(c(mean(x$rmsea.config),
mean(x$rmsea.weak),mean(x$rmsea.strong),mean(x$rmsea.strict)),4,1)
CFI=matrix(c(mean(x$cfi.config),
mean(x$cfi.weak),mean(x$cfi.strong),mean(x$cfi.strict)),4,1)
BIC=matrix(c(mean(x$bic.config),
mean(x$bic.weak),mean(x$bic.strong),mean(x$bic.strict)),4,1)
DF=matrix(c(mean(x$df.1),
mean(x$df.2),mean(x$df.3),mean(x$df.4)),4,1)
table1=cbind(chi,P.VALUE,RMSEA,CFI,BIC,DF)
colnames(table1)=c("chi","P.VALUE","RMSEA","CFI","BIC","DF")
row=matrix(c("configural","weak","strong","strict"),4,1)
table=cbind(row,table1)

# Model Comparison

chi_dif=matrix(c(mean(y$weak.dif),mean(y$strong.dif),
mean(y$strong2),mean(y$strict.dif),mean(y$strict2)),5,1)
P_value=matrix(c(mean(y$p.weak),mean(y$p.strong),
mean(y$p.strong2),mean(y$p.strict),mean(y$p.strict2)),5,1)
DF_diff=matrix(c(mean(y$df.w),mean(y$df.sg),
mean(y$strong2.df),mean(y$df.st),mean(y$strict2.df)),5,1)
CFI_diff=matrix(c(c1,c2,c3,c4,c5),5,1)
RMSEA_diff=matrix(c(r1,r2,r3,r4,r5),5,1)
tables=cbind(chi_dif,P_value,DF_diff,CFI_diff,RMSEA_diff)
colnames(tables)=c("chi_dif","P_value",
"DF_diff","CFI_diff","RMSEA_diff")
```

```
row2=matrix(c("Weak(1v2)","Strong(3v1)",
"Strong(3v2)","Strict(4v1)","Strict(4v3)"),5,1)
table2=cbind(row2,tables)

se_conf=sum(z$se.c)
se_w=sum(z$se.w)
se_sg=sum(z$se.sg)
se_st=sum(z$se.st)
FAIL_count=data.frame(se_conf,se_w,se_sg,se_st)

write.table(table,"C:\\Documents and Settings\\laalvarado
\\My Documents\\Andres\\Thesis\\Case-3, testlets in both groups
\\tables\\tables.xls",sep="\t",row.names=FALSE)
write.table(table2,"C:\\Documents and Settings\\laalvarado
\\My Documents\\Andres\\Thesis\\Case-3, testlets in both groups
\\tables\\table_diff.xls",sep="\t",row.names=FALSE)
write.table(FAIL_count,"C:\\Documents and Settings\\laalvarado
\\My Documents\\Andres\\Thesis\\Case-3, testlets in both groups
\\tables\\table_FAIL.xls",sep="\t",row.names=FALSE)
```

Curriculum Vitae

Luis Andres Alvarado was born on July 13, 1985. He is the first son of Luis and Marisol Alvarado, he graduated from Silva Health Magnet High School in 2003. He then entered the University of Texas at El Paso in hopes of pursuing a degree in Psychology. In 2003 he succeeded his goal and received a bachelors in science. However, in the process of pursuing his degree he became interested in Statistics. It was then that he decided to focus all of his attention to the study of Statistics. It took him two years to get caught up on all the required material for Statistics and in the spring of 2009 he was accepted to the Masters program in Statistics at The University of Texas at El Paso.

It was during his time as a graduate student that he worked as an research assistant at the Center for Institutional Evaluation, Research, and Planning. He also worked as a tutor at the M.A.R.C.S. center at the University library. Both jobs helped him realize the strong passion he has for teaching and research. After graduading Luis plans to takes his two passions and implement them both academically and in business.

Permanent address: 945 South Mesa Hills Apt 3101
El Paso, Texas 79912-4927