ABSTRACT

Title of Dissertation: ROBUST MEANS MODELING:
AN ALTERNATIVE FOR REPEATED
MEASURES DESIGNS

Xiulin Mao, Doctor of Philosophy, 2018

Dissertation directed by: Professor Gregory R. Hancock
Department of Human Development and
Quantitative Methodology

The study presents a series of alternative ANOVA-based methods which offered a remedy to the traditional *F* test to accommodate violations of normality and sphericity assumptions. Specially Robust Means Modeling (RMM), developed from structured means modeling (SMM), a branch of structural equation modeling (SEM), is introduced to circumvent the sphericity assumption while alleviating the violation of normality assumption. Maximum likelihood, Satorra-Bentler scaled chi-square, asymptotic distribution-free (ADF) methods and its corrections, as well as residual-based ADF methods (RES) and its corrections, are included in this RMM category.

A Monte Carlo simulation is designed to evaluate Type I error robustness and power of the ANOVA-based methods and the proposed RMM methods under the fully crossed conditions including degree of non-sphericity, degree of non-normality, sample size, and the number of levels of the repeated measures. The study gains strong ground for RMM methods to be recommended over ANOVA-based methods under almost all conditions except when the model was complex (i.e. 8 levels) and sample size was small (15, 30).

REPEATED MEANS MODELING:
AN ALTERNATIVE FOR REPEATED MEASURES DESIGNS

By

Xiulin Mao

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Gregory R. Hancock, Chair
Professor Jeffrey R. Harring
Professor Hong Jiao
Professor Leigh A. Leslie
Professor Ji Seung Yang

© Copyright by
Xiulin Mao
2018

Acknowledgements

First of all, I would love to extend my greatest gratitude to my advisor, Dr. Gregory Hancock. As my academic father and mentor and my life model, he has led me to the wonderful world of research and has been providing the continuous support and patient guidance through my Ph.D. study and research. Most importantly, he has never lost confidence in me. Without his trust and his encouragement, I would not have reached this stage.

My sincere thanks also go to Dr. Jeffrey Harring, who has spent so much of his time helping me conduct research and Dr. Hong Jiao, who has been encouraging me all these years. I am also deeply thankful to the rest of my committee: Dr. Ji Seung Yang and Dr. Leigh Leslie, for their insightful comments and encouragement, and other professors in our EDMS program: Dr. Laura Stapleton, Dr. Robert Lissitz, and Dr. George Macready who have provided wonderful guidance for me.

Nobody has been more important to me than my family members. I would like to extend my thanks and love to my husband, Yanghua, and my son, Tianze, as well as my parents and parents-in-law, whose love and support are with me in whatever I pursue.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Understanding the sources of stability and change in variables is of interest in virtually every discipline in the social sciences. For example, a psychologist may want to map the development of cognitive abilities in children. In this case, he/she may measure the outcomes of interest repeatedly across time in order to study the change over time and gather enough information to form a trajectory for the development of cognitive abilities in children. On the other hand, each subject can also be measured multiple times on the same dependent variables when he/she is exposed to two or more conditions, such as competing treatments, to test for a difference in those conditions. Both designs are known as *repeated measures designs*, also called *within-subjects designs*, which can be summarized as *N* subjects who are observed on each of *K* successive occasions corresponding to different conditions or different time points (Jensen, 1982). The repeated measures design typically requires far fewer subjects than the between-subject design, which makes the repeated measures design appealing in real world applications. The reason for this benefit is that measurements within the same subject are virtually always positively correlated. If this dependence among these measurements can be accounted for properly by statistical analysis, greater precision of parameter estimates, and more efficient inferential analyses can be achieved than in between-subject designs (Lix & Keselman, 2010), and increased power to detect true treatment effects can be obtained (Maxwell & Delaney, 1990).

For testing data gathered in such repeated measure designs, traditional analysis of variance (ANOVA) is still commonly used. Unfortunately, this approach requires stringent assumptions routinely violated with real world data, including normality of score distributions and sphericity.

Sphericity is a condition assumed in repeated measures ANOVA that is required for the resulting test statistic to follow an *F* distribution. It refers to the homogeneity of the treatment-difference variances, or the variances for differences between all the possible pairs of treatment scores of the repeated measures (Huynh & Feldt, 1970; Keselman, Keselman, & Shaffer, 1991). Suppose the repeated measures contain *K* levels of treatments, the variance of a difference between any two levels of treatments can be defined as

$$\sigma^2_{l-j}(Y_l - Y_j) = \sigma^2_l + \sigma^2_j - 2\sigma_{lj} \qquad (1)$$

where $l = 1, 2, ..., K$, $j = 1, 2, ..., K$ and $l \neq j$; Y_l is one set of treatment scores with σ^2_l being its variance, Y_j is another set of treatment scores with σ^2_j being its variance, and σ_{lj} is the covariance of the two set of scores.

The violations of the assumptions of normality and sphericity can render the validity of inferences from the traditional repeated measures ANOVA test somewhat questionable. To address such violations, various adjustments to the ANOVA test have been proposed, but they too have various limitations.

In recent decades, structural equation modeling (SEM) has gained increasing attention across fields such as education, psychology, sociology, and economics due to its versatility in modeling relations among both measured and latent variables. For the proposed study, two aspects of SEM are directly relevant for repeated measures designs. First, approaches within SEM does not need to make assumptions about variances in order to yield a proper test statistic, thereby circumventing the sphericity assumption altogether. Second, a branch of SEM called *structured means modeling* (SMM) can be used with robust rescaling corrections to estimate parameters in repeated measures designs precisely while alleviating the violation of the normality assumption.

During the past few decades, researchers have been taking great endeavors to find the best method for repeated measures designs as they have been widely used in social and behavioral studies. The current dissertation study aims to compare the performance of the adjusted ANOVA-based methods with those derived from the robust SMM framework to determine the optimal strategies for analysis within repeated measures designs. The current study is a simulation, requiring the integration of multiple software packages, and examining a wide variety of real world data scenarios. As such, the current study provides a useful guideline for practitioners in various fields to carry out repeated measures designs or longitudinal studies in a more modern, and more importantly, a more versatile and valid way.

Chapter 2: Literature Review

Traditional Procedures for One-way Repeated Measure Designs

As mentioned earlier, a repeated measures design can be used to either study the change over time (the same subjects receive the same treatment repeatedly across time) or to test for differences in different conditions (the same subjects are measured multiple times on the same dependent variables when they are exposed to two or more conditions). It can be used with only one variable measured repeatedly overtime or in different conditions (one-way repeated measures design) or combined with other conditions (i.e. combined with between-subject design, multivariate repeated measures, etc.). The former is the focus of the current study.

Traditional Unadjusted *F* Test. Suppose in the one-way repeated measures design, the repeated measures factor contains K levels ($k = 1, 2, \ldots, K$) and N subjects ($i = 1, 2, \ldots, N$) are observed and measured repeatedly on this factor, the model can be defined as

$$Y_{ik} = \mu_{\bullet\bullet} + \beta_{\bullet k} + e_{ik} \qquad (2)$$

where $\mu_{\bullet\bullet}$ is the grand mean of the scores; Y_{ik} is the score for the ith subject in the kth treatment level; $\beta_{\bullet k}$ is the treatment effect of the kth level; e_{ik} is the error effect for Y_{ik}. The hypothesis to be tested is

$$\mathrm{H}_0\text{: } \boldsymbol{\mu}_{\cdot 1} = \boldsymbol{\mu}_{\cdot 2} =, \ldots, = \boldsymbol{\mu}_{\cdot K} \qquad (3)$$

where $\boldsymbol{\mu}_{\cdot K}$ is the mean of *K*th treatment level.

With the level of significance (α), the critical value for the traditional unadjusted *F* test is:

$$F[\alpha;\ K-1,(N-1)(K-1)] \tag{4}$$

with $df_{numerator} = K-1$ and $df_{denominator} = (N-1)(K-1)$.

The validity of this test rests on the assumptions of normality, independence of errors, and sphericity. If all of these assumptions are satisfied, the traditional unadjusted *F* test is believed to be most powerful for detecting treatment effects (Keselman, Algina, & Kowalchuk, 2001). However, in reality, assumptions such as normality and sphericity are typically violated.

Sphericity is defined as the homogeneity of all treatment difference score variances; that is, the variances for all possible differences between *K* levels of treatment are equal. This assumption is almost impossible to be satisfied in real-world data as when repeated measures are used, the variances actually tend to increase over time. Moreover, a test of sphericity assumption is also sensitive to violation of the normality assumption (Lix & Keselman, 2010). When the sphericity assumption is violated, the traditional *F* test becomes inflated, causing too many false rejections of the null hypothesis, or increased Type I error rate (e.g., Algina & Keselman, 1997; Keselman et al., 2001; Lix, Keselman, & Keselman, 1996; Maxwell & Delaney, 1990).

And as the degree of deviation from sphericity escalates, the Type I error rate becomes increasingly inflated (Keselman et al., 2001). Therefore, the traditional *F* test fails to be robust to the violation of sphericity assumption (Box, 1954; Huynh & Feldt, 1980; McCall & Appelbaum, 1973; Quintana & Maxwell, 1994). Similarly, when the normality assumption is violated, the traditional *F* test also yields inflated Type I error rates and the power discreases.

In order to provide a remedy to the traditional *F* test, many other ANOVA-based methods believed to be insensitive to violations of the assumptions were proposed, including adjusted degrees of freedom univariate *F* tests, multivariate approach, as well as robust estimator.

Adjusted Degrees of Freedom Univariate *F* Tests. In order to alleviate the violation of the sphericity assumption when using the traditional analytic approach, Box（1954） suggested that the degrees of freedom be adjusted by a reduction factor ε

$$\varepsilon = \frac{K^2(\bar{v}_{jj} - \bar{v}_{..})}{(K-1)[\Sigma\Sigma v_{jl}^2] - (2K\Sigma\bar{v}_{j.}^2) + (K^2\bar{v}_{..}^2)} \qquad (5)$$

where v_{jl} is an element in row j and column l of the population covariance matrix, $\bar{v}_{jj}$ is the mean of the variances (the diagonal elements) in the population covariance matrix, $\bar{v}_{j.}$ is the mean of the entries in the *j*th row of the population covariance matrix, and $\bar{v}_{..}$ is the mean of all elements in the population covariance matrix. The population sphericity parameter is denoted as ε which indicates the extent to which the covariance matrix departs from sphericity. It falls between $1/(K-1)$ and 1 where 1 denotes no violation of sphericity assumption while lower values imply a departure from the assumption (see, e.g., Algina & Keselman, 1997; Quintana & Maxwell, 1994). However, as a population parameter, ε is usually unknown and therefore must be estimated from a sample (Keselman, Algina, Kowalchuk, & Wolfinger, 1999; Quintana & Maxwell). Therefore, an approximate critical value based on the altered df is

$$F[\alpha;(K-1)\varepsilon,\ (N-1)(K-1)\varepsilon] \tag{6}$$

with $df_{numerator} = (K-1)\varepsilon$ and $df_{denominator} = (N-1)(K-1)\varepsilon$.

Three *df* adjustments, widely available in statistical packages, were proposed to estimate this sphericity parameter, including Geisser-Greenhouse (GG) lower-bound adjustment (Geisser & Greenhouse, 1958), Box's $\hat{\varepsilon}$ adjustment (Box, 1954), and the Huynh-Feldt (HF) $\tilde{\varepsilon}$ adjustment (Huynh & Feldt, 1976), which were described below.

Geisser - Greenhouse Lower–Bound Adjustment (GG). Geisser and Greenhouse (1958) acknowledged that the lowest possible population value for ε would be $\varepsilon = 1/(K-1)$ (Maxwell & Delaney, 1990). For example, if $K = 3$, ε is larger than or equal to 0.5. Therefore, the smallest possible degree of freedom achieved through this lower-bound adjustment equals 1 for the numerator and $N-1$ for the denominator. Because a larger *F* critical value could be obtained with smaller degrees of freedom, the critical value corresponding to these degrees of freedom can be determined as $F[\alpha;\,1,\;N-1]$ with $df_{numerator}=1$ and $df_{denominator}=N-1$. Using this GG lower bound adjustment makes a conservative test of the null hypothesis, leading to fewer rejections of the null hypothesis than expected at a nominal α level. The simplicity in calculating the GG lower bound adjustment has led to its widespread use.

Box's Adjustment (BOX). Proposed by Box (1954) and implemented by Geisser and Greenhouse (1958), Box's $\hat{\varepsilon}$ adjustment employed observed sample data to estimate the population value of ε based on the approximate distribution of F presented by Box (1954). Therefore, this adjustment is usually called Box's $\hat{\varepsilon}$ adjustment. As the sample value is generally larger than the theoretical lower bound, it is a less conservative adjustment for df (Maxwell & Delaney, 1990). $\hat{\varepsilon}$ can be calculated as

$$\hat{\varepsilon}=\frac{K^2(\overline{X}_{jj}-\overline{X}_{..})}{(K-1)[\Sigma\Sigma X_{jl}^2]-(2K\Sigma\overline{X}_{j.}^2)+(K^2\overline{X}_{..}^2)} \qquad (7)$$

where X_{jl} is an element in row j and column l of the sample covariance matrix, $\overline{X}_{jj}$ is the mean of the variances (the diagonal elements) in the sample covariance matrix, $\overline{X}_{j.}$ is the mean of the entries in the *j*th row of the sample covariance matrix, and $\overline{X}_{..}$ is the mean of all elements in the sample covariance matrix. Thus, the critical value will be approximated as $F[\alpha;(K-1)\hat{\varepsilon},\ (N-1)(K-1)\hat{\varepsilon}]$ with $df_{numerator}=\hat{\varepsilon}(K-1)$ and $df_{denominator}=\hat{\varepsilon}(N-1)(K-1)$.

Due to the complexity in its calculation, $\hat{\varepsilon}$ was not widely used until the mid-1980s when statistical packages were developed that included it. Compared with the GG lower bound adjustment, Box's $\hat{\varepsilon}$ adjustment is able to control Type I error rates better and is more powerful than GG (Maxwell & Delaney, 1990). The disadvantage of the $\hat{\varepsilon}$ adjustment, however, is that it tends to over-adjust the degrees of freedom and underestimate population ε. In an attempt to correct this bias in $\hat{\varepsilon}$, Huynh and Feldt (1976) developed another adjustment, which is referred to as the Huynh-Feldt (HF) $\tilde{\varepsilon}$ adjustment.

Huynh-Feldt Adjustment (HF). The Huynh-Feldt $\tilde{\varepsilon}$ adjustment is similar to Box's $\hat{\varepsilon}$ adjustment in the way that it also relies on the observed sample data. It can be expressed as function of $\hat{\varepsilon}$ as

$$\tilde{\varepsilon}=\frac{N(K-1)\hat{\varepsilon}-2}{(K-1)[N-1-(K-1)\hat{\varepsilon}]}, \qquad (8)$$

thereby adjusting $\hat{\varepsilon}$ upward. However, it tends to overestimate ε and can sometimes be larger than 1; in such cases it is set equal to 1 given that the upper bound of ε is 1 (Maxwell & Delaney, 1990). Thus, the critical value can be approximated as $F[\alpha;(K-1)\tilde{\varepsilon},\ (N-1)(K-1)\tilde{\varepsilon}]$ with $df_{numerator}=\tilde{\varepsilon}(K-1)$ and $df_{denominator}=\tilde{\varepsilon}(N-1)(K-1)$. Among the three adjustments, the HF $\tilde{\varepsilon}$ adjustment yields the smallest critical value and thus can lead to more rejections of the null hypothesis.

As mentioned earlier, these three *df* adjustments are mainly designed to accommodate violations of the sphericity assumption. In real world conditions, violations of both sphericity and normality are the main challenges practitioners face when they use repeated measures ANOVA. In fact, non-normality seems more of a rule than an exception (Micceri, 1989), and when coupled with sphericity assumption violations may distort the analytic results even further. Regarding the degree to which sphericity assumption is violated, when $\varepsilon \geq .75$, it is considered as moderate violation of sphericity assumption. $\varepsilon = .41$ is regarded as the lower limit of values often found in the beharioral literature. Even though $\varepsilon = 0.08$ theoretically exists, they are very unlikely to occur in real world research (Quintana & Maxwell, 1994).

The Multivariate Approach. While the traditional ANOVA approach treats several measurements over time/occasions as a single dependent variable repeatedly measured, the multivariate approach treats the repeated measures as separate dependent variables by creating difference variables based on repeated measurements (Kieffer, 2002). In this way, heterogeneous covariance structures are allowed to exist and thus the sphericity assumption is able to be circumvented (Keselman et al., 1999). For example, for a one-way repeated measures design with *K* levels, the multivariate approach creates $K-1$ difference variables or contrasts based on the original K

levels and then analyzes the new variable set. The null hypothesis to be tested, using Hotelling's T^2, is that the vector of population means of these $K-1$ difference variables equals a zero vector (McCall & Appelbaum, 1973). The test statistic is

$$F = \frac{N-K+1}{K-1} T^2 , \tag{9}$$

and is compared to the critical value obtained with $df_{numerator} = (K-1)$ and $df_{denominator} = (N-K+1)$.

The multivariate approach has less stringent requirements than the conventional test in that it is not dependent on the sphericity assumption, but only requires that the covariance matrix be positive definite (Keselman et al., 2001). But it does tend to require a larger sample size than the univariate approach to detect an effect given that it regards each measurement by an individual as a separate dependent variable (Kieffer, 2002), which can offset the advantage of repeated measures design. When sample size is not large enough relative to the number of the levels of the repeated measures, the multivariate approach is not as powerful as univariate approach. Moreover, if the sample size is smaller than the number of repeated measures minus one, multivariate statistic cannot be calculated (Fernándes, Vallejo, Livacic-Rojas, Herrero, & Cuesta, 2009). However, the most important disadvantage of the multivariate approach is that it rests on the normality assumption and is quite sensitive to extreme skew (e.g., Berkovits, Hancock, & Nevitt, 2000; Harwell & Serlin, 1977; Kieffer, 2002; Lix & Keselman, 2010), which is relatively common in real data analysis.

ANOVA-Based Robust Estimator. The **t**rimmed means method has been widely used to derive robust measures of central tendency and variability to deal with the issue of non-normality. A trimmed mean is obtained by removing a predetermined portion of the largest and smallest observations on each measurement occasion and computing the mean of the remaining observations (Lix & Keselman, 2010). For example, β -trimmed means are means calculated from ordered data with the desired β proportion of data removed from both tails of the distribution. Usually 20% (i.e., $\beta = 0.2$) trimming is recommended (Wilcox, 1995).

Some researchers have called for caution in the use of trimmed means because instead of testing the equality of the usual population means, the trimmed means method actually modifies the null hypothesis pertaining to the equality of the population trimmed means across repeated measures (e.g., Berkovits et al., 2000). On the other hand, Wilcox (1993) explored the robustness of the trimmed means method against non-normality and found that inferences based on trimmed means method could be more powerful than those based on the traditional means method. Wilcox (1997, 1998) further concluded that the results from the trimmed means method are more robust and may therefore be more accurate and replicable, which should be preferable over conventional statistics. This conclusion was also supported by Berkovits et al. (2000), who found that when the sphericity assumption was violated, trimmed mean methods were able to offer reasonable Type I error control.

Due to the limitations of the ANOVA-based methods exhibited in the previous studies, more endeavors were taken to develop new methods to deal with the violations of the assumptions of normality and sphericity. Structual means modeling is one of the alternatives that can be used for repeated measures design, based on which, robust means modeling (RMM) could be developed.

