Random Permutation Testing Applied to Measurement Invariance with Dichotomous and Likert-type Indicator Variables

By

Benjamin A. Kite

Submitted to the graduate degree program in the Department of Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Committee members

Wei Wu, Chairperson

Holger Brandt

Kelsie Forbush

Paul Johnson

Jacob Fowles

Date defended: December 12th, 2016

The Dissertation Committee for Benjamin A. Kite certifies
that this is the approved version of the following dissertation :

Random Permutation Testing Applied to Measurement Invariance with Dichotomous and Likert-type Indicator Variables

Wei Wu, Chairperson

Date approved: February 10th, 2017

Abstract

An important detail that appears to be frequently overlooked in the SEM literature is how modeling data arising from responses on an ordered-categorical scale can influence measurement invariance testing. Typically tests for measurement invariance are conducted by comparing the fit of two nested models with chi-square difference testing. With ordered-categorical data the chi-square difference statistic measuring the discrepancy between two models does not follow a chi-square distribution (Muthén & Muthén, 2015), therefore chi-square difference testing is inappropriate. The popular solution to this problem is to use a scaling correction on the chi-square difference statistic to improve its chi-square approximation and test the resulting value for statistical significance (e.g., Garnaat & Norton, 2010; Randall & Engelhard, 2010). The purpose of the present research was to introduce and evaluate random permutation testing applied to measurement invariance testing with ordered-categorical data. The random permutation test builds a reference distribution from the observed data that is used to calculate a *p*-value for the observed chi-square difference value. The reference distribution is built by repeatedly shuffling the grouping variable and then saving the chi-square difference between the two models fitted to the resulting data. The present research consisted of two Monte Carlo simulations. The first simulation was designed to determine how many random shuffles of the grouping variable are appropriate. The second simulation was designed to evaluate random permutation testing across a variety of conditions in comparison to existing chi-square difference testing methods. Simulation results, an empirical example, and suggestions for the use of the random permutation test are provided.

Acknowledgements

I would like to thank my academic advisor Wei Wu for her mentorship and guidance during my time at the University of Kansas. I would also like to thank Paul Johnson for his mentorship and guidance while helping me become a better statistical programmer. Thank you to my remaining committee members, Holger Brandt, Kelsie Forbush, and Jacob Fowles for your helpful suggestions and valuable insights throughout this process.

I thank the Center for Research Methods and Data Analysis and the College of Liberal Sciences at the University of Kansas for access to their high performance compute cluster on which many of the calculations reported here were conducted.

Thank you to Po-Yi Chen and Terrance Jorgensen for the frequent discussions about my dissertation topic. You two helped me realize the potential benefits of this line of research.

Thank you to my amazing wife, Jessica. Without your support and encouragement I would not have made it this far in my academic pursuits.

Contents

1 Introduction **1**

1.1 Project Overview 1

1.2 Measurement Invariance 2

1.2.1 Configural Invariance 3

1.2.2 Metric Invariance 4

1.2.3 Scalar Invariance 5

1.3 Ordered-Categorical Data in CFA 6

1.3.1 Issues with Classic CFA Model 6

1.3.2 Ordered-Categorical Model 7

1.3.2.1 Latent Response Variables 7

1.3.2.2 Item Probability Curve 9

1.3.3 Residual Variance Parameterization 9

1.3.3.1 Theta Parameterization 10

1.3.3.2 Delta Parameterization 10

1.3.4 Model Estimation and Implementation 11

1.3.4.1 Thresholds and Polychoric Correlations 11

1.3.4.2 CFA Model Parameters and Fit 14

1.3.5 Ordered-Categorical CFA Assumptions 15

1.4 Measurement Invariance with Ordered-Categorical Data 16

1.4.1 Types of Tests 16

1.4.2 Chi-Square Difference Testing . . . 17
1.4.2.1 Satorra-Bentler Correction . . . 18
1.4.2.2 M*plus* DIFFTEST Command . . . 19
1.5 Present Research . . . 21
1.5.1 Limitations with Model Comparisons . . . 21
1.5.2 Random Permutation Testing . . . 22
1.5.2.1 Application To Measurement Invariance Testing . . . 24
1.5.2.2 Distinction from Bootstrapping . . . 24
1.5.2.3 Unequal Responses Categories Between Groups . . . 25
1.5.2.4 Examples . . . 26
1.5.2.5 Computation Time and Basic Syntax . . . 28
1.5.3 Research Questions . . . 29

2 Method **30**

2.1 Simulation One . . . 30
2.1.1 Data Generation and Analysis . . . 31
2.1.2 Between Replication Conditions . . . 32
2.1.2.1 Response Options . . . 32
2.1.2.2 Factor Loading Invariance . . . 33
2.1.3 Within Replication Conditions . . . 33
2.1.3.1 Number of Random Permutation Shuffles . . . 33
2.1.4 Outcomes . . . 34
2.1.4.1 Unequal Response Categories . . . 34
2.1.4.2 Type I Errors and Power . . . 34
2.2 Simulation Two . . . 35
2.2.1 Data Generation and Analysis . . . 36
2.2.2 Between Replication Conditions . . . 37
2.2.2.1 Latent Variable Distribution . . . 37

2.2.2.2 Response Categories . . . 38
2.2.2.3 Threshold Symmetry . . . 38
2.2.2.4 Sample Size . . . 39
2.2.2.5 Factor Loading Invariance . . . 39
2.2.3 Within Replication Conditions . . . 40
2.2.3.1 Chi-Square Difference Random Permutation Testing . . . 40
2.2.3.2 DIFFTEST Chi-Square Difference in M*plus* . . . 40
2.2.3.3 Satorra-Bentler Scaled Chi-Square Difference in lavaan . . . 41
2.2.3.4 Unadjusted Chi-square Difference Testing . . . 41
2.2.4 Outcomes . . . 41
2.3 Simulation Implementation . . . 42

3 Results **43**

3.1 Simulation One . . . 43
3.1.1 Unequal Response Categories . . . 43
3.1.2 Type I Errors . . . 44
3.1.3 Power . . . 46
3.1.4 Selected Number of Shuffles . . . 48
3.2 Simulation Two . . . 48
3.2.1 Normally Distributed Latent Variables . . . 49
3.2.1.1 Type I Errors . . . 49
3.2.1.2 Power . . . 49
3.2.1.3 Comparison With Existing Testing Procedures . . . 50
3.2.2 Non-Normal Latent Variable Conditions . . . 54
3.2.2.1 Comparisons With Existing Testing Procedures . . . 55

4 Empirical Example **59**

4.1 Data . . . 59

4.2 Analysis and Results . . . 61
4.2.1 Initial Test for Measurement Invariance . . . 61
4.2.2 Follow-up Tests . . . 62
4.3 Additional Testing Procedures . . . 63
4.4 Effect of Measurement Constraints . . . 64
4.5 Conclusions . . . 64

5 Discussion **66**
5.1 Research Questions . . . 66
5.1.1 Random Permutation Error Rate and Power . . . 66
5.1.2 Between Replication Differences . . . 67
5.1.3 Comparison with Scaling Corrections . . . 68
5.1.4 Group Differences on Latent Variable Distribution . . . 68
5.1.5 Number of Random Permutation Shuffles . . . 69
5.1.6 Summary . . . 70
5.2 Suggestions for Random Permutation Testing . . . 70
5.2.1 Response Category Sparseness Across Groups . . . 71
5.3 Limitations . . . 72
5.4 Directions for Future Research . . . 73
5.4.1 Additional Simulation Conditions . . . 73
5.4.2 Number of Shuffles . . . 73
5.4.3 Different Model Estimators . . . 74
5.4.4 Measures of Model Fit Differences . . . 74
5.4.5 Summary . . . 75
5.5 Conclusion . . . 75

A Polychoric Correlation Example **82**

B Random Permutation Test R Syntax **87**

C Further Evaluation of lavTestLRT **89**

List of Figures

1.1 Random Permutation Example 1 27
1.2 Random Permutation Example 2 28

3.1 Type I Error Rates Across Replications 45
3.2 Power Across Replications 47
3.3 Power with Normal Latent Variables 50
3.4 Type I Errors and Power Across Conditions and Testing Methods 53
3.5 Type I Errors and Power Across Conditions with Non-Normal LV 57

A.1 Log Likelihood as a Function of Polychoric Correlation 86

C.1 Evaluations of lavTestLRT with Delta Option 91

List of Tables

3.1 Number of Resamples and Collapses 44

3.2 Proportion of Replications with a Change in Decision with Population Invariance . 46

3.3 Proportion of Replications with a Change in Decision with Non-Invariance 48

3.4 Type I Error Rates with Normal Latent Variables 49

3.5 Power with Normal Latent Variables 49

3.6 Type I Errors Across Testing Methods 51

3.7 Power Across Testing Methods 51

3.8 Comparison of M*plus* DIFFTEST and Random Permutations 54

3.9 Type I Error Rates with Non-Normal Latent Variable 55

3.10 Power with Non-Normal Latent Variable 56

3.11 Comparison of M*plus* DIFFTEST and Random Permutations 58

4.1 Response Frequencies for USA and Canada 60

A.1 Response Frequencies for Two Items 82

A.2 Thresholds Used for Polychoric Correlation Computation 83

A.3 Cumluative Probabilities with Polychoric Correlation of .57 84

A.4 Response Pattern Probabilities with Polychoric Correlation of .57 84

C.1 Rejection Rates for lavTestLRT with Delta Method 90

Chapter 1

Introduction

1.1 Project Overview

An important detail that appears to be frequently overlooked in the SEM literature is how modeling data arising from responses on a Likert-type scale (e.g., 1 = *Strongly Disagree* to 5 = *Strongly Agree*) can influence tests for measurement invariance. Many of the common practices for measurement invariance testing that are appropriate for continuous data are inappropriate when the observed data are dichotomous or Likert-type (hereafter referred to as ordered-categorical). One important consideration with ordered-categorical data is how to compare nested models when testing for measurement invariance. With ordered-categorical data the chi-square difference statistic measuring the discrepancy in fit between two nested models does not follow a chi-square distribution because the assumption of multivariate normality is violated (Muthén & Muthén, 2015), therefore chi-square difference testing is inappropriate. Further, popular rules of thumb for the difference in alternative fit indices ($\Delta AFIs$), such as a confirmatory fit index difference (ΔCFI) greater than .01 (Cheung & Rensvold, 2002), do not perform well (Sass et al., 2014). The popular model comparison method is to use a scaling correction on the chi-square difference statistic to improve its chi-square approximation and test the resulting value for statistical significance (e.g., Garnaat & Norton, 2010; Randall & Engelhard, 2010). The most popular implementation of a

scaling correction is offered in M*plus* (Muthén & Muthén, 2015) with the DIFFTEST command.

The purpose of the present research is to introduce and evaluate random permutation testing, a non-parametric method, applied to chi-square difference testing with ordered-categorical data. Rather than rescaling the observed chi-square difference statistic between two models being compared, the random permutation test builds a reference distribution from the observed data that is used to calculate a *p*-value for the observed chi-square difference value. The important feature of the random permutation test is that no assumptions about the reference distribution are required, this removes the need to apply any scaling correction to the observed chi-square difference statistic. This chapter continues with a basic introduction to different types of measurement invariance and how they are typically evaluated. A detailed review of the unique features of the ordered-categorical CFA is also provided. The process of testing for measurement invariance with ordered-categorical data is discussed before reviewing existing methods for rescaling the chi-square difference statistic for hypothesis testing. A detailed overview of how random permutation testing can be implemented with worked examples is provided. The chapter concludes with a list of research questions that the present research attempts to answer.

1.2 Measurement Invariance

In structural equation modeling (SEM), researchers are often interested in comparing multiple groups, or time-points within individuals, on one or more latent variables. These comparisons can be on latent variable means, variances, covariances, or predictive relationships. An important requirement for such comparisons is the assumption of measurement invariance, or measurement equivalence. Measurement invariance for a set of manifest variables exists if the variables have the same measurement properties, or measurement parameters, across groups. Testing and meeting the assumption of measurement invariance allows any differences observed at the latent variable level (i.e., structural level) to be attributed to true differences in the latent variables rather than differences in the measurement of the latent variables across groups.

Vandenberg and Lance (2000), a frequently cited review of measurement invariance testing practices, outlined the following levels of measurement invariance: 1) equal covariance matrices across groups, 2) same measurement model for each group (configural invariance; Horn & McArdle, 1992; Meredith, 1993), 3) equal factor loadings for identical indicator variables across groups (metric invariance; Steenkamp & Baumgartner, 1998), 4) equal indicator intercepts and factor loadings across groups (scalar invariance; Steenkamp & Baumgartner), and 5) equal unique variances, intercepts, and factor loadings for indicators across groups (strict invariance; Meredith). Schmitt and Kuljanin (2008) provided a more updated review of common practices in measurement invariance testing, and they found that the test of equal covariance matrices across groups was essentially never used in their reviewed studies. Because of its lack of popularity, the test of equal covariance matrices will not be discussed further in this review. Additionally, because strict invariance is difficult to achieve and not necessary for group comparisons at the structural level, the current paper will not discuss strict invariance. The focus of the present paper is limited to tests for configural, metric, and scalar invariance.

1.2.1 Configural Invariance

The typical first step when testing for measurement invariance is to ensure that configural invariance holds across groups. Configural invariance is the least restrictive assumption about measurement equivalence, it asserts that the CFA model used to measure the latent variables(s) is identical in each group; however, no equality constraints are placed on the parameters. In other words, configural invariance implies that the same pattern of fixed and freely estimated measurement parameters exists in measurement models across groups (Horn & McArdle, 1992). Equation 1.1 helps to demonstrate the concept of configural invariance.

$$y_{ig} = v_g + \lambda_g \eta_{ig} + \varepsilon_i \tag{1.1}$$

The term y_{ig} indicates an indicator variable score for person i in group g. The latent variable

score for person i in group g is represented by η_{ig}, v_g represents the indicator intercept in group g, λ_g represents the indicator factor loading for group g, and ε_i is the residual term.

The test for configural invariance is more of a qualitative than quantitative task (Little, 2013). In other words, null hypothesis testing is not typically used to evaluate configural invariance; instead analysts fit a configural invariant model and examine its measurement parameters across groups and the overall model fit in order to make a decision about whether or not configural invariance is tenable. Analysts can use commonly accepted global model fit indices in order to assess configural invariance. Popular global fit index cutoffs are > .95 for the Tucker-Lewis index (TLI) and comparative fit index (CFI) and < .06 for the root mean squared error of approximation (RMSEA; Hu & Bentler, 1999). Once configural invariance is supported, or the initial model is altered to a configural invariant alternative (e.g., poorly performing indicators are removed from the model for all groups), subsequent tests of measurement invariance are permissible.

1.2.2 Metric Invariance

After configural invariance has been supported, the analyst can begin to make assertions about the equivalence of measurement parameters across groups. The first assertion is typically that the factor loadings for each indicator variable are equal across all groups/time-points, this is commonly referred to as metric invariance (Steenkamp & Baumgartner, 1998). Factor loading invariance has also been referred to as "weak invariance" (e.g., Widaman & Reise, 1997); however, weak invariance has also been used to refer to what this paper defines as configural invariance (Horn & McArdle, 1992). For the sake of clarity, the less confusing term "metric invariance" is used throughout this paper. Equation 1.2 provides a mathematical expression of the metric invariance model.

$$y_{ig} = v_g + \lambda \eta_{ig} + \varepsilon_i \tag{1.2}$$

This equation is similar to Equation 1.1 , but the subscript $_g$ for the factor loading (λ) has

been removed. In the metric invariance model, each indicator variable only receives a single factor loading estimate, and that estimate is used for measurement in all groups.

Researchers typically compare the metric model to the configural invariance model, and this is possible because the metric model is nested within the configural model. The metric invariance model is more restricted than the configural model, therefore it will have worse model fit. Evaluating metric invariance involves testing the null hypothesis that the indicator variable factor loadings are equal across groups. If the configural model does not have model fit that is significantly improved from the metric model, the observed data are not unlikely given the null hypothesis of equal factor loadings across groups and the null hypothesis is not rejected. Failing to reject the null hypothesis of equal factor loadings allows the analyst to assert that factor loadings are equal in the population. The null hypothesis test for metric invariance is typically done by assessing the statistical significance of the chi-square difference between the metric and configural ($\chi^2_{metric} - \chi^2_{configural}$) model with $df_{metric} - df_{configural}$ degrees of freedom.

1.2.3 Scalar Invariance

In order to make inferences about mean differences on latent variables across groups, it must first be established that the indicator variable intercepts (i.e., expected values when the latent variable is 0) are equal in all groups. Failure to meet this requirement will render tests of latent variable mean differences difficult to interpret because the effect of group membership cannot be separated from the true latent variable difference. Scalar invariance is tested by determining if the assumption of equal indicator intercepts across groups is tenable. Equation 1.3 shows the assumptions of scalar invariance which builds off of the metric invariance assumption shown in Equation 1.2.

$$y_{ig} = v + \lambda \eta_{ig} + \varepsilon_i \tag{1.3}$$

In Equation 1.3 the indicator intercept no longer has the group subscript $_g$, this indicates that a single intercept estimate is provided for each indicator variable and that estimate is applied to

respondents in all groups. Fitting a scalar invariance model allows the analyst to set a reference group to have a latent variable mean fixed to a certain value (typically 0) and then estimate the latent variable means for all other groups. The scale of the latent variable will be based on the scale of the reference group.

Scalar invariance is typically tested by comparing the fit of the scalar invariance model to the metric invariance model. If the scalar model fit is not significantly worse than the metric model, then scalar invariance is tenable; this is possible because the only difference between the scalar and metric invariance models is the constraints applied to the indicator variable intercepts. Any deterioration in model fit can be attributed to the added constraints on the intercepts. Scalar invariance can also be tested by comparing the scalar invariance model and the configural invariance model, this approach provides a joint test of metric and scalar invariance.

1.3 Ordered-Categorical Data in CFA

1.3.1 Issues with Classic CFA Model

There are numerous issues with applying the classic CFA model to ordered-categorical indicator variables. First and foremost, the linear relationship specified between the latent variable(s) and each indicator variable is not appropriate because a one-unit increase in a Likert-type indicator is meaningless because of unequal interval widths (O'Brien, 1985). In other words, the distance or difference between 1 (*Disagree*) and 2 (*Neutral*) on a Likert-type scale is not necessarily equivalent to the difference between 2 (*Neutral*) and 3 (*Agree*). Further, the intercept value estimated in the classic CFA model has little meaning with ordered-categorical indicators. For example, an intercept value of 2.45 would have a difficult interpretation on a five-point Likert-type scale with anchors of 1 (*Disagree*) and 5 (*Agree*). These issues require a model that can specify non-linear relationships between the ordered-categorical indicator variables and latent variables.

In many research scenarios, researchers may not be concerned about theoretical issues with applying the classic CFA model to ordered-categorical data as long as the model results can be trusted

and are useful. In addition to theoretical concerns, issues in parameter estimation occur when the classic CFA model is applied to ordered-categorical data. In short, measurement parameters are underestimated (Babakus et al., 1987; Johnson & Creech, 1983; Muthén & Kaplan, 1985) and model fit indices are biased (Brown, 2006). The underestimated factor loadings can be attributed to underestimated correlations between indicator variables, and the biased model fit indices arise from assumption violations of the classic CFA model. These issues can be overcome by using a model that correctly models the theorized data generation process and uses an estimation method with tenable assumptions.

1.3.2 Ordered-Categorical Model

1.3.2.1 Latent Response Variables

The classic CFA model assumes that latent variables have linear relationships with indicator variables. With an ordered-categorical model, rather than specifying linear relationships between the manifest variables and latent variables, the model specifies linear relationships between the latent variables and latent response variables (LRV; Muthén & Asparouhov, 2002) that underlie each manifest variable. The concept of a LRV is based on the assumption that each ordered-categorical manifest variable has a latent, typically normally-distributed, variable that gives rise to the observed response. For example, responses to a survey question about agreement with a statement on a five-point Likert-type scale with anchors of 1 (*Strongly Disagree*) and 5 (*Strongly Agree*) are believed to be determined by each respondents latent level of agreement with the statement.