Introduction to Robust Means Modeling

Structured Means Modeling (SMM). Structured means modeling is a branch of structural equation modeling (SEM) that has become an ever-increasingly popular data analytic method across a wide variety of fields due to its versatility in modeling relations among both measured and latent variables. Growing out of, but more powerful and flexible than, multiple regression, SEM can be treated as a more general and flexible model subsuming methods such as canonical correlation, multiple regression, MANOVA, ANOVA, and *t*-tests, and is able to deal with complex situations involving, for example, interactions, measurement error, multilevel models, and correlated error terms.

Covariance is generally regarded as the basic statistic of SEM as the main goal of SEM is to understand patterns of covariances among a set of observed variables using a proposed model. If only covariances are analyzed, means of observed variables are irrelevant and the intercepts are typically omitted from the structural equations (Hancock, 1997). Some researchers, especially those who use ANOVA heavily in their analysis, therefore, are under the impression that SEM deals solely with covariances. However, this view is too limited as means can also be estimated and analyzed in SEM for both latent and observed variables (Kline, 2011) and this can be achieved with SMM. SMM aims to model variables' mean structure along with the covariance structure in order to facilitate inference regarding populations' underlying construct means (Hancock, 2001). SMM uses equations involving means along with the accompanying variable intercepts and these new equations constitute the mean structure, which is added to the model's basic covariance structure and estimated in addition to covariance structure. Therefore, SMM includes intercept parameters as well as variances, covariances, and path coefficients (Thompson & Green, 2006). In other words, by using SMM, variables' mean structure and covariance structure can be modeled simultaneously, and tests of hypotheses about means on a latent variable or observed variable, as well as the error covariance structure, can be achieved. This actually makes ANOVA a special case of SEM given that ANOVA is only concerned with means of observed variables.

To understand how SMM operates, assume latent variable ξ has p observed variables as indicators, $x_1,\ x_2,\ ...,\ x_p$ in vector $\mathbf{x}$; $\mathbf{X}$ values for a given individual may be expressed in a $p \times 1$ vector as

$$\mathbf{x} = \mathbf{T} + \mathbf{\Lambda}\xi + \mathbf{\delta}, \tag{10}$$

where $\mathbf{T}$ is a $p \times 1$ vector of intercept values, $\mathbf{\Lambda}$ is a $p \times 1$ vector of loadings relating ξ to its indicators, and $\mathbf{\delta}$ is a $p \times 1$ vector of random measurement errors. Thus, the expected values for the indicators can be derived as

$$\mathrm{E}[\mathbf{x}] = \mathbf{\mu} = \mathbf{T} + \mathbf{\Lambda}\kappa, \tag{11}$$

where $\mathbf{\mu}$ is the $p \times 1$ population mean vector for the observed variables, and κ is the mean of ξ. Assuming ξ is independent of the measurement errors, the model implied variance-covariance matrix can be derived as

$$\mathbf{\Sigma}\text{=}\mathbf{\Lambda\Phi\Lambda}'\text{+}\mathbf{\Theta}, \tag{12}$$

where $\mathbf{\Phi}$ is the variance for ξ and $\mathbf{\Theta}$ is the $p \times p$ error variance-covariance matrix.

For repeated measures design, two aspects of SMM are directly relevant. First, similar to SEM, SMM makes no assumptions about (co)variances and allows errors to correlate with each other if necessary, thereby circumventing the sphericity assumption altogether. Second, a special case of SMM is that there is no latent factor ξ, and only the observed variables are in the model. In this case, the original SMM with a latent variable can be simplified to a measured variable mean structure model

$$\mathbf{x} = \mathbf{T} + \mathbf{\delta}. \tag{13}$$

Accordingly, the expected values for the indicators can be derived as

$$\mathrm{E}[\mathbf{x}] = \boldsymbol{\mu} = \mathbf{T}, \tag{14}$$

and the model implied (co)variance can be derived as

$$\boldsymbol{\Sigma} = \boldsymbol{\Theta}. \tag{15}$$

Graphically, this model can be presented as a constant number of one, as well as the error terms, having direct bearing on the observed variables (See Figure 1).

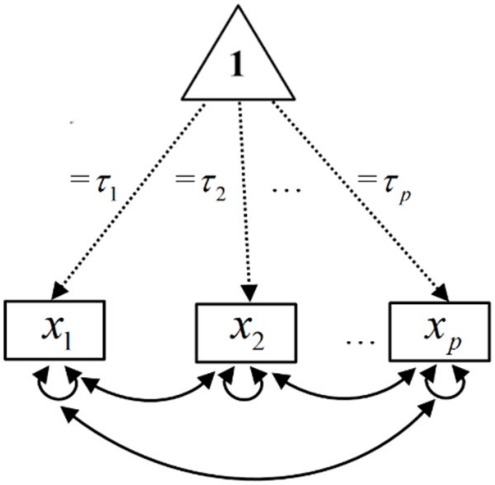

Figure 1: The path diagram of null hypothesis

The null hypothesis for the omnibus test for repeated measures thus becomes

$$\mathrm{H}_0: \quad \tau_1 = \tau_2 = \ldots = \tau_p. \tag{16}$$

To test the equality of the means/intercepts, a constraint can be imposed forcing intercepts equal across time points/measure occasions while allowing for covariances among variables/errors. The maximum likelihood (ML) estimator is the traditional method used in SMM estimation and it assumes that the observed variables are continuous and multivariate normal.

Robust Means Modeling (RMM). For SEM, obtaining parameter estimates (estimation) and evaluating the estimates' sampling variability as well as the behavior of test statistics (evaluation) are the two major statistical tasks (Satorra, 1990). The conclusions gained in the statistical analysis need to be justified by two types of assumptions: structural and distributional. Structural assumptions are related to the specifications of the models. Usually it is assumed that the model is correctly specified, which is fairly reasonable in the current study given the simplicity of the measured variable mean structure models. Distributional assumptions are mainly concerned with the form of the distribution of the observed variables. For example, it is typically assumed that the observed variables are normally distributed, but in reality this assumption does not hold in most cases (e.g., Micceri, 1989). Violation of assumptions may make some conclusions of precision of the estimators and the calculation of the level of significance invalid, or it may distort the analysis (Satorra, 1990). In the present study, it is assumed that the structural assumptions hold, while the issues related with violation of distributional assumption will be the focus.

In order to introduce and compare various estimators, a general condition of

SEM can be laid out as follows: For a one-way repeated measures design with p measured variables (or occasions), assume the sample mean vector $\overline{\mathbf{X}}$ and the $p \times p$ covariance matrix $\mathbf{S}$ are obtained from a random sample of size n from a population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma_0}$. This model yields $p^* = p(p+3)/2$ unique means, variances, and covariances. For the *i*th individual this random sample can be written as $\mathbf{X}_i = (x_{i1}, x_{i2}, ..., x_{ip})'$ for $i = 1, 2, ..., n$ which is obtained from $\mathbf{X} = (x_1, x_2, ..., x_p)'$. The general null hypothesis for SEM to be tested is H_0 : $\boldsymbol{\Sigma_0 = \Sigma(\theta)}$, where $\boldsymbol{\Sigma(\theta)}$ is a covariance matrix written as functions of q free model parameters in vector $\boldsymbol{\theta}$. For simplicity, for any vector of model parameter estimates $\hat{\boldsymbol{\theta}}$, the corresponding model-implied covariance matrix $\boldsymbol{\Sigma(\hat{\theta})}$ can be written as $\hat{\boldsymbol{\Sigma}}$. The purpose in the parameter estimation is to yield a vector of parameter estimates that minimize the function denoted as $F\mathbf{(S, \hat{\Sigma})}$ capturing the discrepancy between elements in $\hat{\boldsymbol{\Sigma}}$ and $\mathbf{S}$. Meanwhile, the goodness-of-fit test statistics are also formulated as a function of this discrepancy and are of the form $T = c(n-1)F$, where F is the minimum value of $F\mathbf{(S, \hat{\Sigma})}$, n is the sample size, and c is a scaling factor. When the null hypothesis is true, the asymptotic distribution of *T* is a χ^2 distribution with $p^* - q$ degrees of freedom (e.g., Fouladi, 2000). However, there are also some other test statistics proposed whose distribution is not approximated by a χ^2 distribution. For example, the distributions of two corrected test statistics proposed by Yuan and Bentler (1997, 1998, 1999) which will be addressed in later section are approximated by an *F* distribution.

If the maximum likelihood (ML) method is used for SMM estimation, the fit function to be minimized is

$$F_{ML} = \ln|\hat{\mathbf{\Sigma}}| + tr(\mathbf{S}\hat{\mathbf{\Sigma}}^{-1}) - \ln|\mathbf{S}| - p + [\overline{\mathbf{X}} - \hat{\boldsymbol{\mu}}]'\hat{\mathbf{\Sigma}}^{-1}[\overline{\mathbf{X}} - \hat{\boldsymbol{\mu}}], \tag{17}$$

and the test statistic associated with this fitting function is

$$T_{ML} = (n-1)F_{ML}. \tag{18}$$

In this case, the scaling factor *c* is 1. The asymptotic distribution of T_{ML} can be approximated by a χ^2 distribution with $p^* - q$ degrees of freedom when null hypothesis H_0 is true and when data are multivariate normal.

The robustness of ML to violations of distributional assumptions has been the focus of a number of studies. To start, it has been observed that under certain conditions of distribution violations, ML behaves reasonably well. For example, Chou, Bentler, and Satorra (1991) found that the ML test statistics and standard errors were quite robust to the violation of the normality assumption when data were either symmetric and platykurtic, or asymmetric with zero kurtosis. Across a broader class of nonnormal distributions, however, ML test statistics and standard errors have been found to be biased even though the ML estimates are fairly consistent (e.g., Chou et al., 1991; Curran, West, & Finch, 1996; Finney & DiStefano, 2013; Fouladi, 2000; Hu, Bentler, & Kano, 1992; Nevitt & Hancock, 2001, 2004; Powell & Schafer, 2001; Yuan & Bentler, 1997, 1998, 1999), with the direction of bias appearing to be dependent on whether the data are leptokurtic or platykurtic (e.g., Browne, 1984; Chou et al., 1991; Finney & Stefano, 2013). Therefore, it is usually recommended that robust estimators be used when the normality assumption does not hold. The use of robust estimators in SMM then leads to what is termed here robust means modeling (RMM) (Fan & Hancock, 2012).

Application of RMM to one-way repeated measures design.

Originally, robust estimators were applied to traditional SEM where means were unrestricted (i.e., ignorable) and covariance structures were the focus. But with SMM, both means and covariance structures need to be taken into account. In order to employ conventional SEM software for covariance structures to analyze both means and covariance, Satorra (1992) and Browne and Arminger (1995) suggested modeling mean and covariance structure simultaneously by replacing the sample covariance structure $\mathbf{S}$ with an augmented moment matrix

$$\mathbf{S}^* = \begin{pmatrix} \mathbf{S} + \overline{\mathbf{X}}\overline{\mathbf{X}}' & \overline{\mathbf{X}} \\ \overline{\mathbf{X}}' & 1 \end{pmatrix}. \tag{19}$$

Correspondingly, the population augmented moment matrix can be specified as

$$\mathbf{\Sigma}^* = \begin{pmatrix} \mathbf{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}' & \boldsymbol{\mu} \\ \boldsymbol{\mu}' & 1 \end{pmatrix}. \qquad (20)$$

Then the newly augmented matrices can be used to analyze the means and covariance structure simultaneously.

The robust estimators relevant to the current study generally fall into two categories: (a) using ML for parameter estimates but adjusting the test statistic; (b) and abandoning ML for distribution-free estimators to account for the non-normality.

ML-Based Adjusted Test Statistics. Both the Satorra-Bentler scaled χ^2 test statistic and adjusted χ^2 test statistic were developed based on the ML test statistics.

Satorra-Bentler scaled Chi-square test statistic (SB1). It is now generally accepted that whether or not the normality assumption holds, ML parameter estimates are consistent. However, this robustness does not apply to test statistics obtained under non-normality (e.g., Satorra, 1992). Therefore, Satorra and Bentler (1988, 1994) developed a scaled χ^2 test statistic (T_{SB1}) to adjust the ML-based χ^2 by taking into account the observed data's distributional characteristics so that the test statistic's distribution behavior should more closely approximate the theoretical χ^2 reference distribution.

Let $vech(\cdot)$ be an operator that transforms a symmetric matrix into a vector by stacking the columns of the non-redundant (diagonal and lower triangular) elements of the matrix so that $\mathbf{s} = vech(\mathbf{S}^*)$ and $\hat{\boldsymbol{\sigma}} = vech(\hat{\mathbf{\Sigma}}^*)$. Let $\dot{\hat{\boldsymbol{\sigma}}}$ be the $p^* \times q$ matrix of partial derivatives of the $p*$ elements in $\hat{\boldsymbol{\sigma}}$ evaluated at the vector of final model parameter estimates $\hat{\boldsymbol{\theta}}$ (i.e., Jacobian matrix), $\mathbf{W}$ be the normal theory ML weight matrix at the minimum of F_{ML} obtained by the function of $\hat{\boldsymbol{\Sigma}}^{-1} \otimes \hat{\boldsymbol{\Sigma}}^{-1}$ ($\otimes$ denotes the Kronecker product) , and

$$\hat{\mathbf{U}} = \mathbf{W} - \mathbf{W}\dot{\hat{\boldsymbol{\sigma}}}(\dot{\hat{\boldsymbol{\sigma}}}'\mathbf{W}\dot{\hat{\boldsymbol{\sigma}}})^{-1}\dot{\hat{\boldsymbol{\sigma}}}'\mathbf{W}. \tag{21}$$

The test statistic can be obtained by

$$T_{SB1} = [(p*-q)/tr(\hat{\mathbf{U}}\hat{\boldsymbol{\Gamma}})]T_{ML} \tag{22}$$

where $\boldsymbol{\Gamma}$ is the population matrix of $\mathbf{s}$ which is a symmetric $p*\times p*$ fourth-order moment weight matrix. $\hat{\boldsymbol{\Gamma}}$ is a consistent estimator of $\boldsymbol{\Gamma}$ and is obtained when the population moments are replaced by the corresponding sample moments. Let $\mathbf{Y}_i = vech[(\mathbf{X_i} - \overline{\mathbf{X}})(\mathbf{X_i} - \overline{\mathbf{X}})']$ and $\mathbf{S_Y}$ be the corresponding sample covariance matrix of $\mathbf{Y}_i$. Then an estimator for $\boldsymbol{\Gamma}$ is $\mathbf{S_Y}$. Thus, the test statistic can be expressed as

$$T_{SB1} = [(p*-q)/tr(\hat{\mathbf{U}}\mathbf{S_Y})]T_{ML}, \tag{23}$$

which is evaluated as a χ^2 distribution with $p*-q$ degrees of freedom when H_0 is true. In this way, the mean of the sampling distribution of T_{SB1} is adjusted closer to the expected mean under the model (Bentler, 1996).

The advantage of this adjustment is that it takes the covariance matrix of the sample (co)variances, which captures the degree of non-normality of the sample data, into consideration and therefore performs better in controlling Type I error rates under a wide variety of distribution conditions, including normal data (Chou et al., 1991; Curran et al., 1996; Fouladi, 2000; Hu et al., 1992; Nevitt & Hancock, 2004). But there are also some cautions against this statistic. Chou et al. (1991) pointed out that because the model evaluated in their study was simple and the non-normality conditions investigated were limited, the performance of the S-B scaled test statistic with more complex models and more conditions of non-normality needed further investigation. Yuan and Bentler (1998) claimed that this statistic worked well when distribution assumptions were violated, but its asymptotic distribution was generally unknown. Hu et al. (1992) and Bentler and Yuan (1999) suggested using the S-B scaled test statistic in medium to large samples ($n > 120$), but not in small samples.

Satorra-Bentler adjusted Chi-square test statistic (SB2). Besides adjusting the mean of the sampling distribution, Satorra and Bentler (1988, 1994) also proposed adjusting both the mean and the variance of the statistic to better approximate a χ^2 distribution. This statistic is then called adjusted χ^2 statistic and can be obtained as

$$T_{SB2} = [d'/tr(\widehat{\mathbf{U}}\widehat{\mathbf{\Gamma}})]T_{ML} , \tag{24}$$

where

$$d' = [tr(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}})]^2 / tr[(\hat{\mathbf{U}}\hat{\mathbf{\Gamma}})^2]. \tag{25}$$

The studies on the behavior of the adjusted χ^2 statistic were limited due to lack of available software. Based on the result from their simulation study, Satorra and Bentler (1994) concluded that this adjusted χ^2 statistic performed “remarkably well” (p. 413). However, the findings in Nevitt and Hancock (2004) did not favor this adjusted statistic and indicated comparably low power and attenuated Type I error rates across most conditions in their study. Based on these findings, Nevitt and Hancock decided not to recommend this method. Fouladi (2000) also indicated that this adjusted statistic was less powerful than the S-B scaled χ^2 test statistic.

Distribution-free Methods.

Browne’s asymptotic distribution-free (ADF) test statistic. Distribution-free methods form another large category of robust estimators. Among them, Asymptotic Distribution-free (ADF) estimation, which is a generalized least squares analysis, was introduced by Browne (1984) to remedy the problems associated with distribution misspecifications since ADF makes minimal or no assumptions about the population distribution of the observed variables. The weight matrix it uses is based on the inverse of the matrix formed by the sample fourth-order moments and sample covariances (Yuan & Bentler, 1997).

Briefly, the discrepancy function to be minimized is

$$F_{ADF} = (\mathbf{s} - \hat{\boldsymbol{\sigma}})' \hat{\boldsymbol{\Gamma}}^{-1} (\mathbf{s} - \hat{\boldsymbol{\sigma}}) \,. \tag{26}$$

The resulting test statistic under the null hypothesis is

$$T_{ADF} = (n-1) F_{ADF} \tag{27}$$

which is referred to as a χ^2 distribution with $p^* - q$ degrees of freedom when H_0 is true.

This estimator has received a great deal of attention ever since it was proposed. Numerous simulation studies (e.g., Anderson & Gerbing, 1984; Chou et al., 1991; Curran et al., 1996; Hoogland & Boomsma, 1998; Hu et al., 1992; Muthen, 1989; Muthen & Kaplan, 1992; Nevitt & Hancock, 2004) confirmed that the ADF statistic was insensitive to the violation of the normality assumption and performed very well when the sample size was very large (e.g., larger than 2500). Unfortunately, however, if the sample size was small to moderate or if the model was complex, it behaved poorly and yielded distorted conclusions about the adequacy of the model.

The fact that ADF uses the inverse of the fourth-order moments of the measured variables to compute parameter estimates, standard errors, as well as test statistics might be the reason why ADF is highly unstable in small to moderate sample sizes (Satorra & Bentler, 1994). Based on the findings of the previous studies and their own meta-analysis, Powell and Schafer (2001) concluded that the ADF test statistic should not be recommended.

Yuan and Bentler adjusted ADF I (YBADF). Because the ADF test statistic has been found to over-reject correct models in small to moderate samples, Yuan and Bentler (1997, 1999) proposed a corrected test statistic which has the same asymptotic distribution as the ADF test statistic

$$T_{YB(ADF)} = T_{ADF} / [1 + (n-1)^{-1} T_{ADF}] \tag{28}$$

and is also referred to as a χ^2 distribution with $p^* - q$ degrees of freedom when H_0 is true. It can be seen that $T_{YB(ADF)}$ gets closer to T_{ADF} when the sample size increases, but it eases the inflation problem T_{ADF} has at smaller samples sizes. Generally, it performs very well at the smaller sample sizes, but this statistic tends to overcorrect the inflation of T_{ADF} (Bentler, 2006, EQS Manual) and it is consistently conservative (Fouladi, 2000).

Yuan and Bentler adjusted ADF II (FADF). Intrigued by the similarity between the quadratic form of T_{ADF} and Hotelling's T^2 statistic, Yuan and Bentler (1999) proposed to scale T_{ADF} so that T_{ADF} can be approximated by an *F* distribution as

$$T_{F(ADF)} = T_{ADF}[n - (p^* - q)] / [(n-1)(p^* - q)] \tag{29}$$

with degrees of freedom $p^* - q$ and $N - (p^* - q)$ for numerator and denominator, respectively, when H_0 is true. $T_{F(ADF)}$ is asymptotically equivalent to T_{ADF}.

Yuan and Bentler (1999) used Monte Carlo simulation to compare the performance of T_{ADF}, $T_{YB(ADF)}$, and $T_{F(ADF)}$ under various distributional forms and sample sizes and concluded that both $T_{YB(ADF)}$ and $T_{F(ADF)}$ behaved better than T_{ADF} and yielded adequate power. They also contended that between $T_{YB(ADF)}$ and $T_{F(ADF)}$, because *F* distribution approximations are much better than the large sample theory based χ^2 approximations, it is understandable why $T_{F(ADF)}$ had improved performance relative to $T_{YB(ADF)}$.

Residual-based ADF test statistic (RES) and its corrections (YBRES, FRES). Besides the aforementioned ADF test statistics, Browne (1984) proposed a residual-based ADF test statistic as

$$T_{RES} = (n-1)\hat{\mathbf{e}}'[\mathbf{S}_{\mathbf{Y}}^{\mathbf{-1}} - \mathbf{S}_{\mathbf{Y}}^{\mathbf{-1}}\hat{\dot{\boldsymbol{\sigma}}}(\hat{\dot{\boldsymbol{\sigma}}}'\mathbf{S}_{\mathbf{Y}}^{\mathbf{-1}}\hat{\dot{\boldsymbol{\sigma}}}_c)^{-1}\hat{\dot{\boldsymbol{\sigma}}}'\mathbf{S}_{\mathbf{Y}}^{\mathbf{-1}}]\hat{\mathbf{e}}, \qquad (30)$$

where $\hat{\mathbf{e}} = \mathbf{s} - \hat{\boldsymbol{\sigma}}$ is a $p^* \times 1$ column vector of residual variances and covariances. T_{RES} can be computed more easily than the ADF test statistics, but as Yuan and Bentler pointed out, similar to the ADF test statistic, T_{RES} is also inflated and at the same time sensitive to model degrees of freedom rather than number of parameters in the model, thus rejecting correct models far too frequently when models are large while sample sizes are small to moderate. In order to solve the issues related to T_{RES}, Yuan and Bentler (1998) proposed corrections to T_{RES} that are similar to the corrections they proposed to adjust T_{ADF}. They proposed replacing the fourth-order moments $\mathbf{S_Y}$ by a new estimate of $\mathbf{\Gamma}$ which has good finite sample properties. The resulting new statistic is

$$T_{YB(RES)} = T_{RES} / [1 + nT_{RES} / (n-1)^2], \qquad (31)$$

following a χ^2 distribution with $p*-q$ degrees of freedom when H_0 is true. $T_{YB(RES)}$ is asymptotically equivalent to T_{RES}. Also, $T_{YB(RES)}$ is numerically smaller than T_{RES} and, therefore, it is expected that the inflated rejection rate of T_{RES} for models with smaller sample sizes can be lessened.

Yuan and Bentler (1998) also noted that the quadratic form of $T_{YB(RES)}$ resembles that of Hotelling's T^2, therefore proposing to rescale T_{RES} to an *F* distribution as

$$T_{F(RES)} = T_{RES}[n-(p*-q)]/[(n-1)(p*-q)], \qquad (32)$$

with degrees of freedom $p*-q$ and $N-(p*-q)$ for numerator and denominator respectively. The distribution of $T_{F(RES)}$ is also asymptotically equivalent to T_{RES}.

Yuan and Bentler (1998) and Bentler and Yuan (1999) employed Monte Carlo simulations to compare the performances of T_{RES}, $T_{YB(RES)}$, and $T_{F(RES)}$ under various distributional and sample size conditions and concluded that both $T_{YB(RES)}$ and $T_{F(RES)}$ were able to correct the over-rejection of T_{RES} for correct models in finite samples. Among the three, $T_{YB(RES)}$ behaved most stably across different conditions but similar to its ADF analog $T_{YB(ADF)}$; it also tended to overcorrect the inflation of T_{RES}. Of all test statistics, $T_{F(RES)}$ performed best over the range of the conditions. Based on their findings, Bentler and Yuan recommended that $T_{F(RES)}$ should be the first choice for practitioners if the sample size is smaller than the number of nonduplicated elements of the sample covariance (i.e., $n < p^*$).

Nevitt and Hancock (2004) were among the few studies that compared all the aforementioned nine test statistics in one study. First of all, their study supported that T_{ML} was not robust to departures to multivariate normality. The comparison between T_{SB1} and T_{SB2} revealed that even though T_{SB1} did not perform very well when the ratio of sample size to the number of free parameters was $n : q < 10:1$, generally speaking, T_{SB1} outperformed T_{SB2}. Therefore, they recommended T_{SB1} in applied modeling situations. This finding is in line with that discovered by Bentler and Yuan (1999) where they found that T_{SB1} broke down under the smallest sample size conditions. At the same time, similar conclusion regarding the performance of T_{ADF} and T_{RES} were obtained that these two statistics were not useful for realistic sample sizes. $T_{F(ADF)}$, $T_{YB(ADF)}$, and $T_{YB(RES)}$ were able to control Type I error rates under some conditions but failed under others. When they failed, the Type I error rates became inflated under some conditions but attenuated under others. $T_{F(ADF)}$ and $T_{F(RES)}$ performed differentially, showing similar behaviors under some conditions but diverging from one another under other conditions. Under certain conditions, $T_{F(RES)}$ showed better performance than $T_{F(ADF)}$. But $T_{F(ADF)}$, $T_{YB(ADF)}$, $T_{YB(RES)}$, and $T_{F(RES)}$ unanimously behaved poorly at their respective sample size lower bounds.

Fan and Hancock (2012) was among the few studies that compared somewhat different sets of ANOVA-based methods (*F* test, Welch's test, the Brown-Forsythe test, James' second-order test, and the Alexander-Govern test) and RMM methods (ADF, SB1,YBADF, FADF, and Bartlett's correction to the ML) applied to between-subjects designs. The study found that RMM was robust in terms of controlling Type I error rates across range of distribution shapes and sample sizes and variance conditions and outperformed ANOVA-based methods for between-subjects designs. Among the RMM methods, both ADF- and ML-based statistics performed well in terms of controlling Type I error rates and power and FADF and YBADF were singled out among RMM methods. Based on their findings, they called for future study on repeated measures designs, which was the focus of the current study.

Chapter 3: Methods

Previous studies showed the potential advantages and disadvantages of ANOVA-based methods as well as RMM methods. The current study then compares all the aforementioned methods except SB2 using a simulation study specified as below.

Test Statistics Examined

The current RMM study not only examines the behavior of the maximum likelihood estimator T_{ML} and the RMM test statistics (i.e, T_{SB1}, T_{ADF}, $T_{F(ADF)}$, $T_{YB(ADF)}$, T_{RES}, $T_{YB(RES)}$, $T_{F(RES)}$), but also compares their performance with other methods proposed in the field of ANOVA to determine whether RMM estimators outperform ANOVA-based methods or vice versa in terms of Type I error rates and power in a one-way repeated measures design. To be more specific, ANOVA-based methods to be examined in this study include the traditional *F* test (*F*), the Geisser-Greenhouse adjusted *F* test (GG), Box's adjusted *F* test (Box), the Huynh-Feldt adjusted *F* test (HF), the β -trimmed method using β = 0.2 (TR), and the one-sample multivariate T^2 test (FM).

Variables Investigated

A Monte Carlo simulation was designed to evaluate Type I error robustness for each aforementioned method under the following fully crossed conditions in a one-way repeated measures design: number of levels (*K*), sample size (*N*), degree of non-sphericity (ε), and degree of non-normality.

Number of levels *(K)*. Different researchers used different numbers of levels for different reasons given that there were no rules of thumb for the number of levels (*K*) used. For example, Quintana and Maxwell (1994) used 5 and 13 because (a) the univariate approach is most valuable when the number of levels is large relative to sample size and (b) so previous research could be paralleled. A series of studies done by Keselman and his colleagues (e.g., Algina, & Keselman, 1997; Keselman, & Keselman, 1988, 1990; Keselman, Keselman, & Shaffer, 1991; Keselman, Kowalchuk, Algina, Lix, & Wilcox, 2000; Lix, Algina, & Keselman, 2003) have used 4 and/or 8 levels.

Therefore, for the current study, the number of levels was set at $K = 4,\ 8$ to reflect those used in past simulation designs that investigated repeated measures designs (see, e.g., Algina, & Keselman, 1997; Berkovits et al., 2000; Keselman, Carriere, & Lix, 1993; Keselman et al., 1991; Keselman, & Keselman, 1988, 1990; Lix, Algina, & Keselman, 2003; Keselman, Kowalchuk, Algina, Lix, & Wilcox, 2000) so that the results are more comparable between the current study and these previous studies.

Sample size (*N*). One of the benefits of repeated measures designs is that they typically require smaller number of subjects in the study relative to between-subjects designs. Quintana and Maxwell (1994) suggested using smaller sample size (e.g., 5, 10, or 20) per level because the univariate approach outperformed the multivariate approach in these situations.

Maxwell and Delaney (1990), on the other hand, suggested a “rough rule of thumb” (p. 603) that sample size should be at least larger than *K*+10 if the multivariate approach is to be used. Based on this guideline, Algina and Keselman (1997) set the sample size as *K*+5, *K*+10, *K*+15, *K*+20, *K*+30, and *K*+40 to investigate the power of univariate and multivariate approaches.

In this study, the behavior of robust estimators from SEM would also be examined. SEM was often regarded as large sample technique because increases in sample size would increase the likelihood of proper model convergence, enhance the accuracy of parameter and standard errors estimates, and improve statistical power. Lei and Lomax (2005) recommended using sample sizes of 100 or more for accurate parameter estimates.

Combining all the factors mentioned above, the sample size of this study will be fixed at $N = 15$, 30, 60, 100, and 200 to reflect various sample size conditions.

Degree of violation of sphericity assumption. The population sphericity index (ε) was set at $\varepsilon = 1$, $\varepsilon = 0.96$, $\varepsilon = 0.75$, and $\varepsilon = 0.48$, indicating conditions ranging from perfect sphericity to a severe violation of sphericity. Covariance matrices for various sphericity conditions were obtained from Keselman and Keselman (1990).

$\varepsilon = 1$ was included as a baseline because $\varepsilon = 1$ indicates that the sphericity assumption was satisfied. $\varepsilon = 0.96$ was picked because covariance matrices with $\varepsilon = 0.96$ exhibit small departures from sphericity. $\varepsilon = 0.75$ represents moderate departures from sphericity, which echo the characteristics of data found in the field of educational and behavioral research (Algina & Keselman, 1997). $\varepsilon = 0.48$ was also included in this study as it corresponds "roughly with the lower limits of values often reported in the behavior literature" (Quintana & Maxwell, 1994, p. 62).

Table 1

Covariance matrices for *k*=4 (Keselman & Keselman, 1990)

ε=1	10.0	5.0	5.0	5.0
		10.0	5.0	5.0
			10.0	5.0
				10.0
ε=0.96	12.0	6.0	5.0	5.0
		10.0	5.0	4.0
			10.0	5.0
				8.0
ε=0.75	18.0	8.0	6.0	4.0
		8.0	5.0	4.0
			7.0	3.0
				7.0
ε=0.48	22.3	10.8	6.5	1.9
		8.3	5.2	3.1
			4.7	2.5
				4.7

Table 1 and Table 2 listed the population covariance matrices for the aforementioned levels of sphericity, which were the same as those in Keselman and Keselman (1990) so that the results were more comparable between the current study and previous studies. These matrices were used to guide data generation. Both of the tables used a constant diagonal value of 10 and a constant off-diagonal value of 5 for the covariance matrices to depict the differenct levels of sphericity (Keselman & Keselman, 1990).

Table 2

Covariance matrices for *k*=8 (Keselman & Keselman, 1990)

ε=1	10.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
		10.0	5.0	5.0	5.0	5.0	5.0	5.0
			10.0	5.0	5.0	5.0	5.0	5.0
				10.0	5.0	5.0	5.0	5.0
					10.0	5.0	5.0	5.0
						10.0	5.0	5.0
							10.0	5.0
								10.0
ε=0.96	11.0	7.0	6.0	6.0	6.0	5.0	5.0	5.0
		11.0	6.0	5.0	5.0	5.0	5.0	5.0
			10.0	5.0	5.0	5.0	5.0	5.0
				10.0	5.0	5.0	5.0	4.0
					10.0	5.0	5.0	4.0
						10.0	4.0	4.0
							9.0	4.0
								9.0
ε=0.75	18.0	8.0	7.0	7.0	6.0	5.0	5.0	5.0
		12.0	8.0	7.0	6.0	5.0	5.0	2.0
			10.0	6.0	6.0	5.0	5.0	2.0
				10.0	5.0	5.0	4.0	4.0
					9.0	5.0	5.0	3.0
						8.0	4.0	4.0
							7.0	1.0
								6.0
ε=0.48	26.1	10.7	10.2	9.9	9.3	6.0	5.9	2.0
		15.8	9.3	8.1	7.9	4.2	3.4	-0.4
			10.8	7.0	6.0	5.5	3.2	2.2
				9.8	5.2	5.6	3.4	2.1
					6.0	3.4	2.6	1.1
						4.7	2.4	2.2

	4.0	1.6
		2.8

Degree of non-normality. As previously mentioned, the traditional *F* test as well as ML estimation rely on the assumption of normality even though ML estimation tends to be robust to normality violations in terms of bias of parameter estimates. Different degrees of non-normality can be represented by different skew and kurtosis values. Lei and Lomax (2005) suggested that the skew and kurtosis values be selected between -2.0 and +3.5. If both values are less than 1.0, the distribution can be regarded as slightly nonnormal, between 1.0 and about 2.3 as moderately nonnormal, and beyond 2.3 as severely nonnormal. Kline (2011) defined non-normality by separating skewness and kurtosis. He suggested that absolute skew values greater than 3 indicate extreme skew and absolute kurtosis values ranging from 8.0 to over 20.0 indicate extreme kurtosis. He further pointed out that absolute kurtosis values greater than 20 may indicate serious problem. Based on Kline's perspective, the skew and kurtosis values proposed by Lei and Lomax can be regarded as fairly mild.

Based on the above, four levels of non-normality were assessed: (a) normal (skewness = 0, kurtosis = 0); (b) moderately non-normal (skewness = 2, kurtosis = 7); (c) severely non-normal (skewness = 3, kurtosis = 21); and (d) elliptical (skewness = 0, kurtosis = 7). Though it is not possible for four conditions to cover the full range of non-normal distributions, these represent a range of conditions encountered in methodological and applied research. In order to generate data for these combined levels of skewness and kurtosis, Fleishman's (1978) polynomial transformation was used. Vale and Maurelli (1983) outlined the procedure to generate the intermediate correlation matrix that accommodated the effect of nonnormalizing on the correlation since non-normal data had intercorrelatons different from the normal data. Then this intermediate correlaton matrix could be applied to the procedures of the data generation of multivariate normal random numbers using

$$Y = -c_2 + c_1Z + c_2Z^2 + c_3Z^3 \tag{33}$$

where Y was the non-normal data generated; c_1, c_2, and c_3 are the coeffieints determined by the intermediae correlation matrix. Each data set was generated by sampling from a population with the skewness and kurtosis as well as other properties listed above.

Design and Execution

The decision regarding the number of replications should be made based on the purpose of the study, the desire to reduce the variance of estimated parameters and the need for adequate power (Harwell, Stone, Hsu, & Kirisci, 1996). If the behavior of standard errors is of interest, more replications may be needed. The number of replications has direct impact on the precision of estimated parameters and more replications produce higher precision in parameter estimates (Bandalos, 2006). Most of the previous studies explicitly indicated the number of replications but provided no justification and the number of replications used ranged from 500 (e.g., Yuan & Bentler, 1998) to 10000 (e.g., Algina & Keselman, 1997). Some studies used 1000 replications (e.g., Fan & Hancock, 2012; Keselman, Algina, Wilcox, & Kowalchuk, 2000; Yuan & Bentler, 1999). Others used 2000 replications (e.g., Nevitt, & Hancock, 2004). For each of the 2× 5 ×4×4 cells of this design, 2000 simulated data sets or replications as the average of the previous studies were generated in SAS (2011). The test statistics for ANOVA-based methods were all calculated and analyzed in SAS. The test statistics for the ML estimator T_{ML} and RMM estimators were obtained from EQS 6.2.

Type I error rate. The rate of false rejection (i.e., Type I error rate) was employed to define test statistic robustness. For each cell of the design, all tests were conducted at $\alpha = 0.05$ level. Type I error robustness were evaluated using Bradley's liberal criterion (Bradley, 1978) where the test's empirical Type I error rate $\hat{\alpha}$ must fall in the interval $.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$ to be considered robust. Therefore, for α =.05, the robustness interval corresponding to Bradley's liberal criterion was $.025 \leq \hat{\alpha} \leq .075$. For all methods under all conditions, the empirical Type I error rate $\hat{\alpha}$ was computed as the number of false rejections out of 2,000 replications or $r = (number\ of\ rejections) / 2000$ and an estimated standard error of $\hat{\alpha}$ could be determined as $SE = [r(1-r)/2000]^{\frac{1}{2}}$. Thus, when $\hat{\alpha}$ =.05, SE = .0049, which is less than 10% of the estimated value.

Power Analysis. In order to understand fully the performance of all these methods, the power of the test statistics investigated was also examined in addition to Type I error rate analysis. Power analysis was carried out based on the results obtained from the Type I error rate analysis. Under Bradley's (1978) liberal criterion, test statistics that yielded empirical Type I error rates between 0.025 and 0.075 were regarded as robust and hence eligible for comparison in a power analysis. Test statistics under some study conditions that maintained Type I error rate beyond the upper bound of 7.5% were removed from power analyses under those conditions. The rationale for this elimination is that a liberal Type I error rate indicates inflated test statistics, which in turn will lead to inflated power estimates which are not comparable with power estimates from the test statistics producing reasonable Type I error rate. On the other hand, test statistics under some study conditions that yielded empirical Type I error rates below the lower bound of 2.5% were retained for power analyses in that these test statistics could potentially maintain acceptable power under those conditions. The results obtained for Type I error rates showed that out of 160 cells (2×4×5×4) for the β-trimmed method using β = 0.2, only four cells showed the Type I error rates between 2.5% and 7.5%. Therefore, the β-trimmed method using β=0.2 was removed from power analysis. Even though 75% of the Type I error rates obtained for the Geisser-Greenhouse adjusted *F* test for *k*=4 (60 cells out of

80 cells, all for sphericity levels over 0.48) and 100% of the Type I error rates for *k*=8 were below 2.5%, this method was still kept for power analysis. For the rest of the methods where sporadic cells provided Type I error rates that were beyond 7.5%, power analysis was carried out for all cells but the results for those cells whose Type I error rates were above 7.5% were removed from further analysis.