The LRVs are linked to the observed responses by using threshold values that divide the continuous LRV into discrete categories. Equation 1.4 shows the relationship between the LRV (y^*) and the observed indicator variable response (y), where G is the number of response categories observed and $c = 1, 2, ..., G-1$. The value of τ_{c-1} represents the minimum value on the continuous LRV where the respondent will respond in category c, whereas τ_c represents the maximum LRV value for a respondent in category c. For every ordered-categorical indicator τ_0 is always fixed to $-\infty$ and τ_G is fixed to $+\infty$. This leaves $G-1$ thresholds parameters that are estimated.

$$y = c, \, if \; \tau_{c-1} < y^* \leq \tau_c \tag{1.4}$$

The formula for the LRV can be seen below in Equation 1.5, which was taken from Muthén and Asparouhov (2002), where y_i^* is the LRV for person i, v is the intercept parameter for the LRV, λ is the factor loading (i.e., the linear relationship between the LRV and the latent variable), η_i is the latent factor score for person i, and ε_i is the residual error.

$$y_i^* = v + \lambda \eta_i + \varepsilon_i \tag{1.5}$$

The expected value for the LRV is given in Equation 1.6, where v is the LRV intercept parameter, λ is the factor loading, and α is the latent variable mean. Equation 1.7 shows the variance of the LRV, where ψ is the latent variable variance and θ is the variance of the residual term ε_i which has a mean of 0.

$$\mu^* = v + \lambda \alpha \tag{1.6}$$

$$\sigma^* = \lambda^2 \psi + \theta \tag{1.7}$$

Typically $\mu^*$and $\sigma^*$are set to be equal to 0 and 1, respectively, in single sample designs. These constraints create standardized, normally distributed LRVs that are easily interpreted.

The theoretical formation of the LRVs allows the SEM to be fitted to the polychoric correlation (Lee et al., 1995) matrix for the indicator variables which provides the lower bound of the correlations between the LRVs. Using the polychoric correlation matrix overcomes the issue of underestimated measurement parameters that is encountered when the classic CFA model is used. When testing for measurement invariance with the ordered-categorical CFA model the data analyst is interested in group differences in factor loadings and thresholds rather than factor loadings and intercepts. The intercepts of the LRVs could be tested for invariance, however these are typically

set to be equal to 0 for all items in all groups because the scale of the LRVs is arbitrary.

1.3.2.2 Item Probability Curve

When conducting measurement invariance testing, one wants to determine if the responses to items are dependent on group membership. When analyzing ordered-categorical indicator variables this requires determining if the item probability curve for a given response is dependent on group membership. Equation 1.8 shows the formula for the probability of an item response (y) being equal to or greater than a category (c) given a latent variable score of η, with F being a distribution function determined by the assumption about the distribution of ε (typically assumed to be normally distributed). When a normal distribution is used, the input is treated as a z-score and the function returns the probability of observing a value equal to or less than the input.

$$P(y \geq c|\eta) = F[-(\tau_c - v - \lambda\eta)\theta^{-1/2}] \tag{1.8}$$

The assumption of measurement invariance asserts than the probability of observing a response (y) greater than or equal to a category (c) is equal across all N groups given that respondents have identical latent variable scores (η; see Equation 1.9). This assertion is made by stating that indicator variable thresholds and factor loadings are equal across groups.

$$P(y \geq c|\eta, Group1) = P(y \geq c|\eta, Group2) = \ldots = P(y \geq c|\eta, GroupN) \tag{1.9}$$

1.3.3 Residual Variance Parameterization

There are two parameterizations for residual, or unique factor, variances that can be used with the ordered-categorical CFA model: Theta and Delta. These different parameterizations are based on the idea that researchers can estimate residual variances in multiple group designs and they can estimate LRV variance in multiple group designs, but they cannot estimate both simultaneously. Theta and Delta parameterization are options in both M*plus* and the R (R Core Team, 2016) pack-

age lavaan (Rosseel, 2012).

1.3.3.1 Theta Parameterization

With Theta parameterization, the residual variances for the LRVs are fixed to 1 in the first group and estimated for all other groups. This leaves the LRV variance (commonly referred to as "scaling factor"; Δ) to be obtained as a remainder. The scaling factor is equal to the inverse of the LRV standard deviation (see Equation 1.10).

$$\Delta = 1/\sqrt{\sigma^*} \tag{1.10}$$

In Equation 1.10, Δ is the scaling factor and σ^* is the variance of the LRV. This can also be rewritten as seen in Equation 1.11.

$$\Delta^{-2} = \sigma^* \tag{1.11}$$

With the theta parameterization, the residual variance for each group (with the exception the first group) is estimated and the scaling factor is obtained as a remainder as shown in Equation 1.12.

$$\Delta_g^{-2} = \lambda_g^2 \psi_g + \theta_g \tag{1.12}$$

Here the residual variance for the first group (θ_1) is fixed to 1, whereas all other groups have θ estimated.

1.3.3.2 Delta Parameterization

Delta parameterization requires the scaling factor to be fixed to 1 in the first group but estimated for all other groups (Muthén & Muthén, 2015), which then leaves the residual variance to be calculated as a remainder. This is shown in Equation 1.13

$$\theta_g = \Delta_g^{-2} - \lambda_g^2 \psi_g \tag{1.13}$$

Specifically, Δ_1 is fixed to 1 in the first group, whereas all other groups have Δ estimated. Then the resulting model parameters are used to compute the residual variance for each indicator variable in each group. The computation of the residual variance is simplified when the scaling factor (Δ_g^{-2}) and the latent variable variance (ψ_g) are set to 1, which puts the residual variance on a scale of 0 to 1 because the factor loading (λ_g) is then bound between -1 and 1 (see Equation 1.14).

$$\theta_g = 1 - \lambda_g^2 \tag{1.14}$$

Importantly, the M*plus* and lavaan defaults are to use Delta parameterization. When conducting measurement invariance testing researchers are forced to constrain either scaling factors or residual variances to be equal across groups for the sake of model identification (e.g., Sass et al., 2014). In these situations, it becomes more clear why Delta parameterization would be preferred; researchers typically avoid constraining residual variances to be equal across groups (Vandenberg & Lance, 2000). Delta parameterization will be used in all Monte Carlo simulations conducted in the proposed research.

1.3.4 Model Estimation and Implementation

There are two commonly used main steps in the process for estimating parameters for the ordered-categorical CFA model. First the variable thresholds and polychoric (or tetrachoric with dichotomous variables) correlation matrix, which estimates the lower bounds of the correlations between all LRVs, need to be computed. Second, the thresholds and polychoric correlation matrix are used with a version of the weighted least squares estimator in order to estimate the model parameters and model fit.

1.3.4.1 Thresholds and Polychoric Correlations

The computation of thresholds and polychoric correlations can be done two different ways. Thresholds and polychoric correlations can be estimated simultaneously using maximum likelihood (ML)

estimation, or they can be estimated in a two-step method (Olsson, 1979). In order to provide more detail, the two-step method is discussed here. The methods discussed here pertain to models with no exogenous manifest variables, which is typically the case with measurement invariance testing models. First the thresholds for each ordered-categorical variable are estimated based on the proportions of responses in each category. The response proportions are then used to compute cumulative proportions that are then used as input in the inverse of the normal cumulative distribution function (CDF) shown in Equation 1.15 which yields threshold estimates. The inverse of the normal CDF is commonly referred to as a probit function (Bliss, 1934). The normal CDF equation which yields the proportion of area of a normal curve to the left of a threshold can be used in R with the "pnorm" function, whereas the probit function which yields the threshold for a given proportion can be used in R with the "qnorm" function.

$$F(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} exp\left(\frac{-(y)^2}{2}\right) dy \tag{1.15}$$

In models with exogenous manifest variables the thresholds can be calculated using ordered-probit regression, for example in R using the "MASS" package (Venables & Ripley, 2002) or "oprobit" in Stata (StataCorp, 2015).

After thresholds have been computed, polychoric correlations are estimated with ML estimation. For any two manifest variables, consider a contingency table with r rows and s columns that shows the frequency of responses for manifest variable 1 (with r response categories) and manifest variable 2 (with s response categories). The ML estimator finds the polychoric correlation ρ that maximizes the likelihood of the observed contingency table. The log likelihood function that is maximized is shown in Equation 1.16.

$$log(L) = log(K) + \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} log(\pi_{ij}) \tag{1.16}$$

In Equation 1.16, n_{ij} is the number of responses in row i and column j of the contingency table, and π_{ij} is the probability of an observation being in row i and column j which is a function of the

thresholds (see Equation 1.17).

$$\pi_{ij} = \Phi_2(\tau_{xi}, \tau_{yj}) - \Phi_2(\tau_{xi-1}, \tau_{yj}) - \Phi_2(\tau_{xi}, \tau_{yj-1}) + \Phi_2(\tau_{xi-1}, \tau_{yj-1}) \tag{1.17}$$

In Equation 1.17, the bivariate normal CDF is applied to the response pattern of interest, as well as adjacent response cells in the contingency table. The values of τ_{xi} and τ_{yi} represent upper threshold values for given responses of i on the two variables of interest (e.g., τ_{x1} is the threshold separating a response of 1 from a response of 2). These probabilities are required to compute the probability of a respondent being in (i, j), rather than the simple cumulative probability provided by $\Phi_2(\tau_{xi}, \tau_{yj})$. The bivariate normal CDF formula can be seen in Equation 1.18.

$$\Phi_2(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} exp\left[-\frac{x^2+y^2-2\rho xy}{2(1-\rho^2)}\right] \tag{1.18}$$

As was previously mentioned, ρ is the parameter that Equation 1.16 estimates by maximizing the log likelihood function. This estimation procedure is applied to all non-redundant pairs of manifest variables to produce the polychoric correlation matrix. Fortunately, researchers wanting to simply estimate the polychoric correlation matrix for their data can use existing software rather than manually implementing the equations provided in this subsection. The R package "polycor" (Fox, 2010) can calculate a polychoric correlation between any two ordered-categorical variables using the methods described here. The "polycor" package can do the two-step method described here as well as the full maximum likelihood approach which estimates thresholds and polychoric correlations simultaneously. In models with exogenous manifest variables, the polychoric correlations are the correlations between the residuals of the endogenous variables as would be obtained with bivariate probit regression (e.g., "biprobit" in Stata). A worked example of manually computing a polychoric correlation estimate can be seen in Appendix A.

1.3.4.2 CFA Model Parameters and Fit

When conducting CFA with ordered-categorical indicators, the maximum likelihood estimator no longer permits the computation of popular alternative fit indices (AFIs) such as the CFI, RMSEA, and TLI. Rather than using an estimation method that maximizes a likelihood function, ordered-categorical models are typically fitted with an alternative estimation procedure. Weighted least squares estimation, which minimizes the discrepancy between the observed and model-implied polychoric correlations and thresholds, is the standard estimator for ordered-categorical variable models. The formula for the WLS estimator can be seen in Equation 1.19.

$$F_{WLS} = (s-\sigma)^T W^{-1}(s-\sigma) \tag{1.19}$$

The observed polychoric correlations and thresholds are represented as s, whereas the model implied polychoric correlations and thresholds are represented by σ. The W matrix represents the weight matrix which summarizes the variances and covariances of the covariances of all elements in s (Brown, 2006). Specifically, W^{-1} is the inverse of a positive definite u x u matrix where $u = p(p+1)/2$, and p is the number of elements in s (Browne, 1984; Wirth & Edwards, 2007). The major limitation with the WLS estimator is the inversion of the weight matrix, which becomes impractical when the number of indicator variables is large.

Because of the limitations with the WLS estimator, currently the most popular estimator for CFA with ordered-categorical indicator variables is weighted least squares with a mean and variance adjusted chi-square statistic. M*plus* was the first SEM software to implement this estimator; the developers gave it the name "WLSMV," which has become the popular shorthand used in other SEM software packages such as lavaan. An important feature of the WLSMV estimator is that it does not require the full weight matrix shown in Equation 1.19. Instead with WLSMV only the diagonal of W, which represents the variances of the covariances of s, is used. Using the diagonal of W is advantageous because the weight matrix no longer has to be positive definite and the entire matrix is not inverted. If p polychoric correlations and thresholds are in s, its diagonal weight

matrix will have $p(p+1)/2$ diagonal elements.

The removed requirement of inverting the weight matrix allows WLSMV to have a much lower sample size requirement than standard WLS. However, WLSMV does require that the number of observations is greater than the number of diagonal elements in the weight matrix. A basic representation of the WLSMV estimator can be seen in Equation 1.20.

$$F_{WLSMV} = (s-\sigma)^T W_D (s-\sigma) \tag{1.20}$$

In Equation 1.20 the W_D term represents the diagonal weight matrix. Further, the WLSMV estimator adjusts the chi-square statistic, standard errors, and model fit indices. This is done by using the full weight matrix W to compute standard errors and a mean and variance adjusted chi-square statistic. All subsequent discussions of the performance of the ordered-categorical CFA model pertain to evaluations with the WLSMV estimator. More information about the WLSMV estimator can be found in Muthén and Muthén (2015). Because of its popularity with ordered-categorical indicator variables, the present research focuses exclusively on models estimated with WLSMV.

1.3.5 Ordered-Categorical CFA Assumptions

When fitting the ordered-categorical model, researchers are making important assumptions about their data. There are six key assumptions that are being made with the ordered-categorical CFA model. Assumptions 1-3 are completely unique to the ordered-categorical CFA model and are vital to correct interpretation of parameter estimates. Assumptions 4-6 also apply to the classic CFA model but are still true in the special case of the ordered-categorical model.

1. Each observed variable has an underlying LRV that is standard normal (see Equation 1.21). In multiple group designs the assumption of a LRV variance equal to 1 can be relaxed in all but one group.

$$LRV \sim N(0,1) \tag{1.21}$$

2. All manifest variables are ordered-categorical, not nominal. Each one-unit increase all vari-

ables must consistently indicate an increase or decrease in a LRV.

3. Any relationship between a LRV and latent variable is linear.

4. The model is correctly specified and identified. Specifically, all relationships between latent variables and manifest variables are included in the model, and the number of parameters estimated does not exceed the number of observed polychoric correlations and thresholds.

5. The unique factors are normally distributed with a mean of 0 and variance of θ (see Equation 1.22). Typical CFA models also assume that unique factors are uncorrelated, but this assumption can be relaxed for certain pairs of unique factors to improve model fit.

$$\varepsilon \sim N(0, \theta) \tag{1.22}$$

6. All latent factors are normally distributed with a mean of 0 and a variance of Ψ (see Equation 1.23). In multiple group designs the mean for a latent variable can be different from 0 in all but one group if the model is properly specified (e.g., a scalar invariance model). Further, the latent variable variance can be fixed to 1 in at least one group for model identification purposes.

$$\eta \sim N(0, \Psi) \tag{1.23}$$

1.4 Measurement Invariance with Ordered-Categorical Data

1.4.1 Types of Tests

There are popular SEM resources for psychology researchers which suggest that researchers test for configural, metric, and then scalar invariance sequentially (e.g., Kline 2016, p. 399; Vandenberg & Lance, 2000). These assertions appear to be made as a simple extension of the popular methods that are applied to continuous, multivariate normal, indicator variables. These recommen-

dations are problematic because the vast majority of variables in psychology research are ordered-categorical, not continuous (Lubke & Muthén, 2004). A review of relevant research found no simulation research evaluating the performance of the free-baseline model building approach to measurement invariance testing with ordered-categorical indicator variables (this is likely due to the fact that the free-baseline approach is difficult to implement in simulation designs). Muthén and Muthén (2015) recommend constraining factor loadings and item thresholds in tandem when testing for measurement invariance. The basis of this recommendation is the fact that the item probability curve (see Equation 1.8), or the probability of an item response given the latent variable, is dependent on both measurement parameters. This issue with both parameters influencing the item probability curve was discussed in detail by Muthén and Asparouhov (2002). Despite the strong recommendations by Muthén and Muthén, some researchers still apply the free-baseline approach to measurement invariance testing when using the ordered-categorical CFA model (e.g., Cyders, 2013, Skriner & Chu, 2014). In order to stay consistent with currently recommended best practices, the current research will focus on testing metric and scalar invariance in simultaneously.

After testing for metric and scalar invariance simultaneously, researchers can follow-up significant tests with further evaluations of measurement invariance. Individual indicator variables can have their measurement parameter constraints manipulated to allow nested model comparisons testing for invariance in individual variables. In addition to being able to test individual variables for invariance, individual indicator measurement parameters can also be tested. The factor loadings for one or multiple indicator variables can be constrained to equality across groups in one model and freely estimated in another, this creates a test for metric invariance. These follow-up tests use the same testing procedures discussed in the following sections when comparing the nested models for statistical significance.

1.4.2 Chi-Square Difference Testing

Chi-square difference testing is frequently used when testing for measurement invariance (Brown, 2006; Vandenberg & Lance, 2000). The chi-square fit of the more restricted model is compared to

the chi-square fit of the less parsimonious model. When the multivariate normality assumption is met, the chi-square difference between two nested models follows a chi-square distribution. The degrees of freedom for the chi-square difference statistic is equal to the difference in degrees of freedom between the two nested models. This allows a null hypothesis test of the equality of the group measurement parameters which are constrained to equality in the more restricted model. The *p*-value from this test informs the researcher of the probability of the observed data given that the null hypothesis of equal measurement parameters across groups is true. This *p*-value is unbiased when the chi-square difference statistic follows the chi-square distribution. When multivariate normality is not present, the chi-square difference statistic is no longer chi-square distributed and should not be used for null hypothesis testing (Satorra, 1999). This is true even when the individual chi-square values are scaled to account for non-normality (Satorra & Bentler, 2001). Fortunately, there are corrections to the chi-square difference statistic that can be applied to improve its chi-square approximation and allow null hypothesis testing with models estimated with WLSMV. A limitation with the methods discussed here is that they are only readily available in certain software packages.

1.4.2.1 Satorra-Bentler Correction

The Satorra-Bentler (SB) correction offers a method for improving the chi-square approximation of chi-square difference values when multivariate normality is violated. A valuable feature of the SB correction is the flexibility that allows it to be used with WLS or ML estimators. The general concept of the SB correction is expressed in Equations 1.24 and 1.25.

$$T_D = T_1 - T_0 \tag{1.24}$$

$$\bar{T}_D = T_D / \tilde{c} \tag{1.25}$$

Here T_D is the chi-square difference between nested model (T_1) and its parent model (T_0). The test statistic used for hypothesis testing is $\bar{T}_D$, which is computed using T_D and a scaling parameter $\tilde{c}$. The computation of the scaling parameter is intensive when an asymptotic distribution free (ADF) estimator such as WLSMV is used. See Satorra (1999) and Satorra & Bentler (2001) for more detailed discussions and illustrations of the computations involved in the SB correction statistic.

The SB correction can be seen most frequently with models using some form of ML estimation (e.g., Carpenter et al., 2010; Osman et al., 2014). When using the WLSMV estimator the SB correction for chi-square difference testing can be applied using the "lavTestLRT" function from the lavaan package in R. When two nested models are supplied to the lavTestLRT function, the correction outlined by (Satorra, 1999) is applied to produce a scaled chi-square difference statistic. The present research is designed to evaluate the performance of the lavTestLRT function specifically. A review of the literature suggested that the present research is the first Monte Carlo simulation to evaluate the performance of the lavTestLRT function. The lavTestLRT function was evaluated using all of the default options in version 0.5-22.