For those test statistics that entered the power analysis, a new series of data was generated using the same sample sizes, distributions, (co)variances, and sphericity levels as the Type I error analysis. For each of the cells, 1000 simulated data sets (replications) were generated in SAS (2011). In order to create the nonnull condition, the first group's means was moved by an amount that induced a difference between the first and second group equivalent to a specific Cohen's *d*. In order to determine a target Cohen's *d*, a pilot study was done with 500 replications. It was found out that when the Cohen's *d* was larger than 0.2, the results of power for the large sample sizes ($n > 60$) were very close to or equaled 100% (i.e., 1.00), and thus were not useful for discriminating the performance of different methods. When the Cohen's *d* was 0.1 and 0.2, the power levels at all sample sizes became distinguishable, with Cohen's *d* of 0.1 predictably yielding generally lower power than the Cohen's *d* of 0.2. Therefore, in this study, Cohen's *d* of 0.2 was used to guide the generation of the data.

Chapter 4: Results

Non-convergence

In the Type I error portion of the study, rates of non-convergence of all test statistics were tracked for all cells. The results were summarized in Table 3 and Table 4. Overall, no non-convergence occurred for ANOVA-based methods examined in this study, although some SEM-based methods did fail to converge in some cases. Among all SEM-based methods, T_{ML} and T_{SB1} did not encounter any non-convergence across all conditions. Because the test statistics $T_{F(ADF)}$ and $T_{YB(ADF)}$ are derived based on T_{ADF}, then $T_{F(ADF)}$ and $T_{YB(ADF)}$ should produce the same non-convergence rates as T_{ADF}. Similarly, because the test statistics $T_{YB(RES)}$ and $T_{F(RES)}$ are derived from T_{RES}, they should behave the same with respect to convergence.

Table 3 shows the rates of non-convergence when k =4 and n =15 and convergence existed for all methods at all higher sample sizes in this study. It can be seen that non-convergence only occurred when n =15 for distribution-free test statistics ($T_{F(ADF)}$, $T_{YB(ADF)}$, T_{ADF}) across all distributions and sphericity levels specified in this study. Generally speaking, the rates of non-convergence increased when the violation of sphericity assumption becames more severe (i.e., when ε gets smaller) across all distributions, with the exceptions of moderately non-normal distribution (skewness = 2, kurtosis = 7) and elliptical distribution (skewness = 0, kurtosis = 7). The rates of non-convergence also increased when the violation of normality assumption became more severe across all sphericity levels. The highest rate of non-convergence (42.0%) occurred when the distribution was severely non-normal (skewness = 3, kurtosis = 21) and population covariance matrices departed from sphericity most severely (ε = 0.48); conversely, as expected, the lowest rate of non-convergence (3.1%) occurred when the distribution was normal and the sphericity assumption was satisfied or is violated to the least extent (ε = 0.96).

Table 3

Rates of non-convergence (%) for *k*=4, *n* =15

ε	skew=0, kurt=0	skew=0, kurt=7	skew=2, kurt=7	skew=3, kurt=21
	ADF	ADF	ADF	ADF
1	3.4	6.4	9.2	18.4
0.96	3.1	5.2	8.3	19.7
0.75	3.8	7.9	7.5	24.0
0.48	6.6	17.3	20.8	42.0

When *k*=8, Table 4 shows that non-convergence only occurred for all distribution-free test statistics ($T_{F(ADF)}, T_{YB(ADF)}, T_{ADF}$) and all residual-based ADF test statistics (T_{RES}, $T_{YB(RES)}$, $T_{F(RES)}$) when sample sizes were 15 and 30 across all distributions and sphericity levels specified in this study. No distribution-free test statistics converged when sample sizes were 15 and 30 across all distribution conditions and all sphericity levels, which indicated that these test statistics failed to converge when the sample size was small. For the residual-based ADF test statistics, generally speaking, the rates of non-convergence decreased when the sample size got larger. When the sample size was 15, there was a general trend for the rates of convergence to increase when the departure from normality became greater, but this trend was not very consistent across distribution conditions. For example, the rates of non-convergence for severely non-normal distributions and moderately non-normal distributions were greater than those for normal distributions and elliptical distributions. But the rates of non-convergence for severely non-normal distributions (20.5%, 20.7%, 18.8%, and 16.0%) were smaller than those for moderately non-normal distributions across all sphericity levels (22.2%, 22.1%, 20.0%, and 18.2%). When sample size was 30, there seems to have been no obvious influence by either the distribution conditions or sphericity levels on the rate of non-convergence for each sample size condition. But the non-convergence rates when *n* = 30 were much smaller than those when *n* = 15.

Table 4

Rates of non-convergence (%) for ADF and RES and their corrections with *k*=8

		skew=0, kurt=0		skew=0, kurt=7		skew=2, kurt=7		skew=3, kurt=21	
	ε	ADF	RES	ADF	RES	ADF	RES	ADF	RES
n=15	1	100	15.5	100	17.2	100	22.2	100	20.5
	0.96	100	16.9	100	16.9	100	22.1	100	20.7
	0.75	100	16.5	100	15.3	100	20	100	18.8
	0.48	100	16.1	100	14.7	100	18.2	100	16.0
n=30	1	100	15.4	100	8.9	100	12.2	100	11.0
	0.96	100	12.5	100	9.3	100	12.7	100	9.0
	0.75	100	13.2	100	10.4	100	11.7	100	8.6
	0.48	100	12.8	100	9.2	100	8.4	100	6.4

Type I Errors rates

There are four types of distributions under investigation in this study: normal, elliptical, moderately non-normal, and severely non-normal. Tables 5, 7, 9, and 11 present the results of the Type I error rates under these four types of distributions when the number of levels *k* equals 4. Tables 6, 8, 10, and 12 present the results of the Type I error rates under the four types of distributions when the number of levels *k* equals 8. Among these eight tables, Table 5 and Table 6 present the results of the Type I error rates for the normal distribution, while Table 7 and Table 8 present the results of the Type I error rates for elliptical distribution under sphericity levels of 1, .96, .75, .48 and sample sizes of 15, 30, 60, 100, 200. Table 9 and Table 10 present the results for the moderately non-normal distribution, while Table 11 and Table 12 present the results for severely the non-normal distribution under four sphericity levels (1, .96, .75, .48) and five sample sizes (15, 30, 60, 100, 200).

Table 5: Empirical Type I Error Rates (%) for Normal Distribution with *k*=4

	ε	*F*	Box	HF	GG	FM	TR	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.0	4.6	3.6	4.4	0.9	4.9	6.7	6.2	13.3	1.2	4.7	6.2	11.8	0.8	4.1
	1.0	4.9	3.8	4.5	0.5	4.0	10.3	5.0	12.4	1.3	4.0	5.0	10.2	0.8	3.4
	0.8	6.6	4.4	5.7	1.4	5.7	11.5	6.8	14.2	1.7	5.4	6.8	13.2	1.1	4.6
	0.5	9.2	5.2	5.6	3.1	4.9	15.3	6.4	14.9	1.2	5.0	6.4	12.7	0.6	4.3
n=30	1.0	5.4	4.3	5.3	1.0	4.8	8.3	5.5	8.7	3.2	4.8	5.5	7.9	3.0	4.7
	1.0	5.5	4.6	5.3	0.7	5.4	12.9	6.0	9.2	3.4	5.4	6.0	8.4	3.0	4.9
	0.8	6.1	4.5	4.9	1.7	5.6	13.5	6.0	9.1	4.1	5.6	6.0	8.2	3.2	5.2
	0.5	9.2	4.9	5.1	3.7	4.7	15.8	5.0	8.4	3.5	4.7	5.0	7.7	3.0	4.3
n=60	1.0	4.6	4.3	4.5	0.7	5.0	9.6	5.1	6.3	4.3	5.0	5.1	6.0	4.1	4.9
	1.0	5.5	5.3	5.5	1.5	5.3	15.6	5.5	7.1	4.4	5.3	5.5	6.8	4.1	4.9
	0.8	6.9	4.9	5.1	1.9	5.3	16.6	5.5	7.1	4.7	5.3	5.5	6.5	4.2	5.2
	0.5	8.0	4.9	4.9	3.8	5.3	15.7	5.6	7.1	4.5	5.4	5.6	6.8	4.2	5.2
n=100	1.0	4.9	4.6	4.8	0.9	5.2	8.8	5.3	6.2	4.7	5.2	5.3	6.1	4.5	5.0
	1.0	4.7	4.2	4.4	1.0	4.8	15.4	5.0	5.8	4.3	4.8	5.0	5.7	4.0	4.6
	0.8	6.5	4.8	4.9	1.6	5.1	15.4	5.3	6.4	4.7	5.1	5.3	6.3	4.5	5.1
	0.5	9.7	5.8	5.8	4.2	5.0	17.1	5.0	5.9	4.5	5.0	5.0	5.8	4.2	4.7
n=200	1.0	5.7	5.4	5.7	1.0	5.3	9.9	5.4	5.9	5.1	5.3	5.4	5.8	5.0	5.2
	1.0	5.3	5.0	5.2	0.9	5.0	16.0	5.2	5.5	4.8	5.0	5.2	5.4	4.6	5.0
	0.8	5.8	4.2	4.2	1.6	5.6	16.1	5.7	6.1	5.3	5.6	5.7	6.1	5.1	5.5
	0.5	8.5	4.9	4.9	3.6	5.3	17.1	5.5	5.9	5.2	5.4	5.5	5.9	5.1	5.3

Note. 1. *F* = traditional *F* test. Box = Box’s adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted

F test. TR = the β -trimmed method using β = 0.2 (TR). FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled χ^2 test. RES = Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

2. The underlined values indicated robust Type error rates which fell between 2.5% and 7.5%.

As is shown in Table 5 for normal distributions for k=4, the Box's adjusted F test (Box), the Huynh-Feldt adjusted F test (HF), and the one sample multivariate T^2 test (FM) were robust across all sample sizes and all sphericity levels, while the traditional F test (F) provided inflated Type I error rates (9.2%, 9.15%, 8%, 9.65%, and 8.5%) across all sample sizes when the sphericity assumption was severely violated (ε =.48). Among F, Box, and HF, F provided largest Type I error rates while Box provided the smallest Type I error rates. The performance of the Geisser-Greenhouse lower bound adjusted F test (GG) was opposite of the traditional F test, performing well only when ε equaled .48 across all sample size conditions (3.1%, 3.7%, 3.75%, 4.2%, and 3.6%). The Type I error rates for the rest of the conditions were below the lower bound of Bradley's liberal criterion (2.5%) for GG. The β-trimmed method using β = 0.2 (TR), however, displayed another extreme result, with Type I error rates higher than the upper bound of Bradley's liberal criterion (7.5%) and with only one cell providing robust result when n = 15 and when there was no violation of the sphericity assumption.

Among all SEM-based methods, the maximum likelihood method (ML), the Yuan and Bentler adjusted ADF II test (FADF), the Satorra-Bentler scaled χ^2 test (SB1), and the Yuan and Bentler adjusted RES II test (FRES) delivered robust Type I error rates across all sample size conditions and all sphericity levels. Browne's asymptotic distribution-free test (ADF) and Residual-based ADF test (RES) provided inflated Type I error rates larger than the upper bound of Bradley's liberal criterion (7.5%) when sample sizes were 15 and 30 across all sphericity levels but provided robust results with sample sizes of 60, 100, and 200 across all sphericity levels. Meanwhile, the Yuan and Bentler adjusted ADF I test (YBADF) and Yuan and Bentler adjusted RES II test (YBRES) provided Type I error rates below the lower boundary of the robustness range with sample size of 15 across all sphericity levels but were robust with sample sizes of 30, 60, 100, and 200 across all sphericity levels. Across all Type I error rates provided by these RMM tests, the order of the magnitude came out with ADF > RES > ML/SB1 > FADF > FRES > YBADF >YBRES.

Among all the methods that provided robust Type I error rates, most of them provided Type I error rates of around 5%.

Table 6: Empirical Type I Error Rates (%) for Normal Distribution with *k*=8

	ε	*F*	Box	HF	GG	FM	TR	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.0	4.9	2.4	4.6	0.0	5.1	13.3	15.9	NC	NC	NC	15.9	98.7	0.0	92.4
	1.0	5.6	3.9	5.1	1.0	5.6	7.5	16.1	NC	NC	NC	16.1	98.4	0.0	92.5
	0.8	7.2	3.5	5.6	0.1	5.1	17.3	16.1	NC	NC	NC	16.1	98.5	0.0	92.9
	0.5	5.7	3.0	5.3	0.1	4.3	15.3	14.7	NC	NC	NC	14.7	98.6	0.0	93.0
n=30	1.0	4.7	3.6	4.4	0.0	4.6	18.8	8.8	NC	NC	NC	8.8	54.0	24.8	33.2
	1.0	4.5	4.1	4.5	0.9	4.9	8.0	8.1	NC	NC	NC	8.1	52.4	23.9	33.2
	0.8	7.5	4.2	5.6	0.3	5.0	19.3	8.3	NC	NC	NC	8.3	52.4	25.6	33.4
	0.5	5.3	3.6	5.1	0.0	5.4	18.9	8.1	NC	NC	NC	8.1	52.1	23.3	32.6
n=60	1.0	4.0	3.3	3.9	0.0	4.1	19.0	5.6	9.8	2.6	4.1	5.6	9.4	2.5	3.8
	1.0	4.9	4.4	4.8	0.6	4.8	8.4	7.0	11.8	4.1	5.6	7.0	11.1	3.5	5.3
	0.8	7.3	4.9	5.5	0.3	5.0	20.3	6.6	10.9	3.7	5.0	6.6	10.5	3.1	4.6
	0.5	5.8	4.6	5.6	0.1	4.9	20.1	6.1	9.9	3.4	4.9	6.1	9.5	3.2	4.7
n=100	1.0	4.8	4.5	4.8	0.0	5.1	19.8	6.0	8.2	4.3	5.1	6.0	7.9	4.0	4.7
	1.0	5.6	5.5	5.6	1.2	5.5	9.0	5.3	7.3	4.3	5.0	5.3	7.0	4.1	4.9
	0.8	7.5	5.2	5.3	0.3	4.6	24.3	5.5	8.2	4.0	4.6	5.5	7.7	3.9	4.5
	0.5	5.6	4.7	5.2	0.1	4.5	21.3	5.4	8.2	3.8	4.6	5.4	7.9	3.8	4.3
n=200	1.0	4.9	4.7	4.9	0.0	4.8	22.2	5.1	6.3	4.6	4.8	5.1	6.3	4.5	4.7
	1.0	5.8	5.7	5.8	0.9	5.7	10.1	5.6	6.7	4.7	5.0	5.6	6.6	4.6	5.0
	0.8	6.4	4.4	4.7	0.3	5.3	21.4	5.5	6.9	4.7	5.3	5.5	6.8	4.7	5.1
	0.5	6.3	5.7	5.9	0.2	5.9	22.1	6.4	7.4	5.4	5.9	6.4	7.4	5.3	5.7

Note. 1. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted

F test. TR = the β-trimmed method using β = 0.2 (TR). FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled χ^2 test. RES = Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

2. The underlined values indicated robust Type error rates which fell between 2.5% and 7.5%.

As is shown in Table 6 for normal distributions for *k*=8, the traditional *F* test, Box, HF, and FM were robust across all sample sizes and all sphericity levels. Among *F*, Box, and HF, *F* provided largest Type I error rates while Box provided the smallest Type I error rates. All the Type I error rates provided by GG were below the lower bound of Bradley's liberal criterion (2.5%). TR, however, displayed another extreme result, with Type I Error rates all higher than the upper bound of Bradley's liberal criterion (7.5%).

All SEM based methods behaved poorly when sample sizes equaled 15 and 30. ADF and its corrections did not converge, thus providing no Type I error rates across all sphericity levels. RES and their corrections (YBRES and FRES) encountered some non-convergence but were able to provide the Type I error rates. When sample size equaled 15, RES and FRES provided the Type I Error rates close to 100% while the Type I error rates provided by YBRES were all 0s. When sample size equaled 30, RES provided the Type I error rates of around 50%, FRES around 30%, and YBRES around 20%. ML and SB1, similarly, delivered inflated rejection rates across all sphericity levels. When *n* =15, ML and SB1 provided Type error rates greater than 14.5% across sphericity levels. When *n* =30, they Type error rates started to become smaller but still greater than 7.5% across sphericity levels. This might be due to the fact that the model became more complicated.

When sample sizes were larger than 30 (60, 100, and 200), ML, YBADF, FADF, SB1, YBRES and FRES were robust across conditions and all sphericity levels. But ADF and RES only provided robust Type I error rates with a sample size of 200. When the sample sizes equaled 60 and 100, the Type I error rates for both ADF and RES were inflated across all sphericity levels with only one cell being robust when the sample size was 100 and sphericity level was 0.96. Among all Type I error rates provided by these RMM tests, the order of the magnitude came out with ADF > RES > SB1 > FADF > FRES > YBADF >YBRES.

Table 7: Empirical Type I Error Rates (%) for Elliptical Distribution with *k*=4

	ε	*F*	Box	HF	GG	FM	TR	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.0	4.8	3.4	4.2	0.5	4.2	7.1	5.2	11.7	1.1	4.3	5.2	10.0	0.6	3.5
	1.0	4.8	3.0	3.9	0.7	4.0	14.7	5.1	12.9	1.1	4.0	5.1	11.1	0.5	3.1
	0.8	5.1	3.2	4.1	1.0	4.4	9.0	5.8	13.1	1.2	4.3	5.8	10.9	0.9	3.6
	0.5	9.1	4.6	5.2	3.0	4.1	17.5	5.3	13.7	1.2	4.3	5.3	11.6	0.7	3.6
n=30	1.0	4.5	3.7	4.3	0.6	4.5	8.3	5.0	8.1	2.6	4.5	5.0	7.5	2.2	3.8
	1.0	5.5	4.4	4.9	1.0	4.6	18.4	4.8	8.5	3.4	4.6	4.8	7.6	2.8	4.4
	0.8	6.1	4.1	4.6	1.7	4.0	11.2	4.4	7.2	2.7	4.0	4.4	6.4	2.3	3.4
	0.5	7.7	4.2	4.4	2.5	4.5	19.5	4.7	8.2	2.8	4.6	4.7	7.3	2.1	3.9
n=60	1.0	4.9	4.0	4.4	1.0	4.1	9.4	4.2	5.7	3.3	4.2	4.2	5.5	3.0	3.9
	1.0	5.3	4.4	4.9	1.2	4.9	22.6	5.1	6.6	4.3	4.9	5.1	6.3	3.8	4.6
	0.8	6.1	4.2	4.6	1.2	4.7	11.6	4.8	6.2	3.9	4.7	4.8	6.0	3.7	4.5
	0.5	10.1	5.7	5.8	3.6	5.1	23.1	5.2	6.9	4.6	5.1	5.2	6.6	4.0	4.9
n=100	1.0	4.9	4.5	4.7	0.8	4.7	10.2	4.9	5.8	4.4	4.7	4.9	5.4	4.1	4.5
	1.0	4.7	4.1	4.5	0.9	4.4	22.8	4.5	5.5	3.7	4.4	4.5	5.1	3.6	4.1
	0.8	6.6	5.3	5.4	2.1	4.9	12.1	4.9	6.2	4.5	4.9	4.9	5.8	4.4	4.8
	0.5	8.5	4.4	4.5	3.3	4.6	23.6	4.7	5.5	4.0	4.6	4.7	5.3	3.9	4.4
n=200	1.0	4.5	4.4	4.4	1.0	4.4	9.8	4.5	4.8	4.3	4.4	4.5	4.8	4.3	4.3
	1.0	4.8	4.5	4.6	0.9	4.4	24.0	4.5	5.3	4.3	4.4	4.5	5.1	4.3	4.4
	0.8	6.7	4.9	5.0	2.3	5.7	11.3	5.8	6.3	5.4	5.8	5.8	6.3	5.3	5.6
	0.5	10.5	6.5	6.5	4.6	4.8	24.3	4.8	5.5	4.6	4.8	4.8	5.4	4.5	4.6

Note. 1. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted

F test. TR = the β-trimmed method using β = 0.2 (TR). FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled χ^2 test. RES = Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

2. The underlined values indicated robust Type error rates which fell between 2.5% and 7.5%.

As is shown in Table 7 for elliptical distributions for *k*=4, Box, HF, and FM were robust across all sample sizes and all sphericity levels while *F* provided inflated Type I error rates across all sample sizes when the sphericity assumption was severely violated (ε =.48) based on Bradley's criteria. Among *F*, Box, and HF, *F* provided largest Type I error rates while Box provided the smallest Type I error rates. GG was the opposite of the traditional *F* test, providing robust Type I error rates only when the sphericity assumption was severely violated (ε =.48) across all sample size conditions. The Type I error rates for the rest of the conditions were below the lower bound of Bradley's liberal criterion (2.5%) for GG. TR, however, displayed another extreme result, with Type I Error rates higher than the upper bound of Bradley's liberal criterion (7.5%) with only one cell proving robust result when n=15 and when there was no violation of sphericity assumption).