1.4.2.2 M*plus* DIFFTEST Command

M*plus* has a unique implementation for scaling chi-square difference values when the WLSMV estimator is used. Muthén and Muthén (2015) suggest that researchers use the DIFFTEST command in M*plus* in order to correctly scale chi-square difference values. This correction can frequently be seen in applied research (e.g., Garnaat & Norton, 2010; Randall & Engelhard, 2010). Further, the DIFFTEST command is recommended by popular SEM textbooks (e.g., Kline, 2016; Little, 2013). The DIFFTEST command applies a modification to the method used by the SB correction that is discussed by Asparouhov & Muthén (2006). The parent model is fitted to the data and matrices containing information about the model are saved in a separate output file. When the nested model is fitted and the text file containing matrices from the parent model is provided, DIFFTEST uses information from both models to compute a scaled chi-square difference statistic. A more detailed, albeit brief, explanation of the computation involved with the DIFFTEST com-

mand can be seen in Asparouhov & Muthén. The computational advantage of M*plus* DIFFTEST over the Satorra-Bentler correction is that the former does not use the possibly large sample statistic variance/covariance matrix (see Asparouhov & Muthén). The main difference between the Satorra-Bentler correction and the M*plus* DIFFTEST procedure that users will notice is the difference in Δdf. With the Satorra-Bentler correction Δdf is adjusted and is not necessarily an integer, whereas with M*plus* DIFFTEST Δdf is the same as what would be observed with an unadjusted test.

After introducing the DIFFTEST command, Asparouhov & Muthén (2006) conducted a small Monte Carlo simulation to evaluate its performance. Data were generated for two groups with six indicator variables with three response options per indicator and equal factor loadings and thresholds across groups (making population measurement invariance true). Using 500 simulation replications with sample sizes of 1,100 and 2,200, configural and scalar invariance models were fitted to the generated data. With a sample size of 1,100 measurement invariance was rejected 6% of the time, whereas with a sample size of 2,200 measurement invariance was rejected 4.4% of the time. These values were close enough to the nominal value of 5% for the authors to conclude that the DIFFTEST command performed appropriately with Type I error control.

More recently, Sass et al. (2014) evaluated the M*plus* DIFFTEST command along with other model comparison criteria when testing five-point ordered-categorical data in a Monte Carlo simulation. The authors found that the popular ΔCFI cutoff value of .01 (Cheung & Rensvold, 2002) and the $\Delta RMSEA$ cutoff of .01 (Chen, 2007) had inflated Type I error rates when comparing nested models estimated with WLSMV. The scaled chi-square difference value provided by M*plus* DIFFTEST showed Type I error rates between .061 and .090, which is slightly higher than the desired rate of .050. Further, while also simulating five-point ordered-categorical data, Suh (2015) showed that the M*plus* DIFFTEST command is sensitive to group differences in latent variable distributions when conducting measurement invariance testing in the item response theory (IRT) framework. Specifically, Suh showed that Type I error rates are inflated when the studied groups have greatly different latent variable distributions. Specifically, when one group has a normally

distributed latent variable ($M = 0$, $SD = 1$, $skewness = 0$, $kurtosis = 0$) and the other had a non-normally distributed latent variable ($M = 0$, $SD = 1$, $skewness = 1.5$, $kurtosis = 3.5$) the Type I error rate was .22. These results show that the M*plus* DIFFTEST command has an issue with Type I error control, and the limitation of sensitivity to assumption violations.

1.5 Present Research

1.5.1 Limitations with Model Comparisons

Unfortunately, the majority of past research (e.g., Cheung & Rensvold, 2002; Vandenberg & Lance, 2000) that is frequently cited by psychology researchers pertains to methods that were evaluated for situations with continuous indicator variables, which are uncommon in psychology research. Research establishing new recommendations for model comparison methods with ordered-categorical data would make an important contribution to the literature. Sass et al. (2014) showed that popular cutoff values for $\Delta AFIs$ do not perform well, therefore their use when conducting measurement invariance testing with ordered-categorical indicators should be discouraged. The popular choice for model comparisons for researchers is chi-square difference testing with a scaling correction as implemented by the M*plus* DIFFTEST command (e.g., Garnaat & Norton, 2010; Randall & Engelhard, 2010). Researchers conducting their analysis in R using the lavaan package can use the lavTestLRT function to perform a similar scaling correction. Research providing a viable alternative to M*plus* DIFFTEST and lavTestLRT for chi-square difference testing would make a valuable contribution to the literature. There are two important limitations with existing scaling corrections. One limitation is that the scaling correction available to researchers is determined by the statistical software being used. Another limitation is the performances of the existing scaling corrections. Previous simulation research has shown that the M*plus* DIFFTEST correction has slightly inflated Type I error rates, and no simulation work evaluating the performance of lavTestLRT can be found. A chi-square difference testing procedure that is not dependent on a certain software package, and can consistently control Type I errors would clearly make a valuable contribution as a tool for

researchers.

The present research is designed to evaluate the random permutation testing approach to chi-square difference testing when testing for measurement invariance with ordered-categorical indicator variables. The random permutation testing approach is not computationally intensive beyond what is required to fit a single CFA model with the WLSMV estimator, therefore it would be relatively easy to implement in any statistical software package. Further, the theory behind random permutation testing procedure is centered around keeping the Type I error rate at .05, therefore it should control Type I errors as well or better than existing scaling corrections.

1.5.2 Random Permutation Testing

Random permutation testing can be applied to chi-square difference testing with models estimated with WLSMV to overcome the issue of the difference statistic not following a chi-square distribution. The focus of the present research is demonstrating how this approach works and evaluating its performance. The proposed random permutation test is a non-parametric method based on the idea of building a reference distribution under the assumption that groups have the same measurement parameters. In other words, the reference distribution is built under the assumption that the null hypothesis of no effect of group membership on measurement is true. This reference distribution is used to calculate a *p*-value when testing the null hypothesis. The permutation testing method was first used more than eighty years ago when Fisher (1935) applied it with paired data, and then shortly afterwards Pitman (1937) applied it to test the statistical significance of correlation coefficients. A variety of other uses for permutation testing have been covered by Higgins (2004). The benefit of permutation testing is that building a non-parametric reference distribution alleviates many of the assumptions of standard parametric hypothesis tests. This is advantageous in situations where researchers know an assumption (e.g., all variables are normally distributed) about their data is violated. The major limitation with random permutation testing is an increase in computation time resulting from repeatedly manipulating data and computing a test statistic to build a reference distribution.

When testing for the effect of group membership on a test statistic, a null distribution can be built by randomly shuffling the grouping variable and saving the resulting test statistic after each shuffle. If there is no effect of group membership on measurement parameters, the observed test statistic (calculated from the original data) is just as likely as the values created by randomly shuffling the grouping variable and the observed value will only be in the upper 95th percentile five percent of the time. This should keep the Type I error rate of the test procedure at .05. Building a null distribution this way is especially useful when the distribution of the test statistic is unknown. For a true permutation test the grouping variable would be reshuffled for all possible combinations of group membership. The formula for the total possible number of permutations can be seen in Equation 1.26, where n_1 and n_2 represent the sample sizes for two groups, and N is the total sample size.

$$Combinations = \frac{N!}{n_1!n_2!} \tag{1.26}$$

In many research designs using all possible combinations of group membership with permutation testing would result to far too many combinations to analyze in a reasonable amount of time. Instead of using all possible permutations, one can sample permutations by randomly shuffling the grouping variable a fixed number of times (Higgins, 2004). The p-value obtained with the random shuffles serves as an estimate of the p-value that would be obtained if all possible grouping possibilities were used. The p-value obtained from comparing the observed statistic to the randomly permuted reference distribution will have a standard error that is a function of the p-value and the number of permutations. This can be seen in Equation 1.27, where p is equal to the p-value and R is equal the number of permutations used (Higgins).

$$SE_p = \sqrt{\frac{p(1-p)}{R}} \tag{1.27}$$

This equation shows that the number of permutations used is an important factor to consider. The effect of the number of permutations used will be explored in the present research.

1.5.2.1 Application To Measurement Invariance Testing

Random permutation chi-square difference testing would require the original data to have the grouping variable randomly shuffled *m* times. After each of the *m* shuffles the resulting data are used to fit the configural and scalar invariance models, and then the chi-square difference value between the two models is saved. After all random permutation shuffles are complete the observed chi-square difference value is compared to the *m* permuted chi-square difference values. The proportion of randomly permuted chi-square difference values greater than the observed chi-square difference value is the *p*-value. A *p*-value less than or equal to .05 indicates that the the observed data are unlikely given that the null hypothesis is true, therefore the null hypothesis is rejected and the researcher concludes that differences in measurement exist across groups.

The important benefit of the random permutation chi-square difference test is that it can be implemented in any software package without requiring complex formulas. The random permutation test only requires the ability to fit individual CFA models, the ability to randomly shuffle the grouping variable, and the ability to do this repeatedly and save the results. Even if the statistical program used to fit the CFA models does not have the ability to randomly shuffle the grouping variable and repeat the fitting process automatically, statistical programs such as R or general programming languages such as Python (Rossum, 1995) can handle those tasks.

1.5.2.2 Distinction from Bootstrapping

The random permutation testing method clearly resembles the more popular bootstrapping approach (Efron, 1979) to building empirical distributions. Because bootstrapping does not allow group membership to change, it cannot approximate the sampling distribution of a population where group membership does not influence measurement parameters. Using random permutation testing and bootstrapping to build a chi-square difference reference distribution should yield similar distributions when there are no group differences in population measurement parameters. Differences between the two methods would be seen when differences in population measurement parameters exist. The chi-square reference distribution that would be created with bootstrapping

would be influenced by both sampling variability and group differences in measurement parameters. The benefit of random permutation testing is that randomly shuffling the grouping variable removes the effect of group measurement parameter differences on the chi-square difference reference values. This is desired because the reference distribution should approximate the sampling distribution of the chi-square difference statistic where group membership does not influence measurement parameters.

1.5.2.3 Unequal Responses Categories Between Groups

Applying the random permutation test to ordered-categorical data presents the unique problem of random permutation shuffles possibly having unequal response categories for at least one indicator. When this occurs, the same thresholds cannot be estimated in all groups making testing the equality of the thresholds impossible. There are two potential solutions for this problem. The first method is to collapse response categories when necessary to ensure that both permuted groups have an equal number of response categories for all items. This method will allow all possible permutations to be eligible for sampling when building the reference distribution, however a potential disadvantage will be variations in the number of parameters estimated due to collapsing and fewer thresholds being estimated. For each category collapsed, the degrees of freedom difference between the configural and scalar invariance models decreases by 1. This could result in reference distributions that are negatively biased, which could increase Type I error rates. Although collapsing categories is the only option when this occurs with observed data, another option exists when this is observed with a random permutation shuffle of the grouping variable. An alternative method for dealing with sparseness is to simply discard the random permutation shuffle with unequal response categories for any items and replace it with a draw where an equal number of response categories is observed for both groups. This process will ensure that all random permutations have a chi-square difference statistic with degrees of freedom equal to the test statistic from the observed data. In the present research all instances of unequal response categories occurring in data generated from a population model used the aforementioned collapsing method. All instances of unequal response categories

observed while conducting the random permutation test used resampling.

1.5.2.4 Examples

A basic example of random permutation testing can be created by simulating four dichotomous variables that measure a single latent variable. This example uses data generated for two groups with factor loadings of .60 and thresholds of 0 for all four variables. Because the data were generated with the same measurement parameters, configural and scalar invariance models would not be expected to have a statistically significant difference in model fit. The chi-square difference between the configural and scalar models for the original data is 1.656. A reference distribution is needed to assess the probability of the observed value of 1.656 or greater, given that the null hypothesis is true and the factor loadings and thresholds are equal across both groups. After shuffling the grouping variable, fitting the configural and scalar models, and saving the chi-square difference each time, the resulting chi-square difference values can be used to build the distribution shown in Figure 1.1. Here 94.8% of the values in the reference distribution are greater than 1.656, therefore the null hypothesis of equal factor loadings and thresholds across groups cannot be rejected.

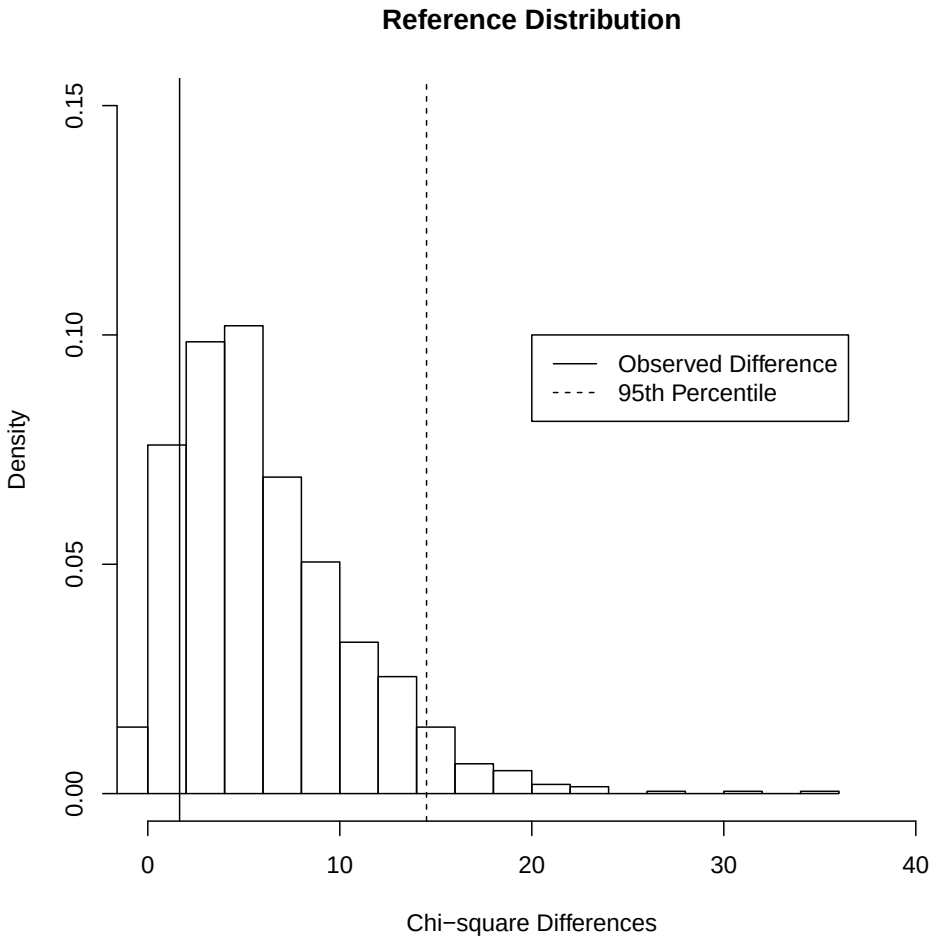

Figure 1.1: Random Permutation Example 1

For an example of random permutation testing when the null hypothesis of equal factor loading and thresholds is false, the first example can be altered so that one group has factor loadings of .30, rather than .60, on two of the four indicators. The generated data have a chi-square difference of 16.829. The grouping variable in the original simulated data can be shuffled 1,000 times to build a reference distribution for the chi-square difference statistic, which is shown in Figure 1.2. Here only 2.9% of the values in the reference distribution are greater than 16.829, therefore the null hypothesis of equal factor loadings and thresholds is rejected.

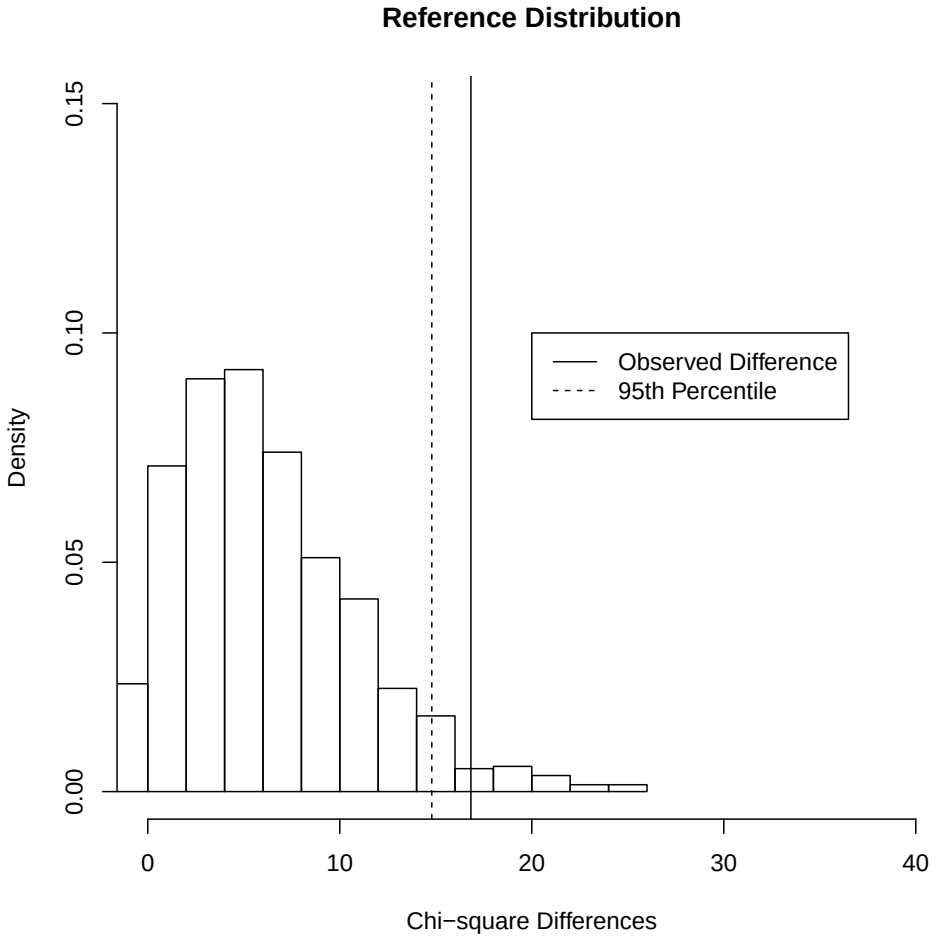

Figure 1.2: Random Permutation Example 2

1.5.2.5 Computation Time and Basic Syntax

The random permutation testing procedure does have the downside of increased computation time. Although the Monte Carlo simulations used in the present research were computationally demanding, running the random permutation test once to test a hypothesis is easily done on a single computer. To provide an idea of the expected computation time required, the second example in the previous section required about six minutes to run with 1,000 shuffles on a single computer (without any multicore processing). The R syntax used to create the second example (including code to generate the data) can be seen in Appendix B.

1.5.3 Research Questions

The present research serves as an evaluation of the random permutation test applied to measurement invariance chi-square difference testing with ordered-categorical indicator variables. The random permutation test is tested and compared with existing chi-square difference adjustment options. The present research addresses the following questions about random permutation testing applied to measurement invariance. These questions guide the Monte Carlo simulations that are outlined in the subsequent chapter.

- How does chi-square difference testing using random permutations perform in terms of Type I error rate and power? Are Type I errors close to .05? If Type I errors are close to .05, does the test show reasonable power that increases as a function of sample size?
- How do sample size, the number of response categories, and threshold symmetry influence the performance of chi-square random permutation testing?
- How does the random permutation testing approach perform compared to DIFFTEST in M*plus* and other chi-square difference test implementations?
- Does the random permutation test outperform DIFFTEST in M*plus* when groups have different latent variable distributions?

In addition to the four main research questions, there was an important question about random permutation testing implementation that needed to be answered. This question was answered first because because the answers were used to inform the design of the second Monte Carlo simulation.

- How does the number of permutations used influence the performance of the random permutation test? What is an appropriate number of random group shuffles to use?

After these questions are answered with two Monte Carlo simulations, an empirical example is used to show how the random permutation test compares with existing chi-square difference testing procedures.

Chapter 2

Method

The research questions for the present research were answered with two separate Monte Carlo simulations. The Monte Carlo simulation work was split into two parts to ease the computational strain created by running thousands of random permutation tests. For the sake of clarity, it is important to outline terminology before discussing the simulation designs. In the present research, the term "shuffle" is used to refer to a random permutation shuffle, whereas the term "replication" will be used to refer to a drawing a single sample from the population parameter space and analyzing it with random permutation testing as well as other testing approaches. Each simulation consisted of a certain number of replications for each simulation condition (i.e., combination between replication variables), and each individual replication had a certain number of random permutation shuffles used to build a reference distribution for random permutation testing. Further, all tests for measurement invariance consisted of testing metric and scalar invariance in tandem by comparing the fit of a configural invariance model to a scalar invariance model.

2.1 Simulation One

The first Monte Carlo simulation was used to answer research question five:

- How does the number of permutations used influence the performance of the random per-

mutation test? What is an appropriate number of random group shuffles to use?

The goal of this simulation was to determine how the number of random permutation shuffles used can influence null hypothesis rejection rates. Prior to conducting a larger simulation comparing competing chi-square difference testing procedures, the appropriate minimum number of permutations for the present research was identified. Because random permutations are time consuming, it is important to avoid using an excessive number of permutations in each replication when conducting a Monte Carlo simulation. Importantly, this simulation simply aimed to determine the appropriate number of permutations for subsequent Monte Carlo simulations following the same basic structure. Providing a rule of thumb or guidelines for applied researchers was not the goal of the present simulation. A simple 2 (response options) x 2 (factor loading invariance) design with four between replication conditions and 1,000 replications in each condition was used. The number of conditions in this simulation were limited because of the computational demands using 1,000 replications with 1,000 random permutation shuffles. The response option and factor loading invariance variables were chosen because they were expected to have the largest influence on the shape of the reference distribution created in each replication. In each between replication condition, ten different random permutation tests were conducted with different numbers of random permutation shuffles (100-1,000 in increments of 100).

2.1.1 Data Generation and Analysis

Data for two groups were simulated in the R program with the lavaan package using the "simulateData" function. A single latent common factor with eight indicator variables was simulated, with 150 cases present in each group. Eight indicators were used in order to represent a reasonable number of indicators that are used to measure a latent variable in applied research. The latent common factor had a mean of 0, a variance of 1, and was normally distributed for both groups in the population model. In order to handle generated data where a variable had unequal observed responses across groups, categories were collapsed to ensure that response options matched in all "observed" data. For example, if generated data showed responses of 1, 2, 3, 4, 5 in group A, but

1, 2, 3, 4 in group B, all values of 5 in group A were recoded to be 4. This was done to mimic what researchers are likely to do when faced with this issue in applied research.

There were two models fitted and compared in all replications: a configural invariance model and a scalar invariance model. The configural model had the factor loadings and thresholds freely estimated for both groups. The latent variable in each group had its estimated mean and variance fixed to be 0 and 1, respectively. Further, the variances of the LRVs were fixed to 1 in both groups. The scalar invariance model had the factor loadings and thresholds constrained to equality across groups. Constraining the measurement parameters across groups allowed the latent variable mean and variance to be estimated in the second group rather than fixed to 0 and 1. Further, in the scalar invariance model the variances of the LRVs were still constrained to be 1 for model identification purposes. The model comparison for the original simulated data, as well as data in each random permutation, was conducted using the lavaan package in R by fitting the configural and scalar invariance models with the "cfa" function. The popular WLSMV estimator was used for all models.

2.1.2 Between Replication Conditions

2.1.2.1 Response Options

Evaluating the effect of response categories on the required number of permutations is important because increasing the number of response categories increases the number of parameters estimated in a model, which also increases computation time. Data were simulated with two or five response options for each of the eight indicator variables. With two response options the configural and scalar models compared in the present research had a chi-square difference value with 14 degrees of freedom, whereas when five response options are used the chi-square difference had 38 degrees of freedom. A population threshold of 0 was used for two response options and for five response options -1.30, -0.47, 0.47, and 1.30 was used. The thresholds used for five response options were based on the recent simulation work of Sass, Schmitt, and Marsh (2014), and the threshold of 0 for the two response option condition was chosen so that the indicator variables would, on

average, have an equal number of responses in both categories.

2.1.2.2 Factor Loading Invariance

Manipulating factor loading invariance allowed changes in Type I error rates and power based on the number of random permutations to be observed. Factor loading invariance/non-invariance was created by subtracting 0 or .25 from the factor loadings for items 1 and 2 (which were chosen arbitrarily) in the population model in the focal group (i.e., the group receiving the non-invariance manipulation). Specifically, conditions in which the factor loadings were invariant had a factor loading of .60 for items 1 and 2 for both groups; whereas conditions in which the factor loadings were non-invariant had a factor loading of .60 for the reference group and .35 for the focal group for items 1 and 2. The difference value of .25 for factor loadings has been used in previous simulation research (Sass et al., 2014). Items 3-8 always had factor loadings of .60 and the residual variances for indicator variables were always set at $1 - \lambda^2$ in the population model.

2.1.3 Within Replication Conditions

2.1.3.1 Number of Random Permutation Shuffles

The random permutation chi-square difference test required the simulated data in each replication to be randomly shuffled to build a distribution of chi-square difference values under the assumption that the null hypothesis is true. This was achieved by randomly shuffling the grouping variable in the simulated data 1,000 times. In each random shuffle of group membership the configural invariance and scalar invariance model was fitted to the data and the chi-square difference value between the two models was saved. After this was done 1,000 times the original chi-square difference value was compared to subgroups of the values from the random shuffles. Specifically, the first permutation test was conducted using only the first 100 random permutations as a comparison distribution, and then the first 200 random permutations were used for another test. Separate tests were done for values between 100 and 1,000, in intervals of 100. A *p*-value for the observed statistic was

calculated by determining what proportion of the chi-square difference values from each subset of random shuffles exceeded the observed chi-square difference. A *p*-value equal to or less than .05 resulted in measurement invariance being rejected. Unequal response categories observed in random permutation shuffles were handled by resampling permutation shuffles.

2.1.4 Outcomes

2.1.4.1 Unequal Response Categories

As was previously mentioned, unequal response categories can occur when conducting the random permutation test. These permutation shuffles were dealt with by discarding them and replacing them with a new shuffle of the grouping variable. This process was repeated until the new shuffle resulted in groups that had the same number of response categories on all indicator variables. The number of occurrences of unequal response categories in both the originally generated data as well as the data in random permutation testing was recorded in simulation one. This information is provided to give the reader an idea of the potential impact of the decision on how to handle unequal response categories.

2.1.4.2 Type I Errors and Power

The outcome of interest in each condition with each testing method was the proportion of the replications in which measurement invariance was rejected. This rejection rate was calculated by dividing the number of replications where measurement invariance was rejected by 1,000. In conditions where measurement invariance was simulated (i.e., all population model factor loadings and thresholds are equal in both groups), the rejection rate was the Type I error rate. The desired Type I error rate was .05., with values between .036 and .064 considered to be within the nominal range. The nominal range was calculated using the expected nominal value (.05) and the number of replications (1,000) to compute the standard error of the Type I error value of .05 (see Equation

2.1), and multiplying that value by ± 1.96 to yield a 95% confidence interval.

$$SE_{TypeI} = \sqrt{\frac{.05(1-.05)}{1000}} = .007 \tag{2.1}$$

In addition to reporting Type I error rates and power, analysis explored how often an increase in the number of permutations used changes the result of the null hypothesis test. For each of the random permutation sample numbers evaluated (except for the full 1,000), the proportion of replications in which the decision about measurement invariance based on the random permutation test changed when an additional 100 random permutations were used was recorded. It was expected that diminishing returns in added permutations would be observed. In other words, the greatest change in outcomes was expected between 100 and 200 permutations, whereas the change in outcomes between 900 and 1,000 was expected to be minor.

2.2 Simulation Two

The second Monte Carlo simulation was designed to answer research questions one, two, three, and four:

- How does chi-square difference testing using random permutations perform in terms of Type I error rate and power? Are Type I errors close to .05? If Type I errors are close to .05, does the test show reasonable power that increases as a function of sample size?
- How do sample size, the number of response categories, and threshold symmetry influence the performance of chi-square random permutation testing?
- How does the random permutation testing approach perform compared to DIFFTEST in M*plus* and other chi-square difference test implementations?
- Does the random permutation test outperform DIFFTEST in M*plus* when groups have different latent variable distributions?

This simulation evaluated the random permutation chi-square difference test against parametric chi-square difference tests for measurement invariance testing. The simulation was designed to evaluate Type I error rates and power of different chi-square difference tests when testing for metric and scalar (i.e., threshold) invariance in tandem, as suggested by Muthén and Muthén (2015). Simulation conditions were varied with the intent of identifying situations where existing testing procedures do not perform well but the random permutation test does. The simulation design was a fully crossed 2 (latent variable distribution) x 2 (response categories) x 2 (threshold symmetry) x 2 (sample size) x 2 (factor loading invariance) design resulting in 32 between replication conditions used to generate data, each having 1,000 replications. In each replication, four different chi-square difference testing approaches were used to test for measurement invariance.

2.2.1 Data Generation and Analysis

Data for two groups were simulated in the R program. A single latent common factor with eight indicator variables was simulated in all conditions. In order to enhance the flexibility in the data generating process, specifically to allow the latent common factor distribution to be non-normal, the data were simulated manually by using random number generators and linear equations. First the latent variable (η) was generated using the desired distribution and population parameters, then linear equations were used to generate each LRV (y^*; see Equations 2.2 & 2.3). When the latent variable was normally distributed it was generated using the "rnorm" function in R, and when the latent variable was non-normal it will be generated using the power method suggested by (Fleishman, 1978).

$$y_i^* = v + \lambda \eta_i + \varepsilon_i \tag{2.2}$$

$$\varepsilon \sim N(0, 1 - \lambda^2) \tag{2.3}$$

Once the LRVs were generated, the population model threshold values (which are dependent on

the desired number of responses categories) were used to divide the LRVs into discrete categories. For example, in conditions with two response categories with a threshold of 0, LRV scores less than 0 had a response of 1 whereas LRV scores equal to or greater than 0 had a response of 2. In order to handle generated data where a variable had unequal observed responses across groups, categories were collapsed to ensure that response options matched in all "observed" data. For example, if generated data showed responses of 1, 2, 3, 4, 5 in group A, but 1, 2, 3, 4 in group B, all values of 5 in group A were recoded to be 4. This was done to mimic what researchers are likely to do when faced with this issue in applied research.

There were two models fitted and compared in all replications: a configural invariance model and a scalar invariance model. The configural model had the factor loadings and thresholds freely estimated for both groups, whereas the latent variable in each group had its estimated mean and variance fixed to be 0 and 1, respectively. Further, in the configural model the variances of the LRVs (i.e., scales) were fixed to 1 in both groups. The scalar invariance model had the factor loadings and thresholds constrained to equality across groups. Constraining the measurement parameters across groups allowed the latent variable mean and variance to be estimated in the second group rather than fixed to 0 and 1. Further, in the scalar invariance model the variances of the LRVs were still constrained to be 1 for model identification purposes.

2.2.2 Between Replication Conditions

2.2.2.1 Latent Variable Distribution

Data were simulated for both normally distributed and positively skewed latent variables. This is an important manipulation because ordered-categorical CFA estimation methods assume that latent variables are normally distributed. In the normally distributed latent variable condition, the latent variable had kurtosis and skewness values equal to 0 for both groups. In the non-normally distributed condition, the latent variable had a skewness of 1.5 and (excess) kurtosis of 3.5 in the focal group, whereas the reference group had a normally distributed latent variable. These distribution shapes were chosen to replicate the simulation work of Suh (2015) which showed that

the M*plus* DIFFTEST command has inflated Type I error rates, when testing in the IRT framework, when there are large group differences in latent variable distributions.

2.2.2.2 Response Categories

Data were simulated with two or five response options for all of the eight indicator variables. These response option values were chosen to mimic response scales that are likely to be used in applied research. Data with two response options represent response scales where participants are asked to respond "No" or "Yes" to indicate if a statement is true. Data with five response options represent response scales where participants respond on a scale such as "Strongly Disagree", "Disagree", "Neutral", "Agree", or "Strongly Agree" to indicate their level of agreement with a statement.

2.2.2.3 Threshold Symmetry

Data were simulated using population thresholds values that create symmetric or asymmetric indicator variable distributions (based on quantiles from a normally distributed latent variable). Generating data using asymmetric thresholds was done in order increase generalizability to applied research where response distributions are often not symmetric (e.g., Serious Harm Reduction Scale, Martens et al. 2007). Population threshold values of 0 or 0.70 were used when generating dichotomous data. A threshold of 0 creates data where the two response categories are equally likely, whereas a threshold 0.70 creates data where approximately 76% of responses are in the lower response category when both the latent variable and the LRV are normally distributed. The threshold value of 0 is commonly used in simulation research when generating dichotomous data (e.g., Beauducel & Herzberg 2006; Rhemtulla et al. 2012), and the threshold value of 0.70 was chosen as a compromise between what Rhemtulla et al. defined as moderate asymmetry (threshold of 0.36) and extreme asymmetry (threshold of 1.04) levels.

For five response options threshold values of -1.30, -0.47, 0.47, and 1.30 or -0.25, 0.38, 0.84, and 1.28 were used. Values of -1.30, -0.47, 0.47, and 1.30 were used to generate data with a symmetric distribution, whereas -0.25, 0.38, 0.84, and 1.28 generate data that show a large number

cases (approximately 40%) endorsing the lowest possible response when both the latent variable and the LRV are normally distributed. The thresholds for five response options are based on the recent simulation work of Sass, Schmitt, and Marsh (2014) who chose these values in an attempt to create conditions that generalize to applied research. Importantly, the asymmetric conditions should generalize to situations where responses are also grouped towards the upper end of a response scale when the latent variable is normally distributed in both groups. This is because the two groups have no population mean difference on the latent variable therefore the direction of the skew is meaningless if it is the same across groups.

2.2.2.4 Sample Size

Two total sample sizes were used in the simulation, 300 and 600. These values were chosen following typical simulation research conventions when setting sample size values. The values selected fall within the range of those used in previous simulation research (Elosua, 2011; Elosua & Wells, 2013;Lubke & Muthén, 2004; Sass et al., 2014). Both simulated groups had an equal number of observations in all replications.

2.2.2.5 Factor Loading Invariance

Factor loading invariance/non-invariance was created by subtracting 0 or .25 from the factor loadings for items 1 and 2 (which were chosen arbitrarily) in the population model in the focal group (i.e., the group receiving the non-invariance manipulation). Specifically, conditions in which the factor loadings were invariant had a factor loading of .60 for items 1 and 2 for both groups; whereas conditions in which the factor loadings were non-invariant had a factor loading of .60 for the reference group and .35 for the focal group for items 1 and 2. The difference value of .25 for factor loadings has been used in previous simulation research (Sass et al., 2014). Items 3-8 always had factor loadings of .60 and the residual variances for indicator variables were always set at $1 - \lambda^2$ in the population model.

2.2.3 Within Replication Conditions

There were four chi-square difference testing methods compared in each replication; all were used to analyze the same data. For each test a configural invariance model (i.e., freely estimated measurement parameters in both groups) was compared to a scalar invariance model (i.e., factor loadings and items thresholds constrained to be equal in both groups) in order to test for measurement invariance. The models fitted in all four testing methods were identical, the only difference was how the chi-square difference test was conducted.

2.2.3.1 Chi-Square Difference Random Permutation Testing

The permutation chi-square test required the simulated data in each replication to be randomly shuffled to build a distribution of chi-square difference values under the assumption that the null hypothesis is true. In each random shuffle of group membership the configural invariance model and scalar invariance model were fitted to the data and the chi-square difference value between the two models was saved. After this was done the appropriate amount of times (which was determined by simulation two) the original chi-square difference value was compared to the values from the random shuffles. A *p*-value for the observed statistic was calculated by determining what proportion of the chi-square difference values from the random shuffles exceeded the observed chi-square difference. A *p*-value equal to or less than .05 resulted in measurement invariance being rejected. The permutation testing was conducted using the lavaan package in R. Random permutation shuffles in which at least one variable had a different number of observed responses across groups was handled by resampling random permutation shuffles.

2.2.3.2 DIFFTEST Chi-Square Difference in M*plus*

The DIFFTEST method was also used in all evaluations of measurement invariance. Following the instructions of Muthén and Muthén (2015), comparisons between the configural invariance model and scalar invariance model were conducted using the DIFFTEST option in M*plus*. Measurement invariance was rejected in replications that showed a chi-square DIFFTEST *p*-value less than or

equal to .05.

2.2.3.3 Satorra-Bentler Scaled Chi-Square Difference in lavaan

Chi-square difference testing with the SB chi-square difference correction presented by Satorra (1999) was also compared to the random permutation test. The SB correction is easily implemented in R by using the "anova" function, which calls the "lavTestLRT" function in lavaan, to compare the fit of two nested CFA models. Hereafter this correction is simply referred to as the "lavTestLRT" correction. When the WLSMV estimator is used, the chi-square difference statistic is automatically scaled with the SB correction. The correction was used with all of the defaults in the lavTestLRT function in version 0.5-22 of the lavaan package. Measurement invariance was rejected in replications that showed a lavTestLRT chi-square difference *p*-value less than or equal to .05.

2.2.3.4 Unadjusted Chi-square Difference Testing

The final method of chi-square difference testing was an unadjusted test. This testing approach served as a baseline to which the other three approaches can be compared. The unadjusted chi-square difference test was conducted by saving the chi-square value and degrees of freedom from the configural invariance model and the scalar invariance model and then using the difference in those values to calculate a chi-square difference *p*-value. As was the case with the other three testing methods, replications with a *p*-value less than or equal to .05 had measurement invariance rejected. This approach used the lavaan package in all conditions.

2.2.4 Outcomes

The outcome of interest in each condition with each testing method was the proportion of the replications in which measurement invariance was rejected. This rejection rate was calculated by dividing the number of replications where measurement invariance was rejected by 1,000. In conditions where measurement invariance was simulated (i.e., all population model factor loadings

and thresholds are equal in both groups), the rejection rate was the Type I error rate. Testing methods that showed Type I errors within or below the nominal range of .036 to .064 described in the description for simulation one were considered to have acceptable Type I error control. In conditions in which measurement non-invariance was simulated (i.e., measurement parameters differ on items 1 and 2 in the population model), the rejection rate was power. Methods that showed acceptable Type I error rate control were compared on power. Specifically, determining if the random permutation test controls Type I errors and yields power equal to or greater than other testing approaches was the main focus when interpreting the results.

2.3 Simulation Implementation

In order to ensure that the simulations were conducted in a timely manner, the University of Kansas Advanced Computing Facility cluster was used to run multiple between replication conditions simultaneously. Consistent replication of the simulations was ensured by using the portableParallelSeeds (Johnson, 2015) package in R. In cases where a simulation condition would take a long time (e.g., over 30 days) to cycle through all replications and permutation shuffles, the individual replications for that condition were split across computing nodes (e.g., 1,000 replications split into groups of 100 across 10 computing nodes). Using portableParallelSeeds allowed the data generation to be consistent across repeated replications of the simulation, regardless of which computers were used and how the work was divided. Further, it allowed individual replications that triggered error messages to be revisited without requiring the entire simulation to be replicated from the beginning. For example, if a simulation replication towards the end of the simulation triggered an error message and caused the simulation execution to stop, portableParallelSeeds allows the user to look-up the random number seed for that replication and then use it to reproduce the error message in an interactive session for debugging.