Among all SEM based methods, ML, FADF, SB1, and FRES were robust across all sample size conditions and all sphericity levels. ADF delivered inflated Type I error rates larger than the upper bound of Bradley's liberal criterion (7.5%) when sample sizes were 15 and 30 across all sphericity levels but were robust when sample sizes were 60, 100, and 200 across all sphericity levels. RES provided inflated Type I error rates with sample size of 15 across all sphericity levels but was robust with sample sizes of 60, 100, and 200 across all sphericity levels. When sample size was 30, RES was robust except for one cell providing inflated Type I error rate (7.6%)

when ε =.96. Meanwhile, YBADF delivered Type I error rates below the lower boundary of the robustness range with sample size of 15 across all sphericity levels but was robust across all sphericity levels with sample sizes of 30, 60, 100, and 200. YBRES provided Type I error rates below the lower boundary of the robustness range with sample sizes of 15 and 30 across all sphericity levels with only one exception with sphericity level of .96 (Type I error rate = 2.8%) but was robust across all sphericity levels with sample sizes of 30, 60, 100, and 200. Among all Type I error rates provided by these RMM tests, the order of the magnitude came out with ADF > RES > SB1 > FADF > FRES > YBADF >YBRES.

Among all the methods that provided robust Type I error rates, most of them provided Type I error rates of around 5%.

Table 8: Empirical Type I Error Rates (%) for Elliptical Distribution with *k*=8

	ε	*F*	Box	HF	GG	FM	TR	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.0	5.0	1.8	4.0	0.0	3.9	22.3	14.3	NC	NC	NC	14.3	98.3	0.0	93.4
	1.0	5.8	2.5	4.2	0.0	4.5	22.5	15.1	NC	NC	NC	15.1	98.7	0.0	93.4
	0.8	5.7	2.4	4.0	0.5	4.8	21.4	15.1	NC	NC	NC	15.1	98.5	0.0	93.3
	0.5	8.1	3.0	4.1	0.3	3.8	21.9	14.2	NC	NC	NC	14.2	98.9	0.0	93.3
n=30	1.0	4.5	2.8	3.8	0.1	4.0	28.9	6.7	NC	NC	NC	6.7	62.5	31.1	42.0
	1.0	4.9	3.0	3.8	0.0	4.3	28.4	8.0	NC	NC	NC	8.0	63.1	29.0	40.9
	0.8	6.5	2.8	3.8	0.2	4.9	28.1	8.0	NC	NC	NC	8.0	65.3	30.5	43.4
	0.5	8.1	3.7	4.8	0.8	4.5	28.2	7.5	NC	NC	NC	7.5	63.0	31.1	42.1
n=60	1.0	5.3	3.9	4.7	0.0	5.0	33.3	5.9	10.1	3.3	5.0	5.9	9.3	3.0	4.3
	1.0	5.3	3.8	4.4	0.1	4.5	32.9	5.4	8.9	3.1	4.5	5.4	8.3	2.7	4.3
	0.8	7.0	3.9	4.8	0.3	4.6	33.6	5.7	10.3	2.7	4.6	5.7	9.5	2.5	4.1
	0.5	8.9	5.1	5.5	0.5	3.9	32.0	5.5	10.6	2.7	4.0	5.5	9.9	2.5	3.7
n=100	1.0	4.5	3.5	4.2	0.1	4.7	38.1	5.5	8.0	3.9	4.7	5.5	7.5	3.8	4.5
	1.0	5.7	4.1	4.7	0.1	4.5	36.0	5.4	8.0	3.9	4.5	5.4	7.9	3.7	4.3
	0.8	6.4	4.4	4.8	0.3	4.3	34.5	4.7	7.7	3.5	4.3	4.7	7.0	3.4	4.2
	0.5	8.4	4.5	4.9	0.5	4.4	34.6	4.8	6.8	3.6	4.4	4.8	6.2	3.3	4.2
n=200	1.0	5.2	4.6	4.8	0.1	5.3	38.6	5.6	6.4	4.7	5.3	5.6	6.3	4.6	5.1
	1.0	5.3	4.6	4.8	0.0	4.5	38.1	4.7	5.8	4.0	4.5	4.7	5.6	3.8	4.3
	0.8	5.7	3.7	3.8	0.2	3.6	36.3	3.8	4.9	3.4	3.6	3.8	4.7	3.3	3.5
	0.5	8.2	4.5	4.6	0.4	4.4	35.5	4.8	5.8	3.9	4.4	4.8	5.5	3.8	4.2

Note. 1. *F* = traditional *F* test. Box = Box’s adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted

F test. TR = the β -trimmed method using β = 0.2 (TR). FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled χ^2 test. RES = Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

2. The underlined values indicated robust Type error rates which fell between 2.5% and 7.5%.

As is shown in Table 8 for elliptical distributions for *k*=8, HF and FM were robust across all sample sizes and all sphericity levels while *F* provided inflated Type I error rates across all samples sizes when the sphericity assumption was severely violated (ε =.48). Box was robust across all sample sizes and all sphericity levels except when sample size equaled 15 without any violation of sphericity assumption (ε =1). Among *F*, Box, and HF, *F* provided largest Type I error rates while Box provided the smallest Type I error rates. All the Type I error rates provided by GG were below the lower bound of Bradley's liberal criterion (2.5%). TR, however, displayed another extreme result, whose Type I Error rates were all higher than the upper bound of Bradley's liberal criterion (7.5%).

All SEM based methods behaved poorly when sample sizes equaled 15 and 30. ADF and its corrections (YBADF and FADF) did not converge, thus providing no Type I error rates across all sphericity levels. RES and its corrections (YBRES and FRES) encountered some non-convergence but were able to provide the Type I error rates. When sample size equaled 15, RES and FRES provided the Type I Error rates close to 100% while the Type I error rates provided by YBRES were all 0s. ML and SB1, on the other hand, provided inflated Type I error rates higher than the upper bound of Bradley's liberal criterion (7.5%). When sample size equaled 30, RES provided the Type I error rates of around 60%, FRES around 40%, and YBRES around 30%. ML and SB1 delivered robust Type I error rates when the sphericity levels were 1 and 0.48 and inflated Type I error rates when the sphericity levels were .96 and .75.

When sample size was larger than 30 (60, 100, and 200), ML, YBADF, FADF, SB1, YBRES, and FRES were robust across sample size conditions and all sphericity levels. But ADF and RES only delivered robust Type I error rates across sphericity levels when the sample size is 200. When the sample sizes equaled 60 and 100, ADF provided inflated Type I error rates across all sphericity levels with only one cell being robust when the sample size was 100 and sphericity level was 0.48 (Type I error rate = 6.8). Meanwhile, RES delivered inflated rejection rates across all sphericity levels with only two cells providing robust results when the sample size was 100 and sphericity level was .75 and .48. Among all Type I error rates provided by these RMM tests, the order of the magnitude came out with ADF > RES > SB1 > FADF > FRES > YBADF >YBRES.

Among all the methods that provided robust Type I error rates, most of them provided Type I error rates of around 5%.

Table 9: Empirical Type I Error Rates (%) for Moderately Non-normal Distribution with *k*=4

	ε	*F*	Box	HF	GG	FM	TR	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.0	4.3	2.6	3.5	0.5	5.0	6.5	6.2	14.3	1.8	5.2	6.2	11.8	0.9	4.1
	1.0	4.2	2.4	3.4	0.3	4.0	11.4	5.1	11.4	1.4	4.1	5.1	9.4	0.9	3.1
	0.8	5.5	3.1	4.0	1.0	4.4	8.1	5.6	12.2	11.5	4.0	5.6	11.5	0.5	3.5
	0.5	10.7	6.5	6.8	4.3	8.8	19.0	10.3	18.6	4.0	8.7	10.3	16.2	2.6	7.8
n=30	1.0	5.1	3.7	4.2	0.9	4.6	9.7	5.1	8.2	3.3	4.6	5.1	7.2	2.9	4.1
	1.0	4.5	3.2	3.7	0.7	4.6	15.9	4.9	8.2	3.0	4.6	4.9	7.5	2.6	4.0
	0.8	6.1	3.9	4.5	1.2	4.3	10.4	4.8	7.8	3.2	4.4	4.8	7.0	2.7	4.0
	0.5	10.3	7.0	7.0	5.2	7.3	22.1	7.4	11.4	5.4	7.3	7.4	10.7	4.7	6.7
n=60	1.0	4.7	3.9	4.1	1.0	5.1	9.1	5.3	6.6	4.4	5.1	5.3	6.3	4.3	4.8
	1.0	4.4	3.7	3.9	0.9	5.0	19.6	5.2	6.2	4.3	5.0	5.2	5.9	4.1	4.9
	0.8	7.3	4.7	5.0	1.7	4.5	11.5	4.6	6.3	3.7	4.5	4.6	6.1	3.6	4.2
	0.5	9.0	4.6	4.8	3.1	5.7	24.6	6.0	7.6	4.8	5.7	6.0	7.4	4.6	5.6
n=100	1.0	5.2	4.6	4.8	0.9	5.7	8.2	5.7	6.6	5.1	5.7	5.7	6.4	5.0	5.4
	1.0	5.1	4.0	4.4	0.8	4.3	22.8	4.4	5.4	3.5	4.3	4.4	5.1	3.3	4.1
	0.8	7.4	5.3	5.4	1.8	4.9	11.3	4.9	5.7	4.4	4.9	4.9	5.5	4.2	4.7
	0.5	9.8	5.5	5.6	3.6	5.7	28.4	5.8	6.7	5.2	5.7	5.8	6.5	5.1	5.6
n=200	1.0	4.8	4.4	4.6	0.6	5.1	9.5	5.2	5.5	5.0	5.2	5.2	5.4	4.9	5.1
	1.0	4.7	4.4	4.4	0.9	4.7	22.7	4.7	5.2	4.4	4.7	4.7	5.1	4.4	4.6
	0.8	6.4	4.6	4.8	1.7	4.0	11.8	4.0	4.2	4.0	4.0	4.0	4.2	4.0	4.0
	0.5	8.0	4.4	4.4	3.3	5.5	33.9	5.7	5.8	5.4	5.6	5.7	5.8	5.4	5.5

Note. 1. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted

F test. TR = the β -trimmed method using β = 0.2 (TR). FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled χ^2 test. RES = Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

2. The underlined values indicated robust Type error rates which fell between 2.5% and 7.5%.

As is shown in Table 9 for moderately nonnormal data for *k*=4, HF was robust across all sample sizes and all sphericity levels while *F* provided inflated Type I error rates (10.7%, 10.3%, 9%, 9.75%, 8%) when the sphericity assumption was severely violated (ε =.48). GG was to the opposite of the traditional *F* test which was robust only when the sphericity level equaled .48 across all sample size conditions (4.25%, 5.2%, 3.1%, 3.55%, and 3.25%). Box delivered robust Type I error rates across all sample sizes and sphericity levels except when the sample size equaled 15 with sphericity level of 0.96 (Type I error rate = 2.4%). One sample multivariate T^2 test (FM) was robust across all sample sizes and sphericity levels except one cell delivering inflated rejection rate (8.75%) when the sample size equaled 15 with sphericity level of 0.48. Among F, Box, and HF, F provided largest Type I error rates while Box provided the smallest Type I error rates. TR, however, displayed another extreme result, whose Type I Error rates were higher than the upper bound of Bradley's liberal criterion (7.5%) with only one exception (Type I error rate equaled 7.1% when *n*=15 and ε =1).

Among all SEM based methods, ML, FADF, SB1, and FRES controlled the Type I error rates well across all sample size conditions and all sphericity levels with only one inflated rejection rate when sample size was 15 at the sphericity level of .48. Browne's asymptotic distribution-free test (ADF) delivered inflated Type I error rates when sample sizes were 15 and 30 across all sphericity levels but provided robust

Type I error rates with sample sizes of 60, 100, and 200 across all sphericity levels with only one cell providing inflated rejection rate when *n*=60, ε =0.48. Meanwhile the Type I error rates of RES were inflated with sample size of 15 across all sphericity levels but provided robust Type I error rates with sample sizes of 30, 60, 100, and 200 across all sphericity levels with only one cell yielding inflated rejection rate when *n*=30, ε =.48. Conversely, YBADF and YBRES provided Type I error rates below the lower boundary of the robustness range with sample size of 15 across all sphericity levels except one robust cell at ε =0.48. These two tests were robust with sample sizes of 30, 60, 100, and 200 across all sphericity levels. Among all Type I error rates provided by these RMM tests, the order of the magnitude came out with ADF > RES > SB1 > FADF > FRES > YBADF >YBRES.

Among all the methods that provided robust Type I error rates, most of them provided Type I error rates of around 5%.

Table 10: Empirical Type I Error Rates (%) for Moderately Non-normal Distribution with *k*=8

	ε	*F*	Box	HF	GG	FM	TR	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.0	3.7	1.4	2.4	0.0	4.9	15.7	15.3	NC	NC	NC	15.3	97.9	0.0	92.7
	1.0	5.3	1.5	3.0	0.1	5.6	18.5	16.5	NC	NC	NC	16.5	99.1	0.0	94.6
	0.8	7.6	3.2	4.8	0.1	8.5	20.9	22.7	NC	NC	NC	22.7	99.1	0.0	93.7
	0.5	8.8	3.7	4.7	0.5	9.2	22.0	24.4	NC	NC	NC	24.4	99.0	0.0	95.1
n=30	1.0	4.7	2.3	3.2	0.0	4.6	25.4	8.1	NC	NC	NC	8.1	65.8	36.0	46.5
	1.0	4.4	2.5	3.5	0.0	4.7	24.9	8.3	NC	NC	NC	8.3	67.7	32.7	45.0
	0.8	6.7	3.5	4.3	0.1	8.3	26.5	12.1	NC	NC	NC	12.1	68.4	34.5	47.9
	0.5	7.6	3.9	4.4	0.6	9.7	27.2	13.6	NC	NC	NC	13.6	68.0	36.3	49.1
n=60	1.0	5.2	3.8	4.2	0.0	5.1	29.1	6.2	10.3	3.9	5.1	6.2	9.7	3.4	4.9
	1.0	5.8	3.9	4.2	0.0	4.8	30.8	5.9	11.0	3.2	4.8	5.9	10.1	2.8	4.6
	0.8	7.4	4.1	4.8	0.2	7.4	31.5	8.4	12.7	5.8	7.4	8.4	11.9	5.3	7.0
	0.5	9.6	5.2	5.5	0.5	7.8	33.8	9.1	15.2	6.1	7.8	9.1	14.2	5.5	7.2
n=100	1.0	4.7	3.6	4.0	0.0	5.0	31.8	5.8	8.1	3.9	5.0	5.8	7.8	3.8	4.8
	1.0	5.3	4.4	4.7	0.1	4.9	33.7	5.5	7.9	4.0	4.9	5.5	7.5	3.6	4.6
	0.8	6.3	4.3	4.7	0.2	6.5	33.4	7.1	9.2	5.2	6.5	7.1	8.8	4.9	6.2
	0.5	8.9	5.4	5.8	0.6	8.1	35.6	9.0	11.5	7.2	8.1	9.0	11.2	6.8	8.0
n=200	1.0	5.4	4.7	4.8	0.0	4.9	33.2	5.1	6.7	4.3	4.9	5.1	6.5	4.1	4.9
	1.0	5.2	3.9	4.2	0.1	5.0	32.3	5.1	6.4	4.4	5.1	5.1	6.4	4.2	4.7
	0.8	5.4	3.9	3.9	0.2	5.4	36.8	5.7	6.6	4.8	5.4	5.7	6.5	4.7	5.2
	0.5	7.6	4.1	4.1	0.5	6.0	39.7	6.3	7.3	5.7	6.0	6.3	7.2	5.6	6.0

Note. 1. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted

F test. TR = the β -trimmed method using β = 0.2 (TR). FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled χ^2 test. RES = Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

2. The underlined values indicated robust Type error rates which fell between 2.5% and 7.5%.

As is shown in Table 10 for moderately nonnormal data for *k*=8, Box and HF were robust across all sphericity levels when sample sizes were larger than 15. When *n*=15, Box and HF provided model rejection rates smaller than the lower boundary of Bradley's liberal criterion (2.5%) for some cells. FM was robust across all sample sizes when there was no or little violation of sphericity assumption (ε =1, .96). On the other hand, the Type I error rates for FM tended to be inflated when the violation of sphericity assumption became more serious (ε=.75, .48) but when sample sizes increased, this test became more robust when the sample sizes were smaller than 200. When *n*=200, FM was robust across all spheriticy levels. *F*, similar to the previous conditions, was robust across all conditions except when there was a serious violation of sphericity assumption (ε =.48). Among *F*, Box, and HF, *F* provided largest Type I error rates while Box provided the smallest Type I error rates. All the Type I error rates provided by GG were below the lower bound of Bradley's liberal criterion (2.5%). TR, however, displayed another extreme result, whose Type I Error rates were all higher than the upper bound of Bradley's liberal criterion (7.5%).

All SEM based methods performed poorly when sample sizes equaled 15 and 30. ADF and its corrections (YBADF and FADF) did not converge, thus providing no Type I error rates across all sphericity levels. RES and their corrections (YBRES and FRES) encountered some non-convergence but were able to provide the Type I error rates. ML and SB1, on the other hand, provided inflated Type I error rates higher than the upper bound of Bradley's liberal criterion (7.5%). When sample size equaled 15, RES and FRES provided the Type I error rates close to 100% while the Type I error rates provided by YBRES were all 0s. When sample size equaled 30, RES provided the Type I error rates of close to 70%, FRES close to 50%, and YBRES around 30%. When sample size was larger than 30 (60, 100, and 200), YBADF and YBRES were robust across all sphericity conditions. ML, FADF, SB1, and FRES controlled the Type I error rates well with only one or two inflated rejection rates when there was serious violation of sphericity assumption with sample sizes of 60 and 100 and all of these tests were robust when the sample size was 200. Contrary to the performance of the other RMM methods, ADF and RES only provided robust Type I error rates across sphericity levels when the sample size is 200 but delivered inflated rejection rates with sample sizes of 60 and 100. Among all Type I error rates provided by these RMM tests, the order of the magnitude came out with ADF > RES > SB1 > FADF > FRES > YBADF >YBRES.

Among all the methods that provided robust Type I error rates, most of them provided Type I error rates of around 5%.