Chapter 3

Results

3.1 Simulation One

The focus of simulation one was determining the appropriate number of random permutation shuffles for the second and final Monte Carlo simulation. Possible values between 100 and 1,000, in increments of 100, were evaluated. The results reported here lead to 500 being identified as the lowest permissible number of random permutation shuffles for use in simulation two.

3.1.1 Unequal Response Categories

Across the four between run simulation conditions the occurrence of unequal response categories was extremely rare. The initial data generated in each replication never required response categories to be collapsed in order to make response categories match across groups. In all 1,000 replications in each condition all response categories were initially observed for all variables in both groups, therefore no collapsing of categories in the originally generated data was necessary. Across all replications, resampling permutation shuffles was necessary on three occasions with five response options and equal factor loadings in the population model, whereas resampling shuffles was necessary seven times with five response options and population model differences in factor loadings (see Table 3.1).

Table 3.1: Number of Resamples and Collapses

Response Options	Loading Invariance	Number of Resamples	Number of Collapses
2	Yes	0	0
2	No	0	0
5	Yes	3	0
5	No	7	0

3.1.2 Type I Errors

The Type I error rate change across different random permutation shuffle numbers showed rates that were consistently within or close to the nominal range of .036 to .064 (see Figure 3.1). Interestingly, the Type I error rate showed a slight decrease as the number of random permutations used increased. The decrease in Type I error rates appeared to stop after 500 random permutation shuffles. As expected, the proportion of replications which showed a change in the decision about rejecting the null hypothesis of equal factor loadings and thresholds decreased as the initial number of random permutations increased (see Table 3.2). When increasing the number of random permutation shuffles from 500 to 600 the rejection decision with two response options did not change across all 1,000 replications, whereas with five response options the rejection decision only changed in 3 of the 1,000 replications.

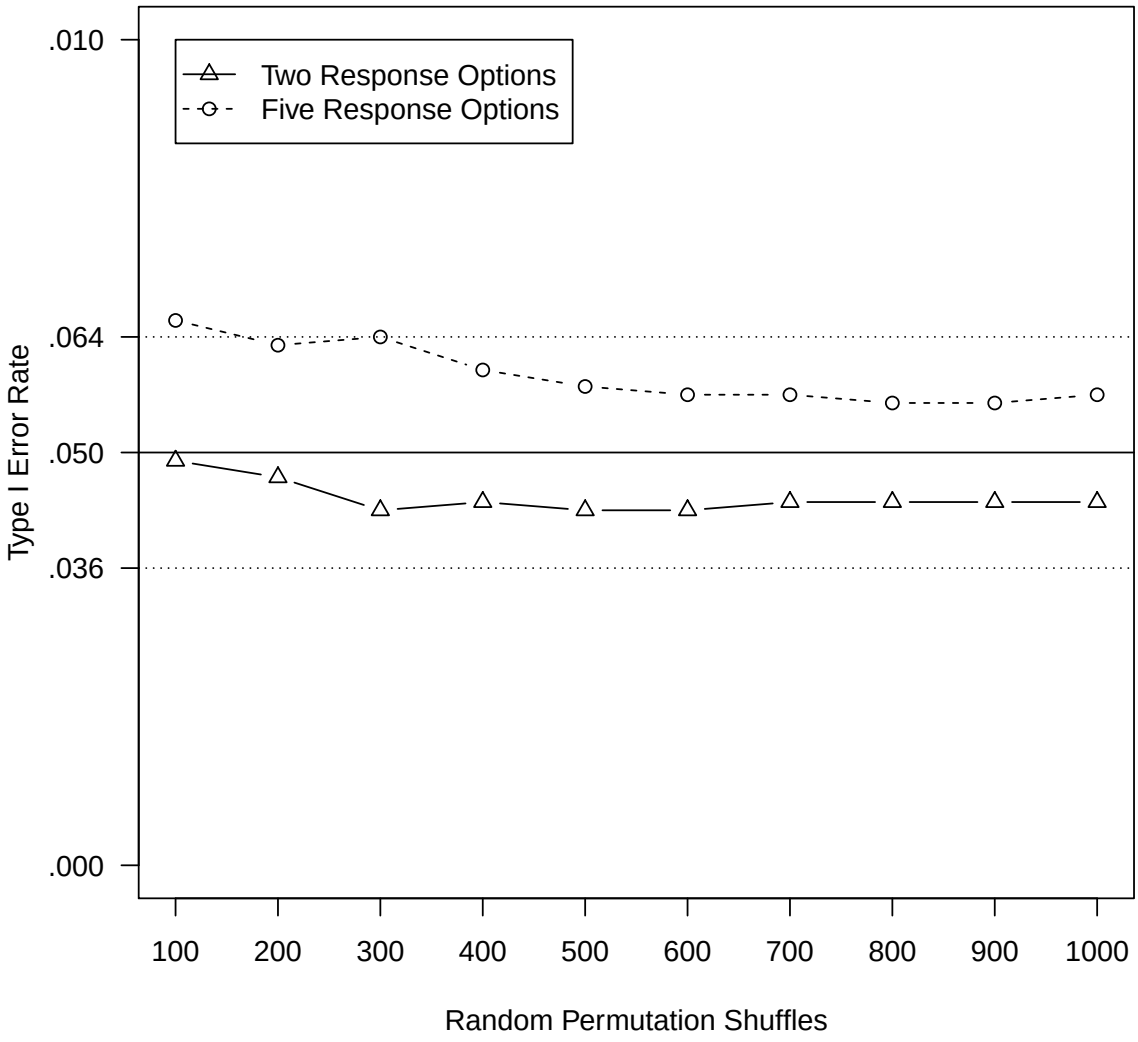

Figure 3.1: Type I Error Rates Across Replications

Table 3.2: Proportion of Replications with a Change in Decision with Population Invariance

Increase in Random Permutations	Changes with Two Responses	Changes with Five Responses
100 to 200	.010	.007
200 to 300	.006	.005
300 to 400	.001	.006
400 to 500	.001	.006
500 to 600	.000	.003
600 to 700	.001	.000
700 to 800	.000	.001
800 to 900	.000	.000
900 to 1000	.000	.001

3.1.3 Power

As was the case with Type I errors, power was not largely influenced by an increase in the number of random permutation shuffles. With both two and five response categories the highest power was observed with 100 random permutation shuffles, with a slight decrease as the shuffles increased (see Figure 3.2). When increasing from 100 to 200 random permutation shuffles, about 3% of the 1,000 replications showed a change in the decision to reject the null hypothesis of equal factor loadings and thresholds. Further, when increasing from 500 to 600 random permutation shuffles a change in the rejection decision was only observed about 1% of the time (see Table 3.3). As the number of random permutation shuffles approached 1,000 the proportion of replications which showed a rejection decision change nearly became zero.

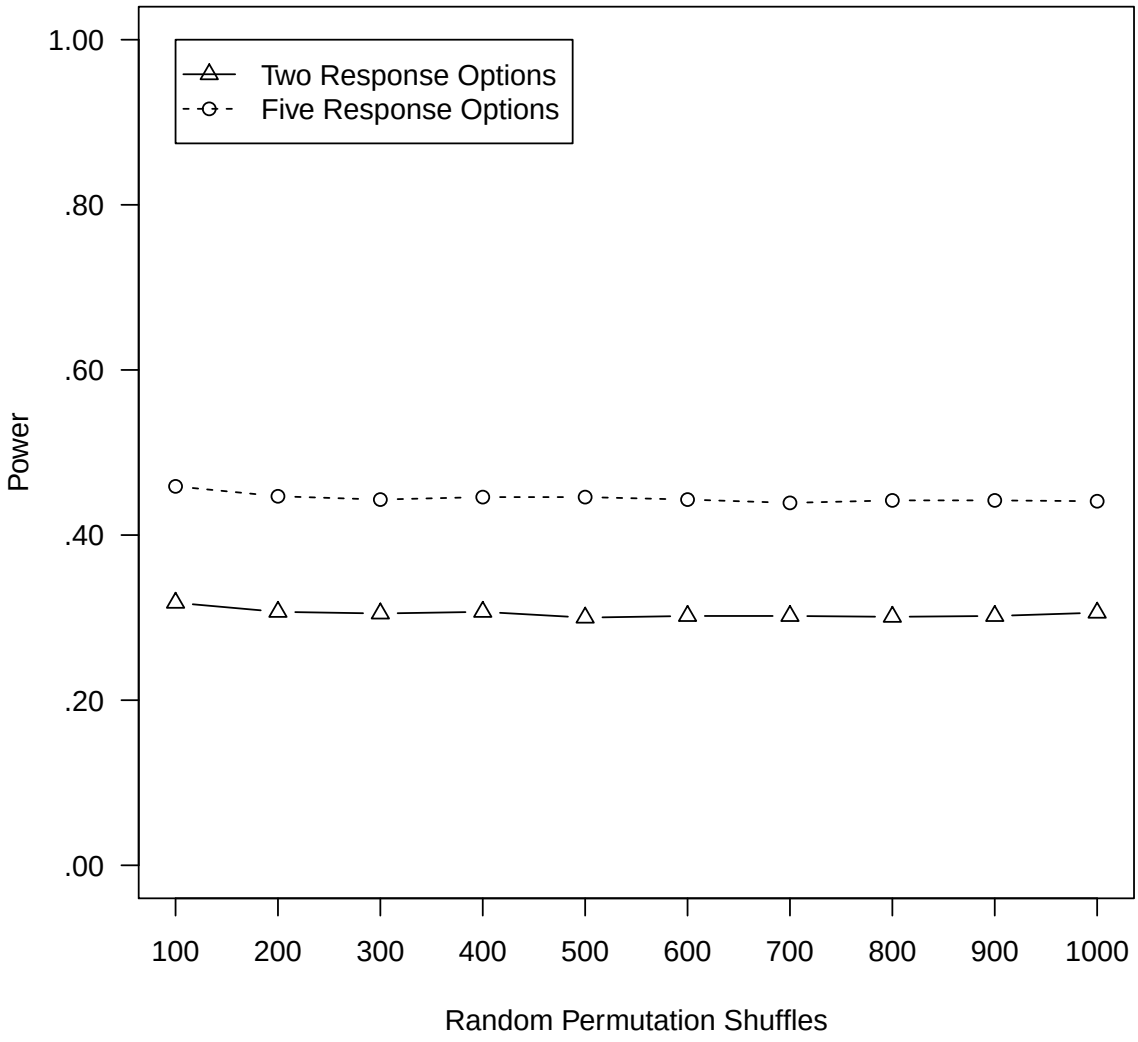

Figure 3.2: Power Across Replications

Table 3.3: Proportion of Replications with a Change in Decision with Non-Invariance

Increase in Random Permutations	Changes with Two Responses	Changes with Five Responses
100 to 200	.033	.028
200 to 300	.018	.016
300 to 400	.014	.015
400 to 500	.013	.008
500 to 600	.010	.009
600 to 700	.008	.008
700 to 800	.011	.005
800 to 900	.005	.004
900 to 1000	.004	.003

3.1.4 Selected Number of Shuffles

Because the focus of the present research is on how random permutation testing controls Type I errors, it is important to use the appropriate number of shuffles to maximize its error control in simulation two. The results of simulation one show that Type I errors decrease as the number of permutation shuffles increases, but that after 500 random permutation shuffles using additional shuffles has little to no impact on results. In order to ensure that random permutation testing could control Type I errors well, 500 permutation shuffles were used in all conditions in simulation two.

3.2 Simulation Two

Based on the results of simulation one, this simulation was conducted with 500 random permutation shuffles. The results are reported first for conditions where the latent variable was normally distributed in each group, and then conditions with a non-normal latent variable in the second group are shown. The results are broken down to discuss Type I error rate control, and then discuss power.

3.2.1 Normally Distributed Latent Variables

3.2.1.1 Type I Errors

Type I errors for the random permutation test were within the nominal range of .036-.064 when the latent variable was normally distributed in both groups (see Table 3.4). Interestingly, all Type I error rates where within one standard error of the nominal value of .05.

Table 3.4: Type I Error Rates with Normal Latent Variables

Sample Size	Response Categories	Symmetric Thresholds	Permutation Type I Error Rate
150	2	Yes	.050
300	2	Yes	.043
150	5	Yes	.053
300	5	Yes	.053
150	2	No	.053
300	2	No	.056
150	5	No	.050
300	5	No	.054

3.2.1.2 Power

Evaluations of power were then conducted in conditions where the latent variables were normally distributed in both groups and factor loading differences were present in the population model (see Table 3.5).

Table 3.5: Power with Normal Latent Variables

Sample Size	Response Categories	Symmetric Thresholds	Permutation Power
150	2	Yes	.279
300	2	Yes	.543
150	5	Yes	.460
300	5	Yes	.786
150	2	No	.214
300	2	No	.406
150	5	No	.342
300	5	No	.707

The results showed that power increased with larger group sizes. Further, power was greater with five response categories per indicator when compared to two response categories. Addition-

ally, power was noticeably lower with non-symmetric indicator variable thresholds. These effects are depicted in Figure 3.3.

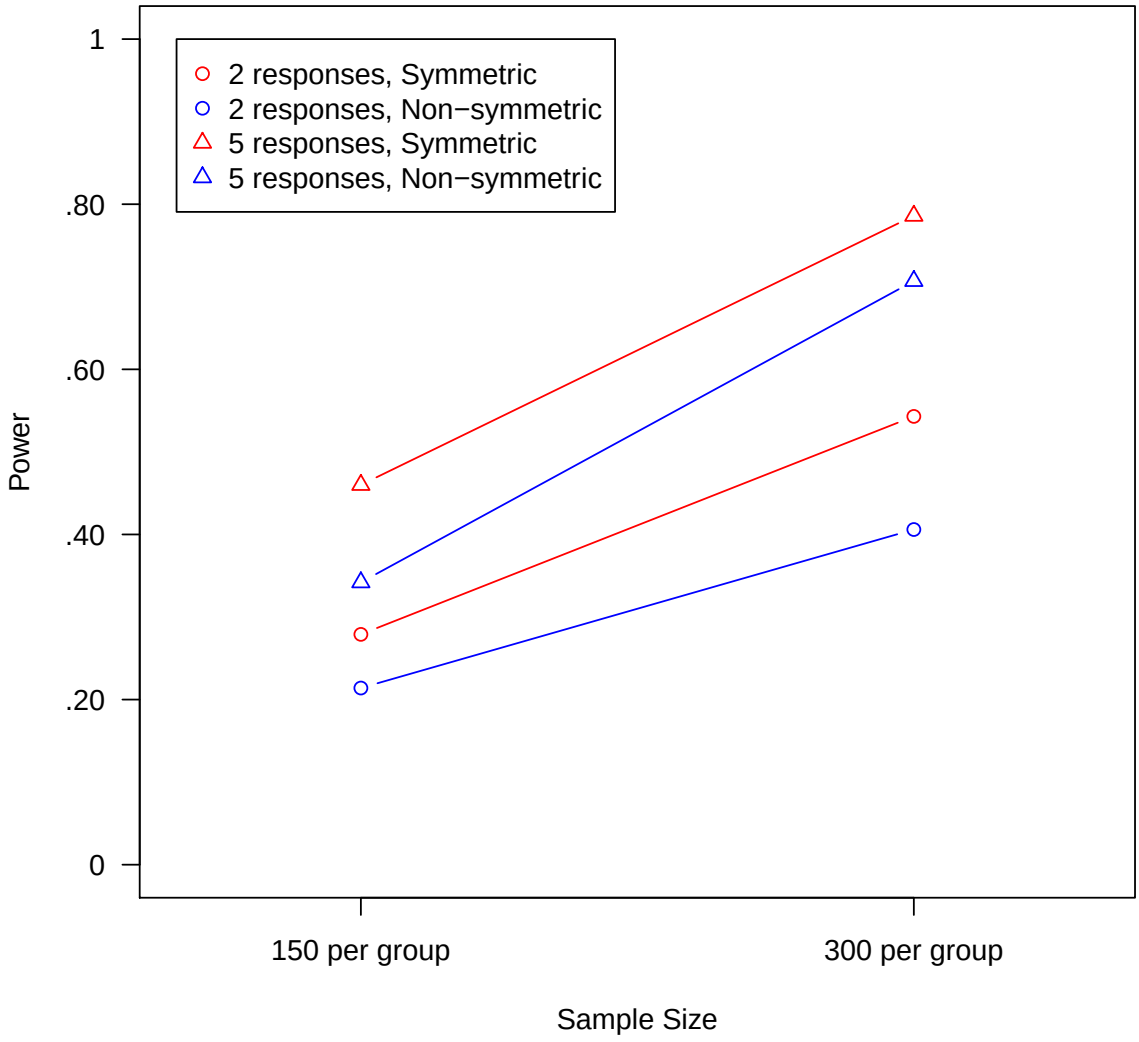

Figure 3.3: Power with Normal Latent Variables

3.2.1.3 Comparison With Existing Testing Procedures

After random permutation testing was shown to perform well with normally distributed latent variables, its performance was compared to existing chi-square difference testing procedures (see

Table 3.6). The M*plus* DIFFTEST correction showed Type I error rates inside of the nominal range in six of the eight conditions with equal population factor loadings and normally distributed latent variables in both groups. The two conditions in which errors were above .064 had two response options per variable and non-symmetric thresholds. The lavTestLRT correction showed error rates that were well below the nominal range and close or equal to 0. As expected, the standard chi-square difference test with no adjustment for the chi-square difference statistic showed error rates that were consistently inflated.

Table 3.6: Type I Errors Across Testing Methods

Condition Info	Random Permutations	DIFFTEST	lavTestLRT	No Adjustment
150, 2, Symmetric	.050	.060	.004	.143
300, 2, Symmetric	.043	.052	.002	.128
150, 5, Symmetric	.053	.062	.000	.131
300, 5, Symmetric	.053	.057	.000	.098
150, 2, Non-Symmetric	.053	.065	.010	.135
300, 2, Non-Symmetric	.056	.078	.004	.139
150, 5, Non-Symmetric	.050	.053	.001	.131
300, 5, Non-Symmetric	.054	.062	.000	.128

Note. The Condition Info column indicates each conditions' group sample size, number of response categories, and threshold symmetry.

Evaluations of power between the random permutation test, M*plus* DIFFTEST, and the lavTestLRT correction showed that the performances of random permutation testing and M*plus* DIFFTEST were remarkably similar with DIFFTEST always being slightly more powerful (see Table 3.7).

Table 3.7: Power Across Testing Methods

Condition Info	Random Permutations	DIFFTEST	lavTestLRT	No Adjustment
150, 2, Symmetric	.279	.319	.084	.452
300, 2, Symmetric	.543	.568	.220	.703
150, 5, Symmetric	.460	.504	.091	.618
300, 5, Symmetric	.786	.811	.347	.890
150, 2, Non-Symmetric	.214	.258	.070	.361
300, 2, Non-Symmetric	.406	.457	.143	.588
150, 5, Non-Symmetric	.342	.370	.049	.519
300, 5, Non-Symmetric	.707	.733	.226	.831

Note. The Condition Info column indicates each conditions' group sample size, number of response categories, and threshold symmetry.

Simultaneous comparisons between Type I error rates and power are displayed in Figure 3.4. The plots in Figure 3.4 show the Type I error rates for each combination of simulation variables on the x-axis, and power on the y-axis. The vertical dotted lines indicate the nominal range for the Type I error rate (.036-.046). These plots show how the different testing approaches simultaneously controlled Type I errors while maximizing power. Plot points that are red indicate testing conditions where Type I errors were above the nominal range, whereas blue points indicate Type I errors below the nominal range. The desired outcome was a many values within the nominal range as possible with high power.

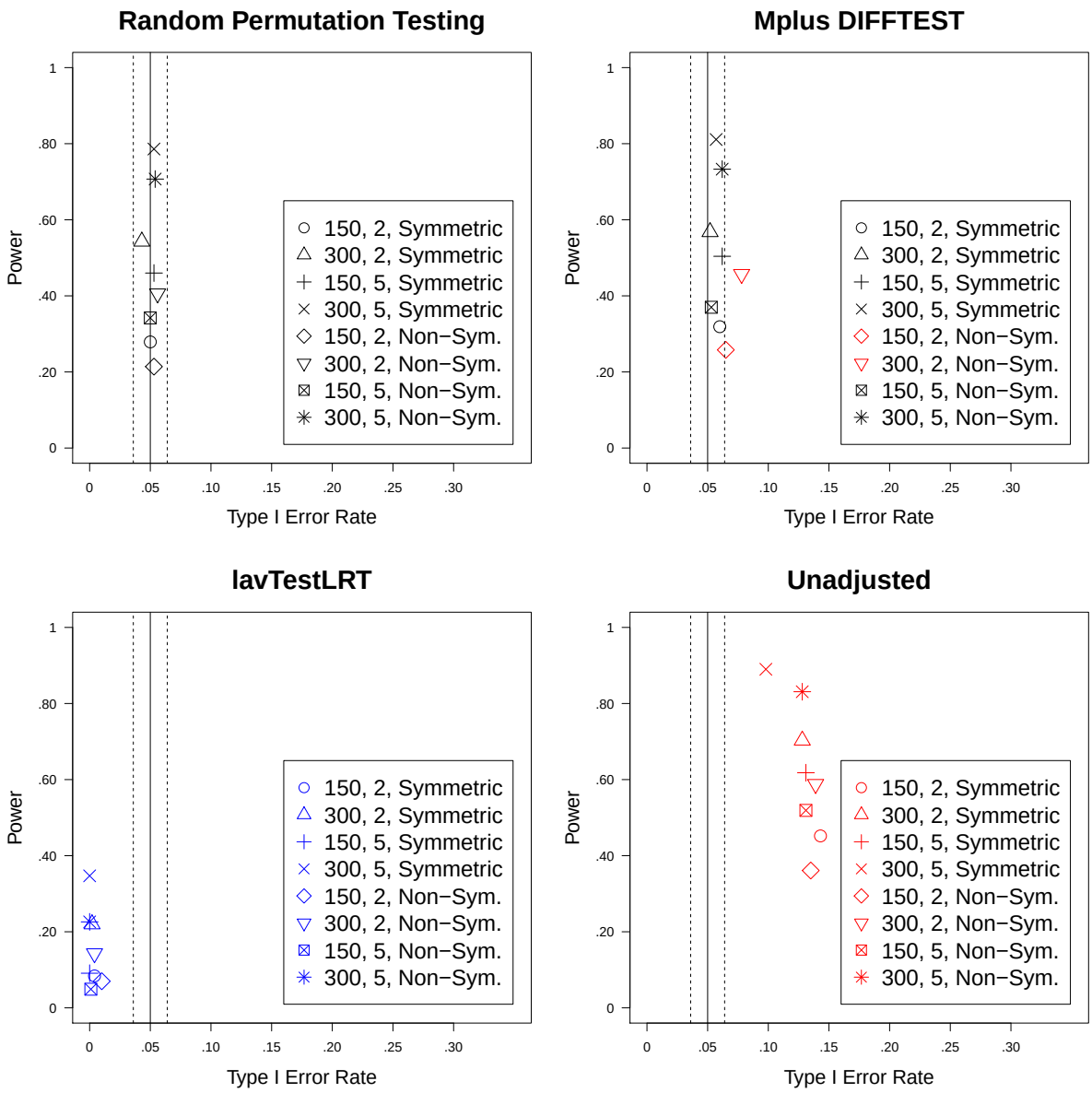

Figure 3.4: Type I Errors and Power Across Conditions and Testing Methods

The plots in Figure 3.4 clearly show that the two best performers were random permutation testing and M*plus* DIFFTEST. The differences in rejection rates between M*plus* DIFFTEST and random permutation testing across all 16 conditions with normally distributed latent variables were further explored using McNemar's test. Specifically, the differences in rejection decisions were evaluated for statistical significance. The results are shown in Table 3.8.

Table 3.8: Comparison of M*plus* DIFFTEST and Random Permutations

Sample Size	Invariance	Response Categories	Symmetric Thresholds	Random Permutations	Mplus DIFFTEST	McNemar's Chi-square
150	Yes	2	Yes	.050	.060	8.1*
300	Yes	2	Yes	.043	.052	7.111*
150	Yes	5	Yes	.053	.062	3.368
300	Yes	5	Yes	.053	.057	1.125
150	Yes	2	No	.053	.065	10.083*
300	Yes	2	No	.056	.078	20.045*
150	Yes	5	No	.050	.053	0.364
300	Yes	5	No	.054	.062	6.125*
150	No	2	Yes	.279	.319	31.688*
300	No	2	Yes	.543	.568	16.457*
150	No	5	Yes	.460	.504	35.558*
300	No	5	Yes	.786	.811	19.862*
150	No	2	No	.214	.258	42.023*
300	No	2	No	.406	.457	45.455*
150	No	5	No	.342	.370	15.848*
300	No	5	No	.707	.733	18.382*

Note. * indicates a chi-square value above the critical value of 3.84

The tests for statistical significance showed that the performances of the M*plus* DIFFTEST procedure and random permutation testing rejections rates were (statistically) significantly different in 13 of the 16 conditions. Five of the conditions with a statistically significant difference in rejection rate showed M*plus* DIFFTEST having a higher Type I error rate. And the remaining eight conditions showed M*plus* DIFFTEST having significantly higher power. In conclusion, the results across conditions with a normally distributed latent variable in both groups showed that the only methods which consistently controlled Type I errors to be within the nominal range was the random permutation test. The M*plus* DIFFTEST procedure showed Type I errors close to or within the nominal range, and had the benefit of more power to detect non-invariance.

3.2.2 Non-Normal Latent Variable Conditions

There were a total of 16 conditions in which the focal group had a latent variable with a mean of 0, variance of 1, skewness of 1.5, and (excess) kurtosis of 3.5. The reference group had a normally

distributed latent variable with a mean of 0 and variance of 1. The random permutation test and the competing chi-square difference tests were evaluated for Type I error rates and power in these 16 conditions.

3.2.2.1 Comparisons With Existing Testing Procedures

The random permutation test showed a Type I error rate within the nominal range in six of the eight possible conditions. The M*plus* DIFFTEST correction showed Type I error rates within the nominal range in three of the eight conditions. As was observed with a normally distributed latent variable in both groups, lavTestLRT had error rates close to 0 and the unadjusted test had consistently inflated error rates (see Table 3.9).

Table 3.9: Type I Error Rates with Non-Normal Latent Variable

Condition Info	Random Permutations	DIFFTEST	lavTestLRT	No Adjustment
150, 2, Symmetric	.053	.066	.013	.113
300, 2, Symmetric	.055	.061	.007	.150
150, 5, Symmetric	.068	.081	.001	.145
300, 5, Symmetric	.080	.100	.002	.192
150, 2, Non-Symmetric	.052	.075	.002	.136
300, 2, Non-Symmetric	.047	.063	.000	.126
150, 5, Non-Symmetric	.045	.054	.001	.111
300, 5, Non-Symmetric	.057	.068	.000	.152

Note. The Condition Info column indicates each conditions' group sample size, number of response categories, and threshold symmetry.

The Type I Error control for the M*plus* DIFFTEST correction and random permutation testing were reasonable enough to warrant an evaluation of power (see Table 3.10).

Table 3.10: Power with Non-Normal Latent Variable

Condition Info	Random Permutations	DIFFTEST	lavTestLRT	No Adjustment
150, 2, Symmetric	.227	.262	.055	.385
300, 2, Symmetric	.475	.511	.191	.636
150, 5, Symmetric	.397	.446	.081	.558
300, 5, Symmetric	.763	.795	.326	.878
150, 2, Non-Symmetric	.254	.302	.058	.429
300, 2, Non-Symmetric	.522	.548	.184	.666
150, 5, Non-Symmetric	.375	.422	.059	.542
300, 5, Non-Symmetric	.736	.764	.208	.870

Note. The Condition Info column indicates each conditions' group sample size, number of response categories, and threshold symmetry.

Simultaneous comparisons between Type I error rates and power in conditions with different latent variable distributions are displayed in Figure 3.5. The plots in Figure 3.5 show the Type I error rates for each combination of simulation variables on the x-axis, and power on the y-axis. The vertical dotted lines indicate the nominal range for the Type I error rate (.036-.064). These plots show how the different testing approaches simultaneously controlled Type I errors while maximizing power. Plot points that are red indicate testing conditions where Type I errors were above the nominal range, whereas blue points indicate Type I errors below the nominal range. The desired outcome was a many values within the nominal range as possible with high power.

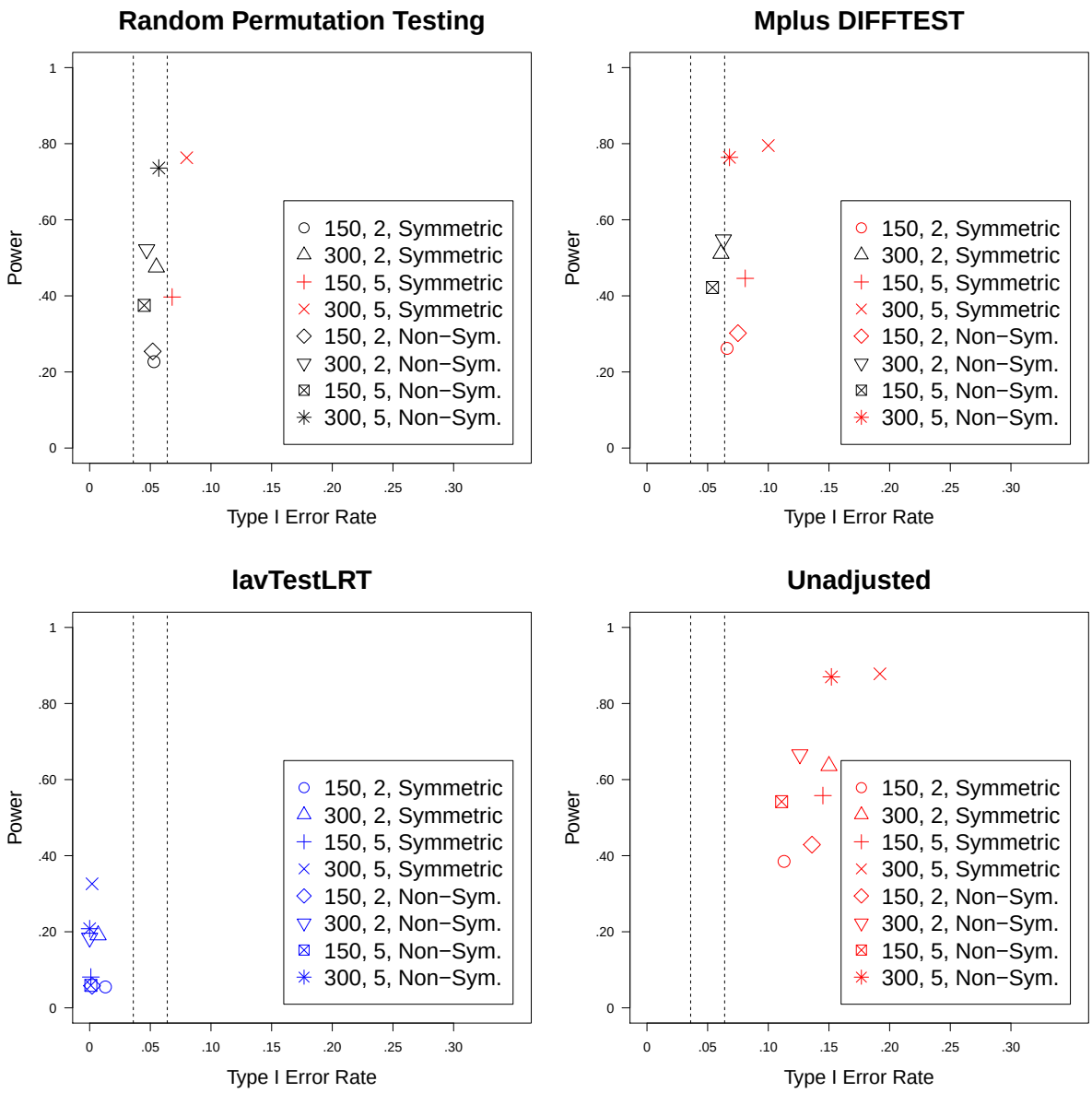

Figure 3.5: Type I Errors and Power Across Conditions with Non-Normal LV

The plots in Figure 3.5 show that although none of the testing approaches showed consistently acceptable Type I error control, the M*plus* DIFFTEST procedure and random permutation testing were by far the best performers in conditions with differences in latent variable distributions. The differences in rejection rates across all 16 conditions between M*plus* DIFFTEST and random permutation testing were further explored using McNemar's test. Specifically, the differences in rejection decisions were evaluated for statistical significance. The results are shown in Table 3.11.

Table 3.11: Comparison of M*plus* DIFFTEST and Random Permutations

Sample Size	Invariance	Response Categories	Symmetric Thresholds	Random Permutations	Mplus DIFFTEST	McNemar's Chi-square
150	Yes	2	Yes	.053	.066	11.077*
300	Yes	2	Yes	.055	.061	2.083
150	Yes	5	Yes	.068	.081	7.579*
300	Yes	5	Yes	.080	.100	16.409*
150	Yes	2	No	.052	.075	21.043*
300	Yes	2	No	.047	.063	14.062*
150	Yes	5	No	.045	.054	4.923*
300	Yes	5	No	.057	.068	6.667*
150	No	2	Yes	.227	.262	29.641*
300	No	2	Yes	.475	.511	30.625*
150	No	5	Yes	.397	.446	43.472*
300	No	5	Yes	.763	.795	28.265*
150	No	2	No	.254	.302	42.481*
300	No	2	No	.522	.548	16.447*
150	No	5	No	.375	.422	41.49*
300	No	5	No	.736	.764	21.441*

The results from McNemar's test showed a significant difference in rejection rate in 15 out of 16 conditions. Seven out of eight possible conditions showed Type I error rates significantly lower (and closer to .05) for the random permutation test. The eight remaining conditions were instances where the M*plus* DIFFTEST procedure showed greater power than random permutation testing.

Chapter 4

Empirical Example

4.1 Data

In order to demonstrate the use of the random permutation test with empirical data and compare it with the M*plus* DIFFTEST procedure and lavTestLRT, publicly available data were gathered for analysis. Data from the 2012 Programme for International Student Assessment (PISA; OECD, 2013) were used in an evaluation of measurement invariance.

The data contained responses from 15-year-old students from 65 different countries. The total number of children who participated in the 2012 PISA survey was over 510,000. A set of eight questions assessed students' Math Self Efficacy (i.e., their belief and confidence in their own math ability); these items were determined to be appropriate for a latent variable model. The questions assessed how confident students would be doing different types of math problems, responses were recorded on a four-point ordered scale with responses of 1 = "Very confident", 2 = "Confident", 3 = "Not very confident", and 4 = "Not at all confident". The items were the following:

1. Using a train timetable to work out how long it would take to get from one place to another.

2. Calculating how much cheaper a TV would be after a 30% discount.

3. Calculating how many square metres of tiles you need to cover a floor.

4. Understanding graphs present in newspapers.

5. Solving an equation like 3x+5= 17.

6. Finding the actual distance between two places on a map with a 1:10,000 scale.

7. Solving an equation like 2(x+3) = (x+3)(x-3).

8. Calculating the petrol consumption rate of a car.

In this example, scores on these indicators were reverse-coded so that higher scores indicate higher levels of confidence. This allows the latent variable to be interpreted as Math Self Efficacy where higher scores indicate a higher level of belief in one's math ability. These items appeared to be theoretically appropriate indicators of a single Math Self Efficacy latent variable, therefore all CFA models fit had eight indicators and a single latent variable.

Rather than conducting a test for measurement invariance with all countries present, only cases from students from the United States of America (USA) and Canada were used. This resulted in a data frame with 21,544 cases from the United States and 4,978 from Canada. In order to reduce the number of cases per group to better represent group sizes typically seen in applied psychology research, 300 students were randomly sampled from each country. The resulting data frame with 600 observations was used to test for measurement invariance of the eight Math Self Efficacy items with the random permutation test procedure. The response frequencies for each indicator variable can be seen in Table 4.1.

Table 4.1: Response Frequencies for USA and Canada

	Canada, USA			
Indicator	"Not at all confident"	"Not very confident"	"Confident"	"Very confident"
1	15, 7	59, 55	126, 148	100, 90
2	18, 8	56, 62	101, 131	125, 99
3	12, 8	57, 78	120, 120	111, 94
4	10, 5	34, 38	141, 138	115, 119
5	7, 2	21, 10	76, 87	196, 201
6	32, 26	94, 108	101, 104	73, 62
7	19, 11	48, 38	107, 102	126, 149
8	26, 12	96, 86	120, 126	58, 76

4.2 Analysis and Results

4.2.1 Initial Test for Measurement Invariance

The first model fitted to the analysis data was a configural invariance model. This model specified the same measurement structure for both countries, but allowed the factor loadings and thresholds in each to be freely estimated. In order to set the scale, the latent variable mean and variance were fixed to 0 and 1, respectively, for both groups. This model fit the data well, although the RMSEA was well above commonly accepted cutoff values, $\chi^2(40) = 245.296$, $CFI = .946$, $TLI = .924$, $RMSEA = .131$. The fit of the configural invariance model was sufficient to warrant further analysis.

The next model fitted to the data was the scalar invariance model. This model had all factor loadings and thresholds constrained to equality across both countries. Further, the constraint of equal measurement parameters allowed the latent variable mean and variance for the Canada group to be freely estimated while the constraints of 0 and 1 remained for the USA group. This model also fit the data well, $\chi^2(70) = 289.807$, $CFI = .942$, $TLI = .954$, $RMSEA = .102$. The difference in the chi-square fit between the two models was 44.511, which was the test statistic for the random permutation test.

The random permutation test was carried out by randomly shuffling the grouping variable 1,000 times. In each shuffle the two aforementioned models were fitted, and their chi-square differences were saved. This built the reference distribution for the chi-square difference statistic. There were a total of five random permutation testing shuffles in which there were unequal response categories across groups. These instances were dealt with by continuing to shuffle the grouping variable until a suitable group membership assignment was observed, therefore a total of 1,005 random permutation shuffles were used. The proportion of cases in the reference distribution greater than the observed test statistic of 44.511 was .022, therefore the null hypothesis of measurement invariance was rejected.

4.2.2 Follow-up Tests

After the initial test found that measurement invariance was not present across the USA and Canada for the Math Self Efficacy indicator variables, a series of additional tests were conducted to determine which item(s) is(are) non-invariant. Eight separate tests of measurement invariance were carried out, each one evaluating a single indicator variable. The same configural model as used for the initial test of measurement invariance was compared to a model where all but one of the indicator variables had its factor loadings and thresholds constrained to equality (i.e., a partial scalar invariance model). Each of the sets of eight models were compared with random permutation testing. The results showed that in six of the eight tests the difference between the configural and partial scalar invariance model was statistically significant, the two exceptions were the difference between the configural and partial scalar models with freely estimated measurement parameters for indicator variables two ("Calculating how much cheaper a TV would be after a 30% discount.") and three ("Calculating how many square metres of tiles you need to cover a floor."). In other words, the model fit did not become significantly worse when all indicators except number two or three had their measurement parameters constrained to equality. These results suggest that both variables are non-invariant.