Table 11: Empirical Type I Error Rates (%) for Severely Non-normal Distribution with *k*=4

	ε	*F*	Box	HF	GG	FM	TR	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.0	3.5	2.1	2.6	0.3	3.3	6.2	4.5	12.1	0.9	3.5	4.5	9.9	0.4	2.8
	1.0	4.4	2.1	2.8	0.4	4.4	15.5	5.6	14.0	1.4	4.4	5.6	12.1	0.8	3.7
	0.8	7.1	4.2	5.1	1.8	6.6	12.1	7.9	16.0	3.1	6.6	7.9	14.5	1.8	5.7
	0.5	10.8	6.9	7.4	4.4	8.4	24.1	10.1	20.0	4.4	9.9	10.1	18.7	3.2	8.3
n=30	1.0	4.1	2.6	3.0	0.4	3.9	8.3	4.5	8.0	2.5	4.0	4.5	6.9	2.0	3.4
	1.0	3.6	2.3	2.6	0.5	4.0	23.7	4.4	7.7	3.0	4.1	4.4	7.1	2.3	3.7
	0.8	7.7	5.6	6.0	2.1	6.9	17.6	7.5	11.1	4.7	7.0	7.5	10.5	3.8	6.3
	0.5	10.2	6.0	6.3	4.5	8.3	27.7	8.6	12.1	6.5	8.4	8.6	11.1	5.9	7.8
n=60	1.0	4.8	3.6	3.9	0.6	4.9	8.6	4.5	6.7	4.0	5.0	4.5	6.9	2.0	3.4
	1.0	4.5	3.3	3.5	0.6	3.6	28.2	3.7	5.0	3.2	3.6	3.7	4.7	2.9	3.5
	0.8	7.1	5.5	5.5	2.0	5.7	23.2	6.0	7.8	5.2	5.8	6.0	7.4	4.8	5.5
	0.5	8.9	5.4	5.5	3.4	7.5	36.1	7.8	9.0	6.8	7.5	7.8	8.6	6.5	7.3
n=100	1.0	3.9	3.2	3.3	0.5	4.1	9.6	4.1	5.1	3.6	4.1	4.1	5.0	3.5	3.9
	1.0	5.2	4.2	4.2	0.8	5.3	33.8	5.4	6.4	5.1	5.4	5.4	6.3	4.8	5.2
	0.8	6.6	5.0	5.0	1.9	5.7	28.9	5.8	7.0	5.0	5.7	5.8	6.7	4.6	5.4
	0.5	9.6	5.7	5.7	4.3	7.4	38.8	7.6	8.5	6.8	7.4	7.6	8.4	6.7	7.1
n=200	1.0	4.6	3.9	3.9	0.8	3.8	11.2	4.0	4.3	3.6	3.9	4.0	4.3	3.5	3.8
	1.0	4.8	4.2	4.3	1.4	5.1	36.1	5.1	5.4	4.9	5.1	5.1	5.4	4.9	5.1
	0.8	6.9	5.5	5.5	2.0	5.7	43.5	5.8	6.1	5.5	5.7	5.8	6.1	5.4	5.6
	0.5	8.9	5.5	5.5	4.0	5.8	44.2	5.9	6.5	5.5	5.8	5.9	6.4	5.4	5.7

Note. 1. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted

F test. TR = the β -trimmed method using β = 0.2 (TR). FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled χ^2 test. RES = Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

2. The underlined values indicated robust Type error rates which fell between 2.5% and 7.5%.

As is shown in Table 11 for severely nonnormal data for *k*=4, HF was robust across all sample sizes and all sphericity levels while *F* provided inflated Type I error rates (10.8%, 10.2%, 8.9%, 9.6%, 8.9%) when the sphericity assumption was severely violated (ε =.48). GG was to the opposite of the traditional *F* test which provided robust Type I error rates only when the sphericity level equaled .48 across all sample size conditions (4.4%, 4.5%, 3.4%, 4.25%, and 4%). Box and FM controlled the Type I error rates well with only one or two cells providing inflated rejection rates when sample sizes equaled 15 and 30. FM delivered inflated Type I errors when there was serious violation of sphericity assumption (ε =.48) but Box produced inflated Type I error when the sphericity assumption was not violated or little violated (ε =1, .96). Among *F*, Box, and HF, *F* provided largest Type I error rates while Box provided the smallest Type I error rates. TR, however, displayed another extreme result, whose Type I error rates were higher than the upper bound of Bradley's liberal criterion (7.5%) with only one robust cell (n=15, ε =1) as an exception.

Among all SEM based methods, YBADF controlled the Type I error rates well with only two cells providing inflated rejection rates when there was no or little violation of sphericity assumption (ε =1, .96) with sample size of 15. FADF and FRES performed well with only two inflated rejection rates when there was serious violation of sphericity assumption (ε =.48) with sample size of 15 and 30. ML and SB1 were robust across all sample sizes except when there was serious violation of sphericity assumption (ε =.48). ADF and RES did not control Type I error rates well except when the sample size was 200. ADF provided inflated rejection rates across all sphericity levels when sample sizes were small (15 and 30) and when there was serious violation of sphericity assumption (ε =.48) when the sample sizes were 60 and 100. RES delivered inflated Type I error rates when sample size was 15 across all sphericity levels as well as when the sphericity assumption was seriously violated with sample sizes larger than 15. YBRES was robust across all conditions with large sample sizes (100 and 200) but provided Type I error rates below the lower boundary of the robustness range when there was no or little violation of sphericity assumption with sample sizes of 30 and 60. When sample size equaled 15, YBRES was robust only when the sphericity assumption was seriously violated. Among all Type I error rates provided by these RMM tests, the order of the magnitude came out with ADF > RES > SB1 > FADF > FRES > YBADF >YBRES.

Among all the methods that provided robust Type I error rates, most of them provided Type I error rates of around 5%.

Table 12: Empirical Type I Error Rates (%) for Severely Non-normal Distribution with *k*=8

	ε	*F*	Box	HF	GG	FM	TR	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.0	3.1	0.6	1.3	0.0	3.1	23.2	13.4	NC	NC	NC	13.4	98.1	0.0	93.3
	1.0	4.2	1.2	2.2	0.0	3.9	24.6	13.9	NC	NC	NC	13.9	98.0	0.0	92.8
	0.8	5.1	1.1	1.9	0.0	5.6	26.0	18.2	NC	NC	NC	18.2	99.0	0.0	95.0
	0.5	8.1	2.9	3.9	0.2	9.4	30.4	24.5	NC	NC	NC	24.5	99.2	0.0	95.8
n=30	1.0	4.2	1.0	1.7	0.0	3.1	39.7	6.3	NC	NC	NC	6.3	78.7	48.4	61.4
	1.0	4.1	1.6	2.0	0.0	4.0	37.0	7.3	NC	NC	NC	7.3	79.2	48.6	69.1
	0.8	5.7	2.1	2.7	0.0	6.9	38.8	10.2	NC	NC	NC	10.2	80.7	50.5	62.8
	0.5	9.1	4.1	4.8	0.6	11.3	42.3	15.4	NC	NC	NC	15.4	82.8	53.7	66.1
n=60	1.0	4.8	2.5	2.9	0.0	5.0	48.6	6.1	10.2	3.5	5.0	6.1	9.5	3.2	4.6
	1.0	5.9	3.4	3.7	0.0	4.9	48.7	5.8	9.9	3.1	5.0	5.8	9.4	3.0	4.5
	0.8	6.0	2.9	3.4	0.2	7.3	48.0	8.8	13.7	5.8	7.3	8.8	12.9	5.1	6.9
	0.5	7.8	3.4	3.8	0.3	8.7	45.8	10.5	15.8	6.8	8.4	10.5	15.5	6.2	8.4
n=100	1.0	4.3	3.3	3.4	0.0	5.2	5.4	5.9	8.5	4.3	5.2	5.9	8.1	4.0	4.9
	1.0	4.5	3.0	3.2	0.0	5.0	52.2	5.5	8.1	4.1	5.5	5.5	7.7	3.6	4.9
	0.8	6.0	3.4	3.6	0.0	7.0	52.8	7.8	10.6	5.8	7.1	7.8	10.3	5.6	6.7
	0.5	8.9	5.2	5.5	0.7	9.1	52.7	9.8	12.5	7.8	9.1	9.8	12.2	7.4	8.8
n=200	1.0	6.2	4.6	4.9	0.1	5.3	55.2	5.7	7.0	5.0	5.3	5.7	6.8	4.7	5.3
	1.0	5.8	4.2	4.5	0.0	5.7	54.8	6.3	7.4	4.9	5.7	6.3	7.2	4.8	5.6
	0.8	6.7	4.3	4.4	0.2	6.8	57.6	7.1	8.2	6.3	6.8	7.1	8.1	6.3	6.7
	0.5	6.8	4.2	4.3	0.4	7.1	56.4	7.3	8.5	6.8	7.1	7.3	8.3	6.5	7.1

Note. 1. *F* = traditional *F* test. Box = Box’s adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted

F test. TR = the β -trimmed method using β = 0.2 (TR). FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled χ^2 test. RES = Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

2. The underlined values indicated robust Type error rates which fell between 2.5% and 7.5%.

As is shown in Table 12 for severely nonnormal data for *k*=8, Box and HF were robust across all sphericity levels when sample sizes were larger than 30. When sample sizes equaled 15 and 30, Box and HF provided model rejection rates smaller than the lower boundary of Bradley's liberal criterion (2.5%) for some cells and were robust when there was severe violation of sphericity assumption (ε =.48). FM tended to provide inflated Type I error rates when there was severe violation of sphericity assumption but remained robust for the rest of the conditions when the sample sizes were smaller than 200. When *n*=200, FM was robust across all spheriticy levels. *F*, similar to the previous conditions, was robust across all conditions except when there was a serious violation of sphericity assumption (ε =.48) when the sample sizes were smaller than 200 but provided robust rejection rates across all sphericity levels when sample size equaled 200. Among *F*, Box, and HF, *F* provided largest Type I error rates while Box provided the smallest Type I error rates. All the Type I error rates provided by GG were close to 0 and below the lower bound of Bradley's liberal criterion (2.5%). TR, however, displayed another extreme result, whose Type I Error rates were all higher than the upper bound of Bradley's liberal criterion (7.5%).

All SEM based methods performed poorly when sample sizes equaled 15 and 30. ADF and its corrections (YBADF and FADF) did not converge, thus providing no Type I error rates across all sphericity levels. RES and their corrections (YBRES and FRES) encountered some non-convergence but were able to provide the Type I error rates. ML and SB1, on the other hand, provided inflated Type I error rates higher than the upper bound of Bradley's liberal criterion (7.5%) under the majority of the conditions. When sample size equaled 15, RES and FRES provided the Type I Error rates close to 100% while the Type I error rates provided by YBRES were all 0s. When sample size equaled 30, RES provided the Type I error rates of close to 80%, FRES around 60%, and YBRES around 50%. When sample sizes were larger than 30 (60, 100, and 200), YBRES was robust across all sphericity conditions. YBADF, FADF, FRES controlled Type error rates well with one or two cells providing inflated rejection rates when there was severe violation of sphericity assumption. ML and SB1 were robust across all sphericity levels when sample size equaled 200 but did not perform well when sample sizes were 60 and 100, which provided inflated rejection rates when ε =.75 and ε =.48. Contrary to the performance of the other RMM methods, ADF and RES performed poorly by providing inflated rejection rates across all conditions with only two cells proving robust result when the sample size equaled 200 and when there was no or little sphericity assumption violation. Among all Type I error rates provided by these RMM tests, the order of the magnitude came out with ADF > RES > SB1 > FADF > FRES > YBADF >YBRES.

Among all the methods that provided robust Type I error rates, most of them provided Type I error rates of around 5%.

Empirical Power

Based on the results from Type I error rates analysis conducted above, it was discovered that Type error rates GG yielded were smaller than the lower boundary of Bradley's liberal criterion (2.5%), even close to 0 in most conditions. However, this method was still included in power analysis. On the other hand, the majority of Type I error rates produced by TR were beyond the upper boundary of Bradley's liberal criterion (7.5%), thus being removed from the power analysis.

Tables 13, 15, 17, and 19 present the results of power analysis under the four types of distributions when the number of levels *k* equals 4. Tables 14, 16, 18, and 20 show the results of power analysis under the four types of distributions when the number of levels *k* equals 8. Among these eight tables, Table 13 and Table 14 present the results of power analysis for the normal distribution, while Table 15 and Table 16 show the results of power analysis for elliptical distribution under sphericity levels of 1, .96, .75, .48 and sample sizes of 15, 30, 60, 100, 200. Table 17 and Table 18 present the results for the moderately non-normal distribution, while Table 19 and Table 20 show the results for severely the non-normal distribution under four sphericity levels (1, .96, .75, .48) and five sample sizes (15, 30, 60, 100, 200).

Table 13: Empirical Power (%) for Normal Distribution with *k*=4

	ε	*F*	Box	HF	GG	FM	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.00	11.0	9.2	10.5	2.7	9.2	14.7		5.0	12.3	14.7		20.3	10.3
	0.96	9.9	7.4	9.3	2.4	9.3	12.5		4.3	10.3	12.5		20.5	8.3
	0.75	11.6	9.0	10.4	4.3	8.6	12.4		4.0	9.8	12.4		21.1	8.2
	0.48		10.8	11.7	7.6	12.3	18.3		8.2	17.6	18.3		29.3	13.2
n=30	1.00	17.1	15.9	16.8	4.8	16.5	16.6		3.4	10.1	16.6		29.0	13.4
	0.96	18.5	16.9	17.8	6.1	15.2	15.2		4.3	9.9	15.2		27.8	12.5
	0.75	20.4	17.7	18.1	10.7	14.7	14.2		4.2	9.8	14.2		26.0	12.1
	0.48		15.4	16.1	12.1	22.9	13.5		3.3	9.7	13.5		26.7	11.7
n=60	1.00	28.3	27.5	28.1	11.0	27.9	34.9	43.2	37.6	38.7	34.9	34.9	37.2	32.5
	0.96	32.1	30.4	31.1	13.8	28.8	36.4	39.8	33.2	33.6	36.4	36.4	39.3	32.6
	0.75	39.8	33.8	34.7	20.8	28.4	30.4	36.1	27.4	29.8	30.4	30.4	34.6	28.3
	0.48		31.2	31.3	25.0	42.8	50.1	52.8	49.3	49.4	50.1	50.1	54.1	49.8
n=100	1.00	49.8	48.8	49.5	25.3	48.8	52.3	54.6	51.1	51.9	52.3	52.3	54.2	51.7
	0.96	55.6	53.9	54.8	30.9	51.5	56.1	57.1	53.2	54.9	56.1	56.1	56.9	54.3
	0.75	56.5	50.9	51.0	35.8	45.5	46.2	48.7	44.6	45.9	46.2	46.2	47.9	44.1
	0.48		46.1	46.6	41.5	68.6	69.7	72.4	68.1	69.5	69.7	69.7	71.2	67.8
n=200	1.00	84.3	84.0	84.2	63.3	83.3	84.7	85.1	84.8	85.4	84.7	84.7	85.2	82.3
	0.96	87.5	86.6	87.2	68.8	85.3	87.7	87.9	86.7	87.3	87.7	87.7	87.9	87.5
	0.75	85.6	82.0	82.1	70.9	77.1	83.4	82.3	82.1	83.1	83.4	83.4	84.3	83.1
	0.48		76.6	76.6	72.0	93.6	95.5	95.4	95.2	95.3	95.5	95.5	96.0	95.3

Note. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted *F* test. FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF

= Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled χ^2 test. RES = Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

As is shown in Table 13 for normal data for *k*=4, the results for some cells were removed from power analysis because these cells provided Type I error rates larger than 7.5%. Therefore, the results for the traditional *F* test across all sample sizes when the sphericity assumption was severely violated (ε =.48) and those for ADF and RES when sample sizes equal 15 and 30 across all sphericity levels were removed from the final analysis for the same reason.

Generally speaking, when sample sizes increased, all the methods became more powerful. For *n*=15, power in the majority of the cells was less than 20% while the majority of the cells provided power greater than 80% except some cells from ANOVA based methods with *n*=200.

For ANOVA based methods, except for ε=0.48, the order of the strength of power estimates came out with F > HF > BOX > GG and HF > FM across all sample sizes. When ε=0.48, with F being removed from comparison, FM provided strongest power and the order of the strength of power estimates came out with FM > HF > BOX > GG across all sample sizes.

For RMM methods, ML and SB1 yielded the same power estimates across all conditions. FADF was consistently more powerful than YBADF and the power estimates provided by YBRES were consistently higher than those by YBADF across all conditions. The order of the strength of power estimates for YBRES, SB1/ML, and FRES yielded YBRES > SB1/ML > FRES across all sample sizes and sphericity levels. Both YBRES and ML/SB1 provided comparatively higher power estimates than FADF across all sample sizes and sphericity levels except when *n*=60, ε=1 and *n*=200, ε=1. Therefore, except when *n*=60, ε=1 and *n*=200, ε=1, the order of the magnitude of power estimates could come out with YBRES > ML/SB1 > FADF > YBADF across all sample sizes and sphericity levels.

When *n*=15, YBRES provided the highest power among all methods that entered into analysis across all sphericity levels with a range from 20.3% to 29.3%. Except for ε=0.48, the order of the strength of power was obtained as YBRES > SB1/ML > *F* > HF > FRES > YBADF across all other sphericity levels. *F* yielded the highest power estimates among all ANOVA-based methods and the power estimates for YBRES were more than twice those for *F*. When ε=0.48, FM (12.3%) yielded the highest power estimates among all ANOVA-based methods and the power estimate for YBRES (29.2%) was more than twice that for FM. BOX provided larger power estimates than YBADF across all sphericity levels, and FRES was more powerful than BOX except when ε=.75, thus having YBRES > SB1/ML > FADF > FM > BOX > YBADF.

When *n*=30, power estimates provided by ANOVA-based methods increased by a greater percentage (more than 50%) than those provided by RMM methods. But YBRES still provided the highest power estimates ranging from 26% to 29% while YBADF provided the lowest power estimates between 3.3% and 4.2% among all methods that entered into analysis across all sphericity levels. Except for ε=0.48, the order of the strength of power was obtained as YBRES > *F* > HF > FM > SB1/ML > FRES > FADF > YBADF across all sphericity levels. *F* (ranging from 17.1% to 20.4%) yielded the highest power estimates among all ANOVA-based methods and the power estimates for YBRES were less than twice those for *F*. When ε=0.48, FM yielded the highest power estimates among (22.9%) all ANOVA-based methods and the power estimate for YBRES (26.7%) was only 3.8% larger than that for FM. Compared with FADF, GG provided smaller power estimates when the data were normal or close to normal but provided larger power estimates when ε=.48 and ε =.75, thus having YBRES > *F* > HF > FM > SB1/ML > FRES > GG > YBADF.

When *n*=60, GG (ranging from 11% to 25%) provided the lowest power estimates among all methods that entered into analysis across all sphericity levels. Except for GG, all other methods provided power estimates of more than 27%. RES provided the equal power estimates as ML and SB1 across all conditions. ADF delivered the highest power estimates levels when ε=1 and ε=.96 and the power estimates came out as 43.2% and 36.4% respectively. The order of the strength of power thus was obtained as ADF > YBRES > RES/SB1/ML > FRES > *F* > HF > BOX > GG. When ε=.75, F became the most powerful method (39.8%) and the order of the magnitude of power estimates came out with *F* > ADF > HF > YBRES > BOX > RES/SB1/ML > FM > FRES > GG with empirical power estimates falling between 20.8% to 39.8%. When ε=0.48, the power estimates provided by RMM methods (ranging from 49.3% to 54.1) were much greater than those provided by ANOVA-based methods (between 25% and 42.8%) with the order of strength of power of YBRES > ADF > RES/SB1/ML > FRES > FADF > BADF > FM > HF > BOX > GG. The power estimate provided by YBRES (54.1%) was more than twice that of GG (25%).

When *n*=100, ADF delivered the highest power estimates (between 48.7% and 72.4%) while GG provided the lowest power estimates (between 25.3% and 41.5%) among all methods that entered into analysis across all sphericity levels except when ε=.75 and the order of the strength of power was obtained as ADF > YBRES > RES/SB1/ML > FADF > HF > BOX > GG. RES provided the equal power estimates as ML and SB1 across all conditions. When ε=.75, *F* provided largest power estimates (56.5%) and the order of the magnitude of power estimates became *F* > FRES > HF > BOX > ADF > YBRES > RES/SB1/ML > FADF > FM > YBADF > GG and these methods provided empirical power estimates between 35.8% and 56.5%.

When *n*=200, generally speaking, RMM methods (ranging from 82.1% to 95.6%) were more powerful than ANOVA-based methods (ranging from 63.3% to 93.6%). GG provided the lowest power estimates across all conditions. The empirical power estimates provided by RMM methods were consistently larger than 80% across all sphericity levels. When ε=.48, all empirical power estimates delivered by RMM methods were greater than 95% while the power estimates provided by ANOVA-based methods except FM were below 80%. The same as *n*=60 and *n*=100, YBRES delivered the highest power among all methods that entered into analysis across all sphericity levels except when ε=.75. When ε=.75, *F* became the most powerful method and provided a power estimate of 95.6%.

Table 14: Empirical Power (%) for Elliptical Distribution with *k*=4

	ε	*F*	Box	HF	GG	FM	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.00	9.5	7.5	8.2	2.2	11.1	13.5		4.3	11.4	13.5		19.9	9.4
	0.96	12.3	9.6	10.9	3.3	8.7	11.4		4.1	9.2	11.4		19.8	7.4
	0.75	13.5	9.7	10.9	4.8	9.2	11.7		3.5	9.3	11.7		19.8	7.5
	0.48		13.1	14.0	9.4	14.4	16.9		7.5	15.3	16.9		28.5	12.9
n=30	1.00	17.3	14.9	16.3	3.9	17.3	18.3		13.1	17.4	18.3	18.3	24.1	16.1
	0.96	20.3	17.9	19.1	6.2	17.3	18.9		14.1	17.4	18.9	18.9	25.6	16.3
	0.75	25.1	21.0	21.3	12.0	20.1	20.8	27.9	15.3	20.5	20.8	20.8	26.9	18.3
	0.48		18.1	19.1	14.0	24.8	26.2		20.9	24.8	26.2	26.2	32.7	23.5
n=60	1.00	32.2	30.7	31.6	14.6	33.2	33.5	38.0	30.7	33.2	33.5	33.5	36.6	32.1
	0.96	34.2	31.8	32.8	14.0	31.6	32.4	35.8	28.7	31.7	32.4	32.4	35.5	31.0
	0.75	36.4	31.2	32.2	19.8	28.8	29.8	34.9	27.0	28.9	29.8	29.8	33.9	27.9
	0.48		30.0	30.4	25.0	49.2	49.6	54.5	47.1	49.3	49.6	49.6	53.4	48.7
n=100	1.00	52.3	51.6	52.0	28.6	52.4	52.4	55.3	51.3	52.5	52.4	52.4	54.9	52.0
	0.96	56.0	54.3	54.7	31.2	55.6	56.2	58.1	53.9	55.6	56.2	56.2	57.7	54.7
	0.75	56.7	51.8	52.4	37.1	46.7	46.9	49.3	45.3	46.7	46.9	46.9	48.7	46.5
	0.48		48.1	48.1	42.6	69.7	70.0	72.8	68.4	69.7	70.0	70.0	72.1	68.9
n=200	1.00	82.1	81.5	81.6	62.9	82.4	82.6	83.2	81.4	82.5	82.6	82.6	83.1	81.9
	0.96	85.8	85.2	85.3	67.5	83.4	83.6	84.0	82.8	83.4	83.6	83.6	83.8	83.2
	0.75	85.7	82.7	82.8	71.4	78.0	78.2	79.2	77.5	78.0	78.2	78.2	78.7	77.9
	0.48		78.3	78.4	74.7	94.6	94.6	95.0	94.3	94.6	94.6	94.6	95.0	94.6

Note. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted *F* test. FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF

= Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled χ^2 test. RES = Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

As is shown in Table 14 for Elliptical Distribution for k=4, the results for some cells were removed from the power analysis because these cells provided Type I error rates larger than 7.5%. Therefore, the results for the traditional F test across all sample sizes when the sphericity assumption was severely violated (ε =.48) and those for ADF and RES when sample size equals 15 across all sphericity levels were removed from the final analysis. Some of the results for ADF were also removed when n=30 (ε =1, ε =.96, and ε =.48) and the results for RES when n=30 and ε =.96 were also removed from the analysis for the same reason.

Generally speaking, when sample sizes increased, all the methods became more powerful. For n=15, power observed in the majority of the cells was less than 20% while the majority of the cells provided power greater than 80% when n=200.

For ANOVA based methods, except for ε=0.48, the order of the strength of power estimates came out with F > HF > BOX > GG and F > FM > GG across all sample sizes. When ε=0.48, with F being removed from comparison, FM provided strongest power and the order of the strength of power estimates came out with FM > HF > BOX > GG across all sample sizes.

For RMM methods, ML and SB1 yielded the same power estimates across all conditions. FADF was consistently more powerful than YBADF and the power estimates provided by YBRES were consistently higher than those by FADF across all conditions (YBRES > FADF > YBADF). The order of the strength of power estimates for YBRES, SB1/ML, and FRES yielded YBRES> SB1/ML >FRES across all sample sizes and sphericity levels. ML/SB1 provided comparatively higher power estimates than FADF across all sample sizes and sphericity levels except when *n*=100, ε=1 and *n*=200, ε=.48. Therefore, except when *n*=100, ε=1 and *n*=200, ε=.48, the order of the magnitude of power estimates came out with YBRES > ML/SB1 > FADF > YBADF across all sample sizes and sphericity levels.

When *n*=15, YBRES provided the highest power estimates (between 19.8% and 28.5) while GG provided the lowest power estimates (between 2.2% and 9.4%) among all methods that entered into analysis across all sphericity levels. Except for ε =0.48, the order of the strength of power was obtained as YBRES > *F* > FRES > YBADF > GG across all other sphericity levels. *F* yielded the highest power estimates among all ANOVA-based methods and the power estimates for YBRES were around twice those for *F*. When ε=0.48, FM yielded the highest power estimates (14.4%) among all ANOVA-based methods and the power estimate for YBRES (28.5%) was around twice that for FM. Both BOX and HF provided larger power estimates than YBADF across all sphericity levels, and BOX was more powerful than FRES except when ε=1, thus having YBRES > SB1/ML > FADF > FM > HF > BOX > YBADF > GG.

When *n*=30, power estimates provided by ANOVA-based methods increased by a greater percentage (at least 50%) than those provided by RMM methods. YBRES provided the highest power estimates (ranging from 24.1% to 32.7%) across all sphericity levels except for ε=.75 and GG provided the smallest power estimates (ranging from 3.9% to 14%) across all conditions and the order of the magnitude of power estimates came out with YBRES > ML/SB1 > FM > FADF > FRES. When ε =.75, ADF became the most powerful method with an estimate of 27.9%, thus having ADF > YBRES > *F* > HF > BOX > ML/RES/SB1 > FM > FRES > YBADF.

When *n*=60, ADF provided the highest power estimates (between 34.9% to 54.5) across all sphericity levels except for ε=.75 and the order of the strength of power was obtained as ADF> YBRES> RES/SB1/ML > BOX > GG. GG provided the smallest power estimates (ranging from 14% to 25%) across all conditions. When ε =.75, *F* (36.4%) became the most powerful and the order of the magnitude of power estimates became *F* > ADF > YBRES > HF > BOX > RES/SB1/ML > FADF > FM > FRES > GG and these methods provided empirical power estimates between 19.8% to 36.4%. When ε=.48, RMM methods provided higher power estimates (between 47.1% and 54.5%) than ANOVA-based methods (between 25% and 49.2%) with the order of strength of power of ADF > YBRES > RES/SB1/ML > FADF > FM > FRES > YBADF > HF > BOX > GG and the empirical power estimates fell in the range between 30% to 54.5%.

When *n*=100, ADF delivered the highest power with estimates between 49.3% to 72.8% while GG provided the lowest power estimates with power estimates between 28.6% and 42.6% among all methods that entered into analysis across all sphericity levels except when ε=.75 and the order of the strength of power was obtained as ADF > YBRES > RES/SB1/ML > HF > BOX > GG. RES provided the equal power estimates as ML and SB1 across all conditions ranging from 46.9% and 70%. When ε=.75, *F* (56.7%) became the most powerful method and the order of the magnitude of power estimates was obtained as *F* > HF > BOX > ADF > YBRES > RES/SB1/ML > FADF/FM > FRES > YBADF > GG and these methods provided empirical power estimates between 37.1% and 56.7%. When ε=.48, RMM methods provided higher power estimates (ranging from 68.4% to 72.8%) than ANOVA-based methods (ranging from 42.6% to 69.7).

When *n*=200, GG provided the lowest power across all conditions with estimates between 62.9% and 74.7%. The majority of the empirical power estimates provided were close to or larger than 80%. When ε=.48, all empirical power estimates delivered by RMM methods were greater than 90% while those provided by ANOVA-based methods except FM were below 80%. YBRES delivered the highest power among all methods that entered into analysis when ε=1 (83.2%) and ε=.48 (95%) and RMM based methods were more powerful than ANOVA based methods. When ε=.96 and ε=.75, *F* became the most powerful method among all the methods that entered into analysis with the power estimates of 85.8% and 85.7% respectively and the order of the magnitude of power estimates came out with *F* > HF >

BOX > ADF > YBRES > ML/SB1/RES > FM > FRES > FADF > YBADF > GG and these methods provided empirical power estimates between 71.4% and 85.7% when ε=.75 and between 67.5% and 85.8% when ε=.96.

Table 15: Empirical Power (%) for Moderately Non-normal Distribution with *k*=4

	ε	*F*	Box	HF	GG	FM	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.00	11.1	7.5	9.8	2.6	10.5	13.3		3.9	10.1	13.3		24.7	9.0
	0.96	8.5	6.4	7.3	1.9	7.2	9.3		2.1	6.9	9.3		17.1	5.9
	0.75	13.5	9.5	10.3	3.9	8.9	10.3			9.4	10.3		18.6	7.8
	0.48		7.7	7.9	4.7				5.6				26.6	
n=30	1.00	17.5	14.8	16.1	4.9	18.4	18.9		14.3	18.7	18.9	18.9	24.6	17.2
	0.96	18.0	14.4	16.1	4.6	14.4	15.6		10.6	14.5	15.6	15.6	22.3	12.8
	0.75	23.7	19.1	19.8	9.9	17.0	17.9		13.3	17.1	17.9	17.9	24.2	15.7
	0.48		12.5	13.1	9.5	22.6	24.1		17.4	18.0	24.1		30.6	21.3
n=60	1.00	32.2	30.4	31.2	13.5	32.8	33.3	38.0	30.3	32.8	33.3	33.3	36.6	32.1
	0.96	35.5	32.7	32.9	14.7	33.9	34.2	38.4	31.4	33.9	34.2	34.2	37.9	32.9
	0.75	40.4	35.3	36.0	21.8	31.2	31.6	35.0	28.5	31.2	31.6	31.6	34.3	30.1
	0.48		25.5	26.4	20.9	44.5	45.1		41.0	44.5	45.1	45.1	49.7	43.4
n=100	1.00	52.0	50.6	51.0	29.0	54.1	54.4	57.0	53.4	54.1	54.4	54.4	56.2	53.5
	0.96	55.0	53.0	53.6	32.8	52.8	53.5	55.8	51.3	53.0	53.5	53.5	55.0	52.6
	0.75	57.6	53.2	53.4	40.3	46.9	47.4	50.4	45.1	46.9	47.4	47.4	49.7	46.4
	0.48		47.2	47.6	39.7	71.6	71.8	74.9	70.7	71.6	71.8	71.8	74.1	71.2
n=200	1.00	83.9	83.3	83.5	64.1	85.1	85.1	86.1	84.7	85.1	85.1	85.1	86.1	84.9
	0.96	85.2	84.2	84.4	68.1	84.6	84.7	85.5	84.1	84.6	84.7	84.7	85.4	84.4
	0.75	87.8	84.5	84.7	72.9	79.2	79.3	80.1	79.0	79.2	79.3	79.3	80.0	79.2
	0.48		84.0	84.0	78.3	97.2	97.2	97.3	97.2	97.2	97.2	97.2	97.3	97.2

Note. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted *F* test. FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF

= Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled χ^2 test. RES = Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

As is shown in Table 15 for Moderately Non-normal Distribution for *k*=4, the results for some cells were removed from the power analysis because these cells provided Type I error rates larger than 7.5%. Therefore, the results for the traditional *F* test across all sample sizes when the sphericity assumption was severely violated (ε =.48) and those for ADF and RES when sample size equaled 15 and those for ADF when sample size equaled 30 across all sphericity levels were removed from the final analysis. The results for SB1, ML, FRES, and FM (*n*=15, ε =.48), the results for RES (*n*=30, ε =.48), and the result for FADF (*n*=15, ε =.48) was also removed from the analysis for the same reason.

Generally speaking, when sample sizes increased, all the methods became more powerful. For *n*=15, power estimates seen in the majority of the cells except YBRES were less than 20% while the majority of the cells showed power greater than 80% when *n*=200.

For ANOVA based methods, except for ε=0.48, the order of the strength of power estimates came out with *F* > HF > BOX > GG and *F* > FM > GG across all sample sizes. When ε=0.48, with *F* being removed from comparison, FM provided strongest power and the order of the strength of power estimates came out with FM > HF > BOX > GG across all sample sizes.

For RMM methods, ML and SB1 yielded the same power estimates across all conditions. FADF is consistently more powerful than YBADF and the power estimates provided by YBRES were consistently higher than those by FADF across all conditions so that the order of magnitude of power estimates was obtained as YBRES > FADF > YBADF. Meanwhile the order of the strength of power estimates for ADF, YBRES, SB1/ML, and FRES yielded ADF > YBRES > SB1/ML > FRES across all sample sizes and sphericity levels. ML/SB1 provided comparatively higher power estimates than FADF across all sample sizes and sphericity levels except when n=200, ε=1 and n=200, ε=.48 (FADF= ML/SB1). Therefore, except when n=200, ε=1 and n=200, ε=.48, the order of the magnitude of power estimates could come out with ADF > YBRES > ML/SB1 > FADF > YBADF across all sample sizes and sphericity levels.

When n=15, YBRES provided the highest power estimates (between 17.1% and 26.6%) while GG provided the lowest power estimates (between 1.9% and 4.7%) among all methods that entered into analysis across all sphericity levels. The order of the strength of power was obtained as YBRES > F > FRES > YBADF > GG across all sphericity levels. F yielded the highest power estimates among all ANOVA-based methods and the power estimates for YBRES were around twice those for F. ML was more powerful than HF and BOX. Both BOX and HF provided larger power estimates than YBADF across all sphericity levels and the order of strength of power became HF > BOX > YBADF > GG.

When n=30, YBRES provided the highest power estimates (ranging from 22.3%

to 30.6%) while GG provided the smallest power estimates (ranging from 4.9% to 9.9%) across all conditions. When ε=1, the order of the strength of power came out with YBRES > SB1/ML/RES > FADF > FM > *F* > FRES > HF > BOX > YBADF ranging from 41.9% to 24.6%. When ε=.96 and ε=.75, the order of the strength of power yielded YBRES > *F* > HF > SB1/ML/RES > FADF > FRES > YBADF > GG with power estimates falling between 4.6% and 22.3% for ε=.96 and between 9.9% and 24.2% for ε=.75. Meanwhile, both HF and BOX were more powerful than FRES. When ε=.48, the order of the magnitudes of power estimates became YBRES > SB1/ML/RES > FM > FRES > FADF > HF > BOX > GG with power estimates falling between 9.5% and 30.6% .

When *n*=60, GG provided the smallest power across all conditions with estimates ranging from 13.5% to 21.8%. ADF provided the highest power estimates when ε=1 and ε=.96 with an estimate of power of 38% and 38.4% respectively and the order of the strength of power was obtained as ADF>YBRES>RES/SB1/ML>HF>BOX>GG. When ε=.75, *F* was the most powerful method with power estimate of 40.4% and the order of the magnitude of power estimates became *F*>HF>BOX>ADF>YBRES>RES/SB1/ML> FM/FADF>FRES>GG and these methods provided empirical power estimates between 21.8% to 40.4%. When ε=0.48, the power estimates provided by RMM methods were all larger than 40% ranging from 41% to 49.7% while the power estimates for all ANOVA-based methods except FM were smaller than 30% ranging from 20.9% to 44.5%.

When *n*=100, GG provided the lowest power estimates (ranging from 29% to 40.3%) among all methods across all conditions while ADF delivered the highest power estimates (ranging from 50.4% to 74.9%) among all methods that entered into analysis across all sphericity levels except when ε=.75 and the order of the strength of power was obtained as ADF>YBRES>RES/SB1/ML>HF>BOX>GG. RES provided the equal power estimates as ML and SB1 across all conditions. When ε =.75, *F* became the most powerful method with an estimate of 57.6% and the order of the magnitude of power estimates became *F* > HF > BOX > ADF > YBRES > RES/SB1/ML > FADF/FM > FRES > YBADF > GG and these methods provided empirical power estimates between 40.3% and 57.6%.

When *n*=200, generally speaking, RMM methods were more powerful than ANOVA-based methods. GG provided the lowest power estimates across all conditions ranging from 64.1% to 78.3%. The majority of the empirical power estimates provided were close to or larger than 80%. When ε=.48, all empirical power estimates delivered by RMM methods were greater than 95% while those provided by ANOVA-based methods except FM were around 80%. YBRES delivered the highest power among all methods that entered into analysis across all sphericity levels except when ε=.75. When ε=.75, *F* became the most powerful method and the order of the magnitude of power estimates became *F* > HF > BOX > ADF > YBRES > RES/SB1/ML > FADF/FM/FRES > YBADF > GG and these methods provided empirical power estimates between 72.9% to 87.8%.