For the model testing indicator variable two, the observed chi-square difference statistic was 27.846, and its p-value with 1,000 permutation shuffles was .055. This finding suggests that measurement of Math Efficacy with the "Calculating how much cheaper a TV would be after a 30% discount." question is different between students in the USA and students in Canada. Further, for the model testing indicator variable three the observed chi-square difference statistic was 27.846, and its p-value with 1,000 permutation shuffles was .092. This finding suggests that measurement of Math Efficacy with the "Calculating how many square metres of tiles you need to cover a floor." question is different between students in the USA and students in Canada. This finding is not surprising because the metric system is commonly used in Canada, but less popular in the USA. When measuring Math Self Efficacy across the USA and Canada, it is important to allow the "Calculating how many square metres of tiles you need to cover a floor." indicator to have different

measurement parameters across groups.

The final models used for testing measurement invariance were the original configural invariance model and a partial scalar invariance model with freely estimated measurement parameters for indicator variables two and three. These models had a chi-square difference of value of 11.181 with a random permutation testing p-value of .265; therefore, measurement invariance testing concluded with indicators two and three confirmed as non-invariant (although other justifiable conclusions could have been reached with different follow-up approaches). Comparisons of Math Self Efficacy across the USA and Canada should be done with the partial scalar invariance model.

4.3 Additional Testing Procedures

Continuing with the methods used in the second Monte Carlo simulation, the empirical data were also tested for measurement invariance using lavTestLRT and M*plus* DIFFTEST. The same initial models used for the omnibus test of measurement invariance were compared using both scaling corrections.

The results using the lavTestLRT function in R showed no statistically significant difference between the configural and scalar model in the initial test of measurement invariance $\Delta\chi^2(10.727) = 16.199$, $p = .122$. Therefore, if the data were analyzed only using the lavTestLRT correction provided in R with the lavaan package the test of measurement invariance would have concluded with measurement invariance being supported. No follow-up tests would have been conducted and the non-invariance found in indicators two and three with the random permutation testing would have gone undetected (with empirical data it is unknown if a Type I error was made by the random permutation test, or a Type II error was made with the lavTestLRT function).

The initial test for measurement invariance conducted in M*plus* showed that the first chi-square test via DIFFTEST comparing the configural model and the scalar model was statistically significant $\Delta\chi^2(30) = 62.025$, $p = .001$. The final follow-up test comparing the partial scalar invariance model (freely estimated parameters for indicator variables two and three) and the configural was

not statistically significant $\Delta\chi^2(22) = 33.554$, $p = .055$. Therefore the M*plus* results supported what was found using random permutation testing, indicators two and three are non-invariant across the USA and Canada.

4.4 Effect of Measurement Constraints

Although the empirical data do not allow a statement about whether or not it is correct to reject measurement invariance, they do allow an example demonstrating how freely estimating or constraining parameters for items 2 and 3 influence the group latent variable means and variances. When using the partial scalar invariance model with the measurement parameters for items 2 and 3 freely estimated, Math Self Efficacy for the Canadian group had a mean of .159 and a variance of .900 (these values were fixed at 0 and 1 for the USA group). If measurement invariance is not rejected and a full scalar invariance model is used, Math Self Efficacy for the Canadian group has a mean of .084 and a variance of .952 (these values were fixed at 0 and 1 for the USA group). This example shows that freeing the measurement parameters for items 2 and 3 to improve model fit results in larger group differences in parameters at the structural level.

4.5 Conclusions

This example demonstrated how random permutation testing can be applied to real data. After initially rejecting measurement invariance, random permutation testing was used to test individual indicator variables for invariance. This testing approach showed that two indicator variables were the source of the non-invariance. Removing the equality constraints from the factor loadings and thresholds for those indicators creates a partial scalar invariance model that can be used to compare the USA and Canada on Math Self Efficacy.

The comparisons with existing scaling procedures demonstrate how the results from M*plus* DIFFTEST and random permutation testing match up well, which is consistent with the results of Monte Carlo simulation two. Further, this example shows that using the lavTestLRT function in R

can lead to a different conclusion from the M*plus* DIFFTEST command when analyzing empirical data.

Chapter 5

Discussion

5.1 Research Questions

The purpose of the present research was to evaluate the use of random permutation testing applied to chi-square difference testing for measurement invariance with ordered-categorical indicator variables. The research was focused on models estimated with the popular WLSMV estimator. When models with ordered-categorical data are estimated with WLSMV, the chi-square difference test requires a scaling correction ensure an unbiased test because the assumption of multivariate normality is violated. The random permutation test was introduced as an alternative that is easily implemented in any statistical software, and as a method that should control Type I errors as well or better than existing methods. Three chi-square difference testing methods were compared to the random permutation test: 1) the correction implemented by the M*plus* DIFFTEST command, 2) the Satorra-Bentler correction as implemented by the lavaan package in R (lavTestLRT), and 3) the unadjusted chi-square difference test. There where five research questions developed for the present research, and these questions were answered with two Monte Carlo simulations.

5.1.1 Random Permutation Error Rate and Power

The first research question was:

- How does chi-square difference testing using random permutations perform in terms of Type I error rate and power? Are Type I errors close to .05? If Type I errors are close to .05, does the test show reasonable power that increases as a function of sample size?

The results of the present research suggest that random permutation testing has acceptable Type I error control. Further, even though only two sample size values were used in the Monte Carlo simulation there was a clear increase in power when larger groups were generated. The initial results showed a promising performance of random permutation testing and the test performed as expected. The answer to research question one is that, when the latent variable variable distributions are normal in all groups, Type I errors are close to .05 and power does increase as a function of sample size as would be expected.

5.1.2 Between Replication Differences

The second research question was:

- How do sample size, the number of response categories, and threshold symmetry influence the performance of chi-square random permutation testing?

As was previously mentioned, power increased as a function of sample size, however there was no clear influence of sample size on Type I error rates. The performance of random permutation testing was slightly influenced by the number of response categories for each item; factor loading differences were easier to detect with five response categories when compared to two response categories. This is likely because of the more sensitive measurement of the LRV when more response categories are used. Further, there was a decrease in power to detect factor loading differences when the indicator variable thresholds were non-symmetric. The most important conclusion from the simulation results was that sample size, number of response categories, and threshold symmetry do not influence the Type I error control of random permutation testing.

5.1.3 Comparison with Scaling Corrections

The third research question was:

- How does the random permutation testing approach perform compared to DIFFTEST in M*plus* and other chi-square difference test implementations?

In the present research the random permutation test showed better Type I error rate control and lower power when compared to the M*plus* DIFFTEST procedure. Further, the random permutation testing procedure consistently outperformed the lavTestLRT correction and an unadjusted chi-square difference test. It was not surprising to see the unadjusted chi-square difference test perform poorly, however the discrepancy between the lavTestLRT correction and the random permutation test was surprising. The lavaan implementation of the Satorra-Bentler correction was shown to be far too conservative with Type I error rates near or equal to 0 and low power. Follow-up analysis showed that an alternative to the default implementation of lavTestLRT performs well under the conditions used in simulation two (see Appendix C); however, the random permutation test still showed better Type I error control when the model is correctly specified (i.e., all latent variables are normally distributed).

5.1.4 Group Differences on Latent Variable Distribution

The forth research question was:

- Does the random permutation test outperform DIFFTEST in M*plus* when groups have different latent variable distributions?

The simulation showed that random permutation testing had better Type I error control under these conditions, but with reduced power when compared to M*plus* DIFFTEST. Importantly, the greatly inflated error rates with DIFFTEST observed by Suh (2015) were not observed in the present research with a CFA measurement model. This is likely due to differences in the data generation process. Suh generated data in the IRT framework where thresholds are interpreted on the scale of

the latent variable, whereas in the SEM framework the thresholds are on the scale of the LRV. In the present research the LRVs had a normally distributed error term, which results in a variable that is closer to normally distributed than the latent variable. Because the thresholds were applied to a variable that had a closer approximation of a normal distribution, the effect of the non-normal latent variable was reduced in the data generating process. In other words, the data generation process in the present research reduced the influence of the non-normal latent variable. The influence of group differences in latent variable distribution appeared to have similar influence on both the random permutation test and the M*plus* DIFFTEST procedure. Type I error rates were slightly inflated for both tests. The increased Type I error rates in these conditions can be attributed to the model misspecification that occurred when the LRVs and common latent variable were assumed to be normally distributed in the model estimation process. The differences in the distributions between the groups cannot be directly shown in the estimated models; instead the differences are forced into the estimated thresholds, even though in the data generation process the threshold were equal. In conclusion, there was no evidence suggesting that the random permutation test substantially outperforms the M*plus* DIFFTEST procedure when groups differ on latent variable distribution. In other words, both testing procedures showed a similar increase in Type I errors in the non-normal latent variable conditions.

5.1.5 Number of Random Permutation Shuffles

The fifth research question was:

- How does the number of permutations used influence the performance of the random permutation test? What is an appropriate number of random group shuffles to use?

The present research found that little change occurred when more than 500 permutation shuffles were used, however this should be generalized with caution. The design of the first simulation which found that 500 shuffles would be acceptable was designed solely to inform the design of simulation two. Using 500 rather than 1,000 permutation shuffles allowed simulation two to create less of a computational strain while running. It's important to acknowledge that the present

research did not manipulate the magnitude of non-invariance, nor did it manipulate population threshold differences. There could be situations not evaluated here where increasing the number of random permutation shuffles improves test performance. In summary, the finding of 500 random permutation shuffles being sufficient should be generalized with extreme caution. In other words, the present research has demonstrated that under certain conditions using 500 random permutation shuffles, but 500 is not necessarily always an appropriate number.

5.1.6 Summary

Overall, the random permutation test performed well in the Monte Carlo simulations. It showed acceptable Type I error control and power that increased as a function of sample size. The power of the random permutation test was also influenced by the number of response categories (greater with five response options) and threshold symmetry (greater with symmetric thresholds). The random permutation test clearly outperformed the lavTestLRT correction (with its default options) and the unadjusted chi-square difference test, and it showed better Type I error rate control with lower power when compared to the M*plus* DIFFTEST command. The M*plus* DIFFTEST command performed better than expected when groups differed on latent variable distributions and showed higher Type I error rates and greater power than the random permutation test. Lastly, the results of the present research suggest that, under the conditions of specified in simulation one, using 500 random permutation shuffles is sufficient and increasing the number of shuffles will rarely change the test result.

5.2 Suggestions for Random Permutation Testing

The results of the present research clearly show how the four chi-square difference testing approaches compare on their ability to control Type I errors. The present research suggests that when researchers desire to the Type I error rate of their test as close to .05 as possible, the random permutation testing procedure could be preferable to the three parametric approaches evaluated.

Because the lavTestLRT function in lavaan (using all of its default options) showed Type I error rates near or equal to zero and the unadjusted chi-square difference test showed error rates above 10%, it appears that the random permutation test should be strongly considered as an alternative testing procedure over these two methods. The poor performance of the unadjusted test is not surprising, however the lavTestLRT function performed far worse than expected. Researchers interested in using the lavTestLRT function should review Appendix C for information on how the performance of the function can be improved in situations similar to those in simulation two. The M*plus* DIFFTEST procedure showed reasonable Type I error control (a few conditions had rates above the nominal range) with greater power than the random permutation test. When both random permutation testing and the M*plus* DIFFTEST procedure are options, researchers need to decide if the improved Type I error control of the random permutation test is worth the decrease in power and the extra computation time required to repeatedly fit the same model to shuffled data.

5.2.1 Response Category Sparseness Across Groups

In the present research all cases of equal response categories across groups in random permutation shuffles were dealt with by resampling. Simulation one showed that resampling was rarely required in permutation shuffles, therefore the impact of using this method in the Monte Carlo simulations cannot be discussed. However, from a theoretical point of view resampling random permutation shuffles should be the preferred method for dealing with unequal responses because it will always provide chi-square difference values with the same degrees of freedom as the test value. If one uses category collapsing (as would likely be done for data obtained from participants), the chi-square difference values will have varying degrees of freedom and therefore be on a different scale. An important and easy to overlook detail about using the resampling approach is that this method narrows the population space of possible combinations of the grouping variable. Rather than simply pulling from all possible combinations of the grouping variable, combinations are pulled from possibilities where the number of observed responses for each indicator variable match for all groups. In unique cases where this is impossible, for instance only one respondent in a

single group endorsed a particular response option for an indicator, the researcher should collapse categories in the observed data prior to conducting the random permutation test.

In circumstances where a single indicator variable has very few responses on an extreme end of a response scale for all groups, sparseness across permutation shuffles will occur frequently. This could be handled by further restricting the possible combinations of the grouping variable to where the group frequencies for the problematic item are equal to the frequencies in the original data. For example, if an extreme response category for an indicator variable has only two responses across all groups, the random permutation shuffles used to build the reference distribution could carry the additional requirement of having two cases with the extreme response in each group. This constraint would likely become impossible if more than one indicator variable shows an extremely low response rate for a category.

5.3 Limitations

Perhaps the biggest limitation with the present research was that only 32 between replication conditions were used. The scope of the present research was focused on evaluating the random permutation test in comparison to existing scaling procedures across a variety of conditions, however the computational demands of the random permutation test led to a carefully chosen set of conditions. More between replication variables (e.g., latent variable mean difference between groups, threshold differences between groups, number of latent variables, etc.) could have been manipulated, and more than two levels of the manipulated variables could have been used.

An additional limitation is the narrow focus of the present research. The present research focused on measurement invariance testing when fitting models with ordered-categorical indicator variables and using the popular WLSMV estimator. The results of the present research should be generalized to scenarios that differ from the variable type and estimator used in the present research with extreme caution. No information about the random permutation test performs with non-normal continuous variables, or with models with a maximum likelihood estimator is provided

by the present research. Further, the results of the Monte Carlo simulations should be generalized with caution to model comparison scenarios other than comparing a configural invariance model to a scalar invariance model. Further, other than the conditions in which one of the simulated groups had a non-normally distributed latent variable, all models were correctly specified in the Monte Carlo simulations. The correct specification of the models limits the generalizability of the current findings to applied research where models are always misspecified to some degree.

5.4 Directions for Future Research

5.4.1 Additional Simulation Conditions

As was mentioned in the limitations section, the present research only used two levels of each simulation variable. Effects found here could be further explored in Monte Carlo simulations focusing on certain study variables that have more levels. Larger sample sizes, larger factor loading differences, population threshold differences, two or more latent variables, different group size ratios, and different numbers of groups are just a few ways that future research could expand on the present research design in an attempt to further evaluate the random permutation test. Additionally, conditions in which models are incorrectly specified (e.g., a minor cross loading in the population that is not modeled), could be included to better generalize to applied research. Manipulating additional variables with more than two levels would help identify additional situations in which random permutation testing or existing parametric tests do not perform well.

5.4.2 Number of Shuffles

Future research should also further explore the appropriate number of random permutation shuffles across a variety of conditions. The value of 500 used in the present research should not be considered the appropriate number across all situations. The standard error of the p-value obtained from random permutation testing could be used to inform researchers of whether or not they have used

a sufficient number of random permutations shuffles. Specifically, researchers could compute the standard error of their *p*-value estimate after running the first 100 random permutations. If their *p*-value is within two or three standard errors of .05, then an additional 50 or 100 random permutations can be used and then the standard error of the resulting *p*-value can be recomputed. An evaluation of how updating random permutation shuffles after reviewing an initial result influences the performance of the test would make an valuable contribution to the literature.

5.4.3 Different Model Estimators

The present research only evaluated the random permutation test applied to models fit with the WLSMV estimator. The random permutation test could be applied to models fit with any estimation procedure, however future research is needed to determine how well the procedure performs in these situations. Using a maximum likelihood estimator would allow alternative measures of model fit that are not available with weighted least squares (i.e., log likelihood, Akaike information criteria, Bayesian information criteria) to be used for the random permutation test. These alternative model fit measures could be used in the random permutation test and compared to the chi-square difference test with the SB correction (Satorra & Bentler, 2001).

5.4.4 Measures of Model Fit Differences

Perhaps the largest advantage of random permutation testing over existing chi-square difference testing procedures is that it makes no assumptions about the sampling distribution of the chi-square difference statistic. This also means that the random permutation test does not require the chi-square difference statistic to have any sampling distribution. Without the requirement of a sampling distribution, the random permutation test can use any measure of model fit and evaluate differences between models (e.g., CFI, TLI, RMSEA, etc). Future research should explore how these fit measures perform as test statistics in the random permutation testing procedure.

5.4.5 Summary

Clearly, additional Monte Carlo simulation research is needed to further evaluate the performance of the random permutation test. The second and main Monte Carlo simulation was small with only 32 between replication conditions. The present research was designed only to determine which variables influenced test performance, future research is required to further probe those effects and offer suggestions. Thinking beyond the scope of the present research–how to improve chi-square difference testing when models are fitted with the WLSMV estimator–perhaps the most important avenue for future research is evaluating model fit measures which do not have known sampling distributions in the random permutation test. The random permutation test performance was overall quite similar to the M*plus* DIFFTEST method (and the implementation of lavTestLRT discussed in Appendix C) when the chi-square difference statistic was used; however, the random permutation test has the ability to use change the change in CFI, TLI, or RMSEA between models. The present research was an evaluation of how random permutation testing performs in a very specific area where limited options exist. Future research should evaluate random permutation testing in situations where a test statistic has no known sampling distribution and compare its performance to available parametric tests.

5.5 Conclusion

The present research provided a promising initial evaluation of random permutation testing to handle chi-square difference testing for measurement invariance testing with ordered-categorical indicator variables. In the main Monte Carlo simulation the random permutation test was able to keep Type I errors close to .05 and it controlled Type I errors better than the parametric testing approaches evaluated. The present research suggests that researchers should consider the random permutation testing procedure a viable option for chi-square difference testing when evaluating measurement invariance with ordered-categorical indicator variables.

Bibliography

Asparouhov, T. & Muthén, B. (2006). Robust chi square difference testing with mean and variance adjusted test statistics. *Mplus Web Notes: No. 10*.

Babakus, E., Ferguson, C. E., & Joreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, 24(2), 222–228.

Beauducel, A. & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in cfa. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(2), 186–203.

Bliss, C. I. (1934). The method of probits. *Science*, 79(2037), 38–39.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, 37(1), 62–83.

Carpenter, L. C., Tompkins, S. A., Schmiege, S. J., Nilsson, R., & Bryan, A. (2010). Affective response to physical activity: Testing for measurement invariance of the physical activity affect scale across active and non-active individuals. *Measurement in Physical Education and Exercise Science*, 14(1), 1–14.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504.

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

Cyders, M. A. (2013). Impulsivity and the sexes: Measurement and structural invariance of the upps-p impulsive behavior scale. *Assessment*, 20(1), 86–97.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1), 1–26.

Elosua, P. (2011). Assessing measurement equivalence in ordered-categorical data. *Psicologica*, 32(2), 403–421.

Elosua, P. & Wells, C. S. (2013). Detecting dif in polytomous items using macs, irt and ordinal logistic regression. *Psicologica*, 34(2), 327–342.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd Ltd.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4), 521–532.

Fox, J. (2010). *polycor: Polychoric and Polyserial Correlations*. R package version 0.7-8.

Garnaat, S. L. & Norton, P. J. (2010). Factor structure and measurement invariance of the yale-brown obsessive compulsive scale across four racial/ethnic groups. *Journal of Anxiety Disorders*, 24(7), 723 – 728.

Higgins, J. J. (2004). *Introduction to modern nonparametric statistics*. Belmont, CA: Cenage Learning.

Horn, J. L. & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18(3), 117–144. PMID: 1459160.

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

Johnson, D. R. & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, 48(3), 398–407.

Johnson, P. E. (2015). *portableParallelSeeds: Allow Replication of Simulations on Parallel and Serial Computers*. R package version 0.96.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling*. New York, NY: Guilford.

Lee, S.-Y., Poon, W.-Y., & Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, 48(2), 339–358.

Little, T. D. (2013). *Longitudinal Structural Equation Modeling*. New York: Guilford Press.

Lubke, G. H. & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, 11(4), 514–534.

Martens, M. P., Pederson, E. R., LaBrie, J. W., Ferrier, A. G., & Cimini, M. D. (2007). Measuring alcohol-related protective behavioral strategies among college students: Further examination of the protective behavioral strategies scale. *Psychology of Addictive Behaviors*, 21(3), 307–315.

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

Muthén, B. & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in mplus. *Mplus Web Notes: No. 4*.

Muthén, B. & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal likert variables. *British Journal of Mathematical and Statistical Psychology*, 38(2), 171–189.

Muthén, L. K. & Muthén, B. O. (1998–2015). Mplus user's guide. seventh edition.

O'Brien, R. M. (1985). The relationship between ordinal measures and their underlying values: Why all the disagreement? *Quality and Quantity*, 19(3), 265–277.

OECD (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD Publishing.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4), 443–460.

Osman, A., Lamis, D. A., Freedenthal, S., Gutierrez, P. M., & McNaughton-Cassill, M. (2014). The multidimensional scale of perceived social support: Analyses of internal reliability, measurement invariance, and correlates across gender. *Journal of Personality Assessment*, 96(1), 103–112. PMID: 24090236.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. ii. the correlation coefficient test. *Supplement to the Journal of the Royal Statistical Society*, 4(2), 225–232.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Randall, J. & Engelhard, G. (2010). Using confirmatory factor analysis and the rasch model to assess measurement invariance in a high stakes reading assessment. *Applied Measurement in Education*, 23(3), 286–306.

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be

treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions. *Psychological methods*, 17(3), 354.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.

Rossum, G. (1995). *Python Reference Manual*. Technical report, Amsterdam, The Netherlands, The Netherlands.

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2), 167–180.

Satorra, A. (1999). *Scaled and adjusted restricted tests in multi-sample analysis of moment structures*. Economics Working Papers 395, Department of Economics and Business, Universitat Pompeu Fabra.

Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4), 507–514.

Schmitt, N. & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, 18(4), 210–222. Research Methods in Human Resource Management.

Skriner, L. C. & Chu, B. C. (2014). Cross-ethnic measurement invariance of the scared and ces-d in a youth sample. *Psychological Assessment*, 26(1), 332–337.

StataCorp (2015). *Stata Statistical Software: Release 14*.

Steenkamp, J. E. M. & Baumgartner, H. (1998). Assessing measurement invariance in cross national consumer research. *Journal of Consumer Research*, 25(1), pp. 78–107.

Suh, Y. (2015). The performance of maximum likelihood and weighted least square mean and variance adjusted estimators in testing differential item functioning with nonnormal trait distributions. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4), 568–580.

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4–70.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer, fourth edition. ISBN 0-387-95457-0.

Widaman, K. F. & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *The science of prevention; methodological advances from alcohol and substance abuse research*, (pp. 281 – 324).

Wirth, R. J. & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12(1), 58–79.

Appendix A

Polychoric Correlation Example

In order to further demonstrate how polychoric correlations are computed, a brief example is provided here for a case where the polychoric correlation between two three-point Likert-type indicators needs to be estimated. The two variables used in the example below have a polychoric correlation of roughly .57, this example will show how this estimate can be chosen when conducting analysis manually (as was previously mentioned, this estimate can quickly be obtained with the "polycor" package in R). The first step is to compute the thresholds for each item. Because there are three response options, there are two thresholds for each variable that need to be estimated. An example of the response frequency table that can be used when calculating this by hand can be seen in Table A.1.

Table A.1: Response Frequencies for Two Items

		Item 2		
	Response Options	1	2	3
Item 1	1	40	10	10
	2	15	40	20
	3	10	15	40

The cumulative proportions are used in the inverse of Equation 1.15 to yield the threshold separating the observed response from the next higher response on the distribution of the LRV. For item 1 the first threshold value needs to separate the lower 30% $(100(40+10+10)/200)$ of the LRV from the upper 70%, the inverse of the normal CDF returns a threshold value of -.524. Further

the second threshold for item 1 needs to separate the lower 67.5% $(100(60+15+40+20)/200)$ of the LRV from the upper 37.5%, this results in a threshold value of .454. Using the same process, the thresholds for item 2 are -.454 and .385.

After the thresholds are computed, those values are used to evaluate the log likelihood of possible values of the polychoric correlation. If one were conducting the analysis manually, many possible polychoric correlations would be used to compute the conditional probability of the observed responses. The accepted polychoric correlation is the one that shows the highest log likelihood value provided by Equation 1.16. The subsequent work for this example shows the calculations when a test polychoric correlation value of .57 is used. The calculated threshold values are used in the bivariate CDF (see Equation 1.18) to find the probability of being in a certain response category or lower when the polychoric correlation is .57. In this example there are nine possible response patterns that require a likelihood estimate for Equation 1.16. The values in Table A.2 represent the values used in Equation 1.18 to compute the bivariate probability for each response pattern given a polychoric correlation of .57. In order to be consistent with the notation in Equation 1.18, "x" represents item 1 threshold values and "y" represents item 2 threshold values.

Table A.2: Thresholds Used for Polychoric Correlation Computation

		Item 2 Response		
		1	2	3
Item 1 Response	1	x = -.454, y = -.524	x = -.454, y = .454	x = -.454, y = ∞
	2	x = .385, y = -.524	x = .385, y = .454	x = .385, y = ∞
	3	x = ∞, y = -.524	x = ∞, y = .454	x = ∞, y = ∞

The cumulative probability for each response pattern is calculated using the binoimal distribution CDF shown in Equation 1.18. The resulting values are shown in Table A.3. The cumulative probabilities for each response pattern are then used in Equation 1.17 to yield the probablity of each individual observed response, the results are shown in Table A.4. Subtracting the bivariate probability of adjacent response patterns provides the probability of being in a single response pattern, rather than having two responses equal to or less than a certain value.

Table A.3: Cumluative Probabilities with Polychoric Correlation of .57

		Item 2		
		1	2	3
Item 1	1	$\Phi_2(-.454, -.524) = .177$	$\Phi_2(-.454, .454) = .290$	$\Phi_2(-.454, \infty) = .325$
	2	$\Phi_2(.385, -.524) = .266$	$\Phi_2(.385, .454) = .523$	$\Phi_2(.385, \infty) = .650$
	3	$\Phi_2(\infty, -.524) = .300$	$\Phi_2(\infty, .454) = .675$	$\Phi_2(\infty, \infty) = 1$

Table A.4: Response Pattern Probabilities with Polychoric Correlation of .57

Response Pattern	Pattern Probability
1, 1	$\pi_{11} = .177$
1, 2	$\pi_{12} = .290 - .177 = .113$
1, 3	$\pi_{13} = .325 - .290 = .035$
2, 1	$\pi_{21} = .266 - .177 = .088$
2, 2	$\pi_{22} = .523 - .266 - .290 + .177 = .144$
2, 3	$\pi_{23} = .650 - .523 - .325 + .290 = .093$
3, 1	$\pi_{31} = .300 - .266 = .034$
3, 2	$\pi_{32} = .675 - .300 - .523 + .266 = .118$
3, 3	$\pi_{33} = 1 - .675 - .650 + .523 = .198$

The resulting probability values are then used in Equation 1.16 to compute the log of the likelihood for the responses given a certain polychoric correlation value. This weights the probability of each response pattern by the number of respondents showing the pattern. The computed log likelihood given a polychoric correlation of .57 is shown in Equation A.1. In this example the scaling parameter (K) shown in Equation 1.16 is omitted because it is a constant.

$$log(L)|(\rho = .57) = 40 * log(.177) + 10 * log(.113) + 10 * log(.035) + 15 * log(.088)$$
$$+ 40 * log(.144) + 20 * log(.093) + 10 * log(.034) + 15 * log(.118) + 40 * log(.198) = -416.75 \tag{A.1}$$

Because polychoric correlations are not computed using a closed-form equation, one can iterate across different possible correlation coefficients to find the value that maximizes the log of the likelihood. Trying values between -1 and 1 shows how the likelihood of the observed data changes

as a function of the polychoric correlation. When creating this example, 2,001 different polychoric correlation values were evaluated, values between and including -1 and 1 in intervals of .001. Changes in the log likelihood as a function of the polychoric correlations can be seen in Figure A.1. In this example the polychoric correlation coefficient that maximizes the log likelihood function is .57, this value produces the highest log likelihood value of -416.75, therefore .57 is the estimated polychoric correlation. A smaller interval between tested polychoric correlations can be used across iterations in order to provide a more precise estimate. When fitting a CFA model with ordered-categorical or dichotomous data, all non-redundant pairs of variables would have their polychoric correlation computed.

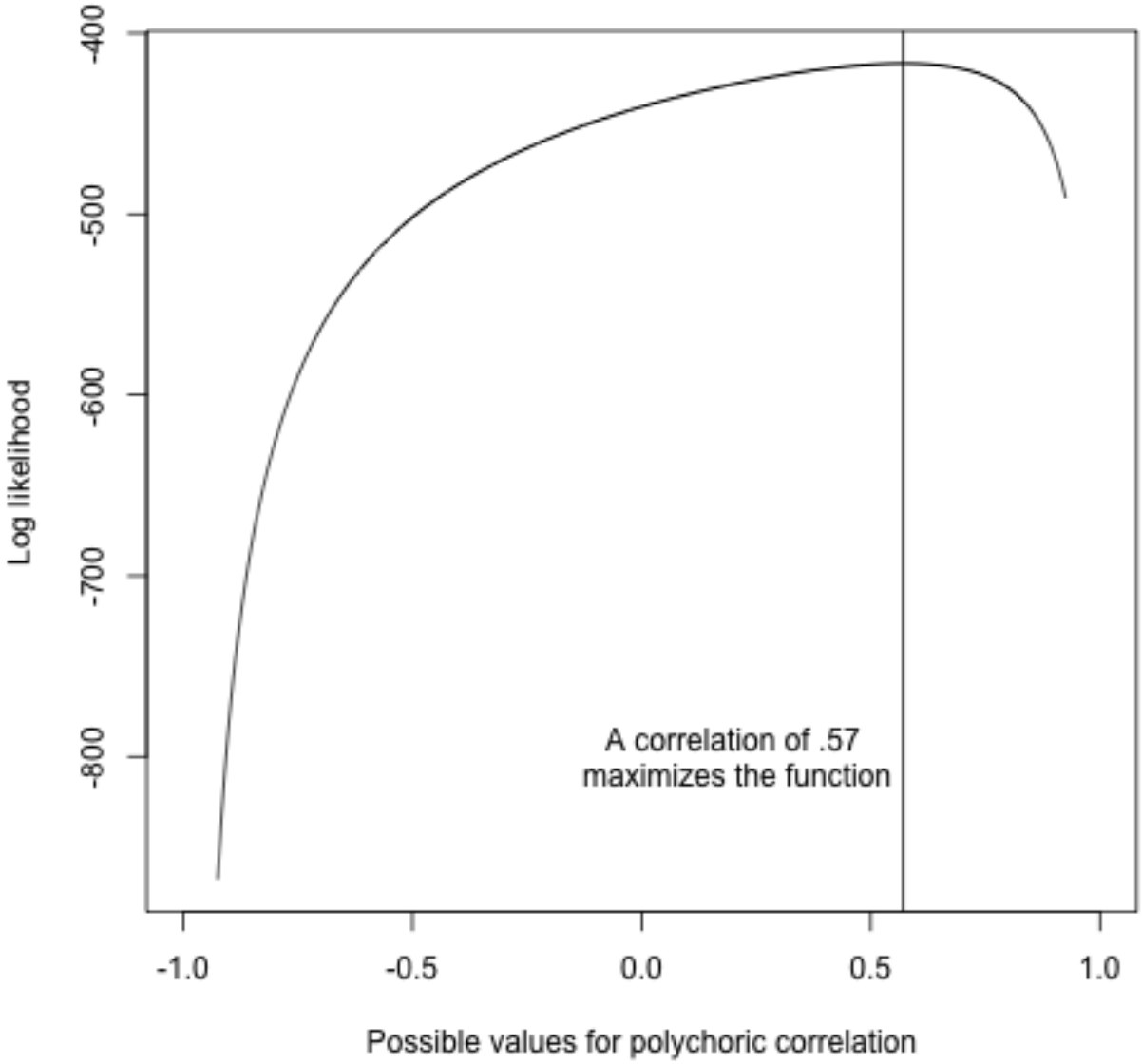

Figure A.1: Log Likelihood as a Function of Polychoric Correlation

Appendix B

Random Permutation Test R Syntax

```
library(lavaan)
library(plyr)
library(simsem)
set.seed(123)
genModel <- "f1 =~ .6*y1 + .6*y2 + .6*y3 + .6*y4
y1 | 0*t1
y2 | 0*t1
y3 | 0*t1
y4 | 0*t1
f1 ~~ 1*f1"
genModel2 <- "f1 =~ .3*y1 + .3*y2 + .6*y3 + .6*y4
y1 | 0*t1
y2 | 0*t1
y3 | 0*t1
y4 | 0*t1
f1 ~~ 1*f1"
dat1 <- simulateData(genModel, sample.nobs = 300)
dat2 <- simulateData(genModel2, sample.nobs = 300)
dat1$group <- 1
dat2$group <- 2
dat <- rbind.fill(dat1, dat2)
```

```
model <- "f1 =~ y1 + y2 + y3 + y4
y1 ~*~ 1*y1
y2 ~*~ 1*y2
y3 ~*~ 1*y3
y4 ~*~ 1*y4"
parentout <- cfa(model, dat, group = "group", ordered = c("y1", "y2", "y3",
    "y4"))
nestedout <- cfa(model, dat, group = "group", ordered = c("y1", "y2", "y3",
    "y4"), group.equal = c("loadings", "thresholds"))
pfit <- fitMeasures(parentout, fit.measures = c("chisq.scaled", "df.scaled"))
nfit <- fitMeasures(nestedout, fit.measures = c("chisq.scaled", "df.scaled"))
chisq.diff2 <- nfit[1] - pfit[1]
df.diff2 <- nfit[2] - pfit[2]

nperms <- 1000
permutedVals2 <- rep(NA, nperms)
for (i in 1:nperms){
dat$group <- sample(dat$group, nrow(dat), replace = FALSE)
        parentout <- cfa(model, dat, group = "group", ordered = c("y1", "y2",
            "y3", "y4"))
        nestedout <- cfa(model, dat, group = "group", ordered = c("y1", "y2",
            "y3", "y4"), group.equal = c("loadings", "thresholds"))
            permutedVals2[i] <- fitMeasures(nestedout, "chisq.scaled") -
            fitMeasures(parentout, "chisq.scaled")
}
mean(ifelse(permutedVals2 > chisq.diff2, 1, 0))
```

Appendix C

Further Evaluation of lavTestLRT

The results in simulation two showed surprisingly low Type I error rates for the scaled chi-square difference test implemented by the lavTestLRT function in lavaan (Rosseel, 2012). The testing procedure evaluated initially showed Type I errors near or equal to zero and low power. The implementation of the lavTestLRT function in simulation two used all of the default options for the function in lavaan; this was done because a literature review did not reveal suggested options for use when comparing two models with ordered-categorical data fitted with the WLSMV estimator. Further review of the lavTestLRT function showed that the "A.method" argument determines how the Jacobian of the constraint function is computed. The default for this function in lavaan version 0.5-22 is "exact", with the only other option being "delta." All 32 between replication conditions from simulation two were re-tested with the lavTestLRT function while specifying "A.method = "delta"." The re-testing was done using the exact same randomly generated data from simulation two, therefore the results are directly comparable. Table C.1 and Figure C.1 show that when the "delta" method is used the lavTestLRT function performs much better under the conditions of simulation two.

Table C.1: Rejection Rates for lavTestLRT with Delta Method

Invariance	Sample Size	Responses	Threshold Symmetry	Skewed LV	Rejection Rate
Yes	150	2	Symmetric	No	.056
Yes	300	2	Symmetric	No	.050
Yes	150	5	Symmetric	No	.054
Yes	300	5	Symmetric	No	.053
Yes	150	2	Non-Symmetric	No	.054
Yes	300	2	Non-Symmetric	No	.065
Yes	150	5	Non-Symmetric	No	.047
Yes	300	5	Non-Symmetric	No	.056
Yes	150	2	Symmetric	Yes	.059
Yes	300	2	Symmetric	Yes	.056
Yes	150	5	Symmetric	Yes	.064
Yes	300	5	Symmetric	Yes	.085
Yes	150	2	Non-Symmetric	Yes	.055
Yes	300	2	Non-Symmetric	Yes	.054
Yes	150	5	Non-Symmetric	Yes	.045
Yes	300	5	Non-Symmetric	Yes	.059
No	150	2	Symmetric	No	.292
No	300	2	Symmetric	No	.543
No	150	5	Symmetric	No	.464
No	300	5	Symmetric	No	.794
No	150	2	Non-Symmetric	No	.225
No	300	2	Non-Symmetric	No	.427
No	150	5	Non-Symmetric	No	.335
No	300	5	Non-Symmetric	No	.712
No	150	2	Symmetric	Yes	.234
No	300	2	Symmetric	Yes	.476
No	150	5	Symmetric	Yes	.412
No	300	5	Symmetric	Yes	.772
No	150	2	Non-Symmetric	Yes	.268
No	300	2	Non-Symmetric	Yes	.531
No	150	5	Non-Symmetric	Yes	.391
No	300	5	Non-Symmetric	Yes	.736

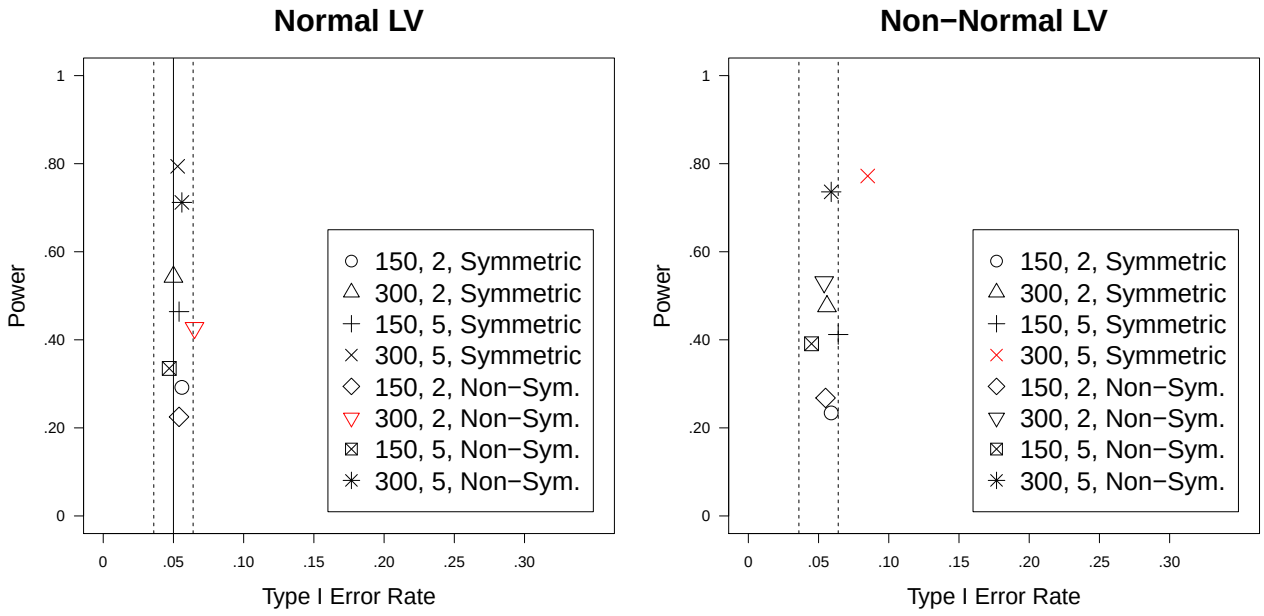

Figure C.1: Evaluations of lavTestLRT with Delta Option