Table 16: Empirical Power (%) for Severely Non-normal Distribution with *k*=4

	ε	*F*	Box	HF	GG	FM	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.00	11.6	7.2	8.5	2.5	10.9	14.3		4.5	10.6	14.3		26.0	9.3
	0.96	10.6	7.7	8.7	2.4	12.4	15.0		3.6	11.3	15.0		25.6	10.3
	0.75	10.5	5.6	7.2	2.5	6.3			3.0	5.5			17.6	5.2
	0.48		4.9	5.6	3.1				5.3				32.3	
n=30	1.00	20.6	16.2	18.0	6.5	22.0	23.3		17.6	22.9	23.3	23.3	29.2	20.2
	0.96	18.7	14.6	15.3	5.6	19.2	20.4		14.3	19.6	20.4	20.4	25.3	17.4
	0.75		11.7	12.8	5.8	14.1	14.9		10.5	14.8	14.9		20.0	12.9
	0.48		9.4	9.9	6.6				24.3			31.0	37.8	
n=60	1.00	35.9	33.6	34.0	15.1	39.3	23.3	45.2	38.0	40.0	23.3	23.3	39.5	20.2
	0.96	36.4	32.6	33.3	17.9	37.4	38.3	44.0	35.1	37.7	38.3	38.3	42.9	36.0
	0.75	38.2	29.3	30.4	18.0	27.6	28.3		25.0	28.0	28.3	28.3	32.5	25.9
	0.48		25.9	25.9	20.0	50.6			47.8	50.9			56.7	49.5
n=100	1.00	50.6	48.6	49.5	27.1	54.7	55.2	58.6	53.5	55.1	55.2	55.2	57.7	54.6
	0.96	58.4	55.7	55.9	32.0	57.8	58.0	61.0	56.5	58.2	58.0	58.0	60.0	57.0
	0.75	60.4	54.4	54.5	37.0	51.8	52.4	56.0	50.5	52.0	52.4	52.4	54.8	51.1
	0.48		50.7	51.2	42.9	77.8			76.0	77.9			79.6	77.5
n=200	1.00	83.6	82.1	82.5	62.7	88.1	88.3	89.4	87.7	88.1	88.3	88.3	89.0	87.7
	0.96	86.9	85.8	85.9	67.8	88.3	88.4	89.0	88.0	88.3	88.4	88.4	88.9	88.3
	0.75	89.4	86.6	86.6	74.5	85.1	85.4	86.0	83.9	85.1	85.4	85.4	86.0	84.6
	0.48		84.0	84.1	77.5	96.8	96.8	97.4	96.8	96.8	96.8	96.8	97.3	96.8

Note. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted *F* test. FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled chi-square test. RES =

Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

As is shown in Table 16 for Severely Non-normal Distribution for *k*=4, the results for some cells were removed from the power analysis because these cells provided Type I error rates larger than 7.5%. Therefore, the results for the traditional *F* test across all sample sizes when the sphericity assumption was severely violated (ε =.48) and those for ADF and RES when sample size equaled 15 and those for ADF when sample size equaled 30 across all sphericity levels were removed from the final analysis. The cells for SB1, ML (*n*=15, ε =.48 and .75; *n*=30, ε =.48; *n*=60, ε =.48; *n*=100, ε =.48), for RES (*n*=30, ε =.75; *n*=60, ε =.48; *n*=100, ε =.48), for FADF , FRES and, FM (*n*=15, ε =.48; *n*=30, ε =.48), for ADF (*n*=60, ε =.48 and .75; *n*=100, ε =.48), and for *F* (*n*=30, ε = .75) were also removed from the analysis for the same reason.

Generally speaking, when sample sizes increased, all the methods became more powerful. For *n*=15, power estimates provided by all the methods except YBRES were less than 20% while the all the methods except GG provided power estimates greater than 80% when *n*=200.

For ANOVA based methods, except for ε=0.48, the order of the strength of power estimates came out with *F* > HF > BOX > GG across all sample sizes. FM, however, provided larger power estimates than *F* in numerous conditions (*n*=15, ε = .96; *n*=30, ε =1 and .96; *n*=60, ε =1 and .96; *n*=60, ε =1 and .96; *n*=100; *n*=200, ε =1 and .96). When ε=0.48, with *F* being removed from comparison, FM provided strongest power and the order of the strength of power estimates came out with FM > HF > BOX > GG.

For RMM methods, ML and SB1 yielded the same power estimates across all conditions. FADF was consistently more powerful than YBADF and the power estimates provided by YBRES were consistently higher than those by FADF across all conditions and thus having YBRES > FADF > YBADF. The order of the strength of power estimates for ADF, YBRES, SB1/ML, and FRES yielded ADF > YBRES > SB1/ML > FRES across all sample sizes and sphericity levels. ML/SB1 provided comparatively higher power estimates than FADF across all sample sizes and sphericity levels except when n=100, ε=.96 and n=200, ε=.48 (FADF= ML/SB1). Therefore, except when n=100, ε=.96 and n=200, ε=.48, the order of the magnitude of power estimates was obtained as YBRES > ML/SB1 > FADF > YBADF across all sample sizes and sphericity levels.

When n=15, YBRES provided the highest power estimates (ranging from 17.6% to 32.3%) while GG provided the lowest power estimates (ranging from 2.4% to 3.1%) among all methods that entered into analysis across all sphericity levels. The order of the strength of power was obtained as YBRES > *F* > FM > FRES > YBADF > GG across all sphericity levels. *F* and sometimes, FM yielded the highest power estimates among all ANOVA-based methods and the power estimates for YBRES were around twice those for *F* and FM. Both BOX and HF provided larger power estimates than

YBADF across all sphericity levels.

When *n*=30, YBRES provided the highest power estimates (between 20% and 37.8) while GG provided the smallest power estimates (between 5.6% and 6.6%) across all conditions. When ε=1, .96, and .75, YBRES, SB1/ML, FADF provided the largest power estimates and the order of strength of power was obtained as YBRES>ML/SB1>FADF>FM>*F*>HF>BOX>GG. FM provided the strongest power among all ANOVA-based methods. When ε=.48, YBRES, RES, and YBADF with power estimates falling above 24.3% were more powerful than all the ANOVA-based methods with power estimates falling below 10%.

When *n*=60, GG provided the smallest power estimates across all conditions with power estimates falling between 15.1% to 20%. ADF was the most powerful method when ε=1 and ε=.96 and provided power estimates of 45.1% and 44% respectively. When ε=1, YBRES, and FADF provided the largest power estimates among all methods, followed by FM, *F,* HM and BOX, with power estimates being larger than 33.6%. The rest methods provided power estimates lower than 30%. When ε=.96, ADF, YBRES, ML/SB1/RES, and FADF provided the largest power estimates which were greater than 37.3%, followed by FM and *F*. When ε=.75, *F* became the most powerful method with a power estimate of 38.2% and the order of the magnitude of power estimates became *F* > YBRES > HF > BOX > RES/SB1/ML > FADF > FM > FRES > YBADF > GG and these methods provided empirical power estimates between 18% and 38.2%. When ε=0.48, YBRES and FADF provided the largest power estimates with power estimates of 56.7% and 50.9% respectively and FM provided the third largest power estimate with an estimate of 50.6% . Except for FM, RMM methods were much more powerful than ANOVA-based methods and the power estimates provided by RMM methods were around two times the estimates provided by ANOVA-based methods.

When *n*=100, GG provided the lowest power estimates among all methods across all conditions with power estimates falling between 27.1% and 42.9%. ADF delivered the highest power when ε=1 and ε=.96 with power estimates being 58.6% and 61% respectively. When ε=1, the power estimates for all RMM methods were over 50%. Among ANOVA methods, only FM and *F* provided power estimates greater than 50%. The power estimate for ADF was more than twice that of GG. The order of magnitude of power estimates became ADF > YBRES > RES/SB1/ML > FM > FRES > YBADF > *F* > HF > BOX > GG and the power estimates ranged from 27.1% to 58.6%. When ε=0.96, the power estimates for all methods except GG were greater than 50% and the power estimate for ADF (61%) was almost the twice that of GG (32%). The order of magnitude of power estimates became ADF > YBRES > *F* > FADF > RES/SB1/ML > FM> FRES > YBADF > HF > BOX > GG. When ε=.75, the power estimates for all methods were greater than 50% except GG, *F* became the most powerful method with a power estimate of 60.4% and the order of the magnitude of power estimates came out with *F* > ADF > YBRES > HF > BOX > ML/SB1/RES > FADF > FM > FRES > GG. When ε=0.48, all the RMM methods that entered into analysis provided power estimates greater than 70%, among which YBRES yielded the largest power estimate of 79.6% and the order of power strength was obtained as YBRES > FADF > FM > HF > BOX > GG. All ANOVA-based methods except FM provided power estimates lower than 52%.

When *n*=200, generally speaking, RMM methods were more powerful than ANOVA-based methods. ADF provided the highest power estimates (ranging from 86% to 97.4%) across all conditions and GG provided the lowest power estimates (ranging from 62.75% to 77.5%) across all conditions except when ε=.75. All methods except GG provided the empirical power estimates larger than 80%. RMM methods tended to yield larger power estimates than ANOVA-based methods except when ε=.75 where *F* became the most powerful method with a power estimate of 89.4% and the order of power strength came out with *F* > HF/BOX > ADF > YBRES > ML/SB1/RES > FM/FADF > FRES > YBADF > GG. When ε=.48, all empirical power estimates delivered by RMM methods were greater than 95% while the power estimates provided by ANOVA-based methods except FM were between 77.5% and 84.1%.

Table 17: Empirical Power (%) for Normal Distribution with *k*=8

	ε	*F*	Box	HF	GG	FM	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.00	8.6	4.4	7.9	0.0	7.8								
	0.96	10.1	5.6	9.7	0.0	7.6								
	0.75	13.7	8.5	11.5	0.4	6.5								
	0.48	14.6	8.7	13.8	0.4	10.9								
n=30	1.00	14.3	11.2	14.0	0.3	11.6								
	0.96	14.2	11.4	13.9	0.4	14.8								
	0.75	21.2	15.0	17.3	1.5	11.4								
	0.48	26.0	20.1	24.8	0.7	24.5								
n=60	1.00	27.8	25.5	27.4	1.5	25.5	29.2		20.5	25.5	29.2		37.0	19.9
	0.96	29.1	26.2	28.3	1.0	29.4	33.7		24.9	29.4	33.7		42.2	23.2
	0.75	34.8	28.2	29.6	5.2	17.6	20.9		14.6	17.6	20.9		29.4	13.3
	0.48	54.9	51.7	54.2	6.1	57.3	61.5		50.2	57.3	61.5		70.2	48.2
n=100	1.00	46.0	44.5	45.9	4.2	42.8	44.7		39.9	42.8	44.7		50.5	39.1
	0.96	44.7	41.3	43.1	4.7	49.2	51.6	60.0	46.1	49.2	51.6	51.6	69.6	45.3
	0.75	58.1	50.6	52.1	14.0	36.0	38.7		32.3	36.1	38.7		43.8	31.3
	0.48	84.8	82.9	84.3	24.5	88.0	88.7		86.0	88.0	88.7		91.3	85.5
n=200	1.00	78.6	78.3	78.6	20.6	77.3	79.1	81.2	76.2	77.3	79.1	79.1	80.5	75.8
	0.96	81.4	80.6	81.1	27.5	87.0	87.5	90.2	85.9	87.0	87.5	87.5	89.9	85.7
	0.75	83.9	79.9	80.2	43.1	66.1	67.0	70.2	65.1	66.1	67.0	67.0	69.8	64.7
	0.48	99.3	99.3	99.3	79.8	99.5	99.5	99.7	99.5	99.5	99.5	99.5	99.6	99.5

Note. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted *F* test. FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled chi-square test. RES =

Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

As is shown in Table 17 for Normal Distribution for k=8, the results for some cells were removed from the power analysis because these cells provided Type I error rates larger than 7.5%. The results for all RMM methods when n=15 and 30 were removed from final analysis. All the results for ADF and RES with sample sizes of 60 and 100 across all sphericity levels except when ε =.96 were removed as well for the same reason.

Generally speaking, when sample sizes increased, all the methods became more powerful. For n=15, power estimates provided by all the methods were less than 15% while the all the methods except GG provided power estimates greater than 60% when n=200.

For ANOVA based methods, the order of the strength of power estimates came out with F > HF > BOX > GG across all sample sizes. FM, however, provided larger power estimates than F in numerous conditions (n=30, ε =.96; n=60, ε =.96 and .48; n=100, ε =.96 and .48; n=200, ε =.96 and .48).

For RMM methods, ML and SB1 yielded the same power estimates across all conditions. FADF was consistently more powerful than YBADF and the power estimates provided by YBRES were consistently higher than those by FADF across all conditions, thus having the order of YBRES > FADF > YBADF. ML/SB1 provided comparatively higher power estimates than FADF across all sample sizes and sphericity levels. Therefore, the order of the magnitude of power estimates came out with YBRES > ML/SB1 > FADF > YBADF > FRES across all sample sizes and sphericity levels.

When *n*=15, no RMM methods entered into analysis. *F* provided the highest power estimates ranging from 8.6% to 14.6% while GG provided the lowest power estimates with power estimates ranging from 0% to 0.4%.

When *n*=30, no RMM methods entered into analysis. *F* provided the highest power estimates (ranging from 14.2% to 26%) except when ε=.96 where FM became the most powerful method with an estimate of 14.8%. GG provided the lowest power estimates with power estimates ranging from 0.3% to 1.5%.

When *n*=60, GG provided the smallest power estimates (ranging from 1% to 6.1%) across all conditions. YBRES was the most powerful across all conditions except when ε=.75 (ranging from 29.4% to 70.2%). When ε=1 and ε=.96, the order of magnitude of power estimates came out to be YBRES > SB1/ML > *F* > HF > BOX > YBADF > FRES > GG with power estimates ranging from 1.5% to 37% and from 1% to 42.2% respectively. Meanwhile, FADF was more powerful than YBADF. The majority of the power estimates was greater than 20%. When ε=.75, *F* yielded the greatest power estimate (34.8%) and ANOVA-based methods tended to be more powerful than RMM methods and the order of magnitude of power estimates became *F* > HF > YBRES > BOX > ML/SB1 > FM/FADF > YBADF > FRES > GG. When ε=.48, RMM methods tended to be more powerful than ANOVA-based methods and the order of strength of power came out to be YBRES > ML/SB1 > FM/FADF > *F* > HF > BOX > YBADF > FRES > GG and power estimates ranged from 6.1% to 70.2%.

When *n*=100, GG provided the lowest power estimates (between 4.2% and 24.5%) among all methods across all conditions. YBRES was the most powerful across all conditions except when ε=.75 with power estimates falling between 43.8% and 91.3%. All methods except GG provided power estimates greater than 30%. When ε=1, the order of magnitude of power estimates came out to be YBRES > *F* > HF > SB1/ML > BOX > YBADF > FRES > GG with power estimates falling between 4.2% to 50.5%. When ε=.96, RMM methods tended to be more powerful than AVOVA-based methods and the order of magnitude of power estimates came out to be YBRES >ADF > SB1/ML/RES > FADF/FM > YBADF > FRES > *F* > HF > BOX > GG, among which YBRES, ADF, SB1/ML/RES yielded power estimates greater than 50%. The power estimates for the rest methods were lower than 50% but greater than 40% except GG. When ε=.75, *F* yielded the greatest power estimate (58.1%) and ANOVA-based methods tended to be more powerful than RMM methods and the order of magnitude of power estimates became *F* > HF > BOX > YBRES > ML/SB1 > FM/FADF > YBADF > FRES > GG. When ε=.48, RMM methods tended to be more powerful than ANOVA-based methods. The power estimates for RMM methods were all greater than 86% and the order of strength of power came out to be YBRES >

ML/SB1 > FM/FADF > YBADF > FRES > *F* > HF > BOX > GG and power estimates ranged from 24.5% to 91.3%.

When *n*=200, GG provided the lowest power estimates among all methods across all conditions between 20.6% and 79.8%. Generally speaking, RMM methods tended to be more powerful than ANOVA-based methods except when $\varepsilon = .75$. Among all methods, ADF provided the greatest power estimates across all conditions except when ε=.75 with power estimates ranging from 70.2% to 99.7%. When ε=1 and .96, the power estimates for all methods except GG were greater than 75%. When ε=.48, the power estimates for all methods except GG were greater than 99%. When ε=.75, the power estimates for all methods except GG were greater than 65% and the order of power strength came out to be *F* > HF > BOX > ADF > YBRES > ML/SB1/RES > FADF/FM > YBADF > FRES > GG, with power estimates falling between 43.1% and 83.9%.

Table 18: Empirical Power (%) for Elliptical Distribution with k=8

	ε	F	Box	HF	GG	FM	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.00	7.7	3.9	5.9	0.1	6.5								
	0.96	9.5	4.4	7.6	0.0	7.1								
	0.75	11.5	5.1	7.6	0.5	6.9								
	0.48		5.6	8.7	0.4	7.3								
n=30	1.00	14.0	9.8	12.5	0.1	11.6	20.2				20.2			
	0.96	15.8	10.7	12.6	0.4	17.5								
	0.75	17.5	11.7	13.1	0.9	11.9								
	0.48		11.9	13.7	1.8	16.7	24.8				24.8			
n=60	1.00	26.7	22.6	24.5	0.8	25.0	30.5		20.8	25.0	30.5		37.0	23.9
	0.96	26.9	21.7	24.1	1.0	32.3	36.5		27.5	32.6	36.5		46.6	31.3
	0.75	31.9	24.7	26.4	3.3	24.4	28.4		20.5	24.4	28.4		36.2	23.7
	0.48		26.9	28.3	5.7	37.7	41.7		32.6	37.7	41.7		51.3	36.5
n=100	1.00	41.5	38.6	40.3	3.7	42.0	44.3		39.9	42.1	44.3	44.3	50.4	41.6
	0.96	47.4	42.9	44.6	5.3	56.8	58.8		53.1	56.8	58.8		62.9	55.7
	0.75	47.7	39.6	41.2	8.2	38.6	40.3		35.1	38.6	40.3	40.3	45.1	37.9
	0.48		43.7	44.4	14.2	59.3	61.4	67.8	56.9	59.5	61.4	61.4	66.6	58.5
n=200	1.00	76.1	74.2	75.2	19.8	76.4	77.3	79.7	75.0	76.5	77.3	77.3	79.5	75.7
	0.96	82.6	81.0	81.4	24.4	90.0	90.4	91.1	89.1	90.0	90.4	90.4	91.1	89.8
	0.75	81.2	76.2	76.6	30.8	71.9	72.7	75.7	70.9	71.9	72.7	72.7	75.5	71.8
	0.48		80.6	80.9	45.5	90.8	91.9	93.3	90.2	90.8	91.9	91.9	92.9	90.7

Note. F = traditional F test. Box = Box's adjusted F test. HF = Huynh-Feldt adjusted F test. GG = Geisser-Greenhouse lower bound adjusted F test. FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled chi-square test. RES =

Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

As is shown in Table 18 for Elliptical Distribution for *k*=8, the results for some cells were removed from the power analysis because these cells provided Type I error rates larger than 7.5%. Therefore, the results for the traditional *F* test across all sample sizes when the sphericity assumption was severely violated (ε =.48) and those for all RMM methods when sample size equaled 15 and those for all RMM methods when sample size equaled 30 across all sphericity levels except for ML and SB1 methods ε =1 and .48 were removed from the final analysis for the same reason.

Generally speaking, when sample sizes increased, all the methods became more powerful. For *n*=15, power estimates provided by all the methods except YBRES were less than 15% while the all the methods except GG provided power estimates greater than 70% when *n*=200.

For ANOVA based methods, except for ε=0.48, the order of the strength of power estimates came out with *F* > HF > BOX > GG across all sample sizes. FM, however, provided larger power estimates than *F* in numerous conditions (*n*=15, ε = .96; *n*=30, ε =.96; *n*=60, ε =1 and .96; *n*=100, ε =1 and .96; *n*=200, ε =1 and .96). When ε=0.48, with *F* being removed from comparison, FM provided strongest power and the order of the strength of power estimates came out with FM > HF > BOX > GG.

For RMM methods, ML and SB1 yielded the same power estimates across all conditions. FADF was consistently more powerful than YBADF and the power estimates provided by YBRES were consistently higher than those by FADF across all conditions, thus yielding YBRES > FADF > YBADF. The order of the strength of power estimates for ADF, YBRES, SB1/ML, and FRES yielded ADF > YBRES > SB1/ML > FRES for all the cells that entered into analysis. ML/SB1 provided comparatively higher power estimates than FADF across all conditions. Therefore, the order of the magnitude of power estimates could come out with YBRES > ML/SB1 > FADF > YBADF across all conditions that entered into final analysis.

When *n*=15, no RMM methods entered into analysis. *F* provided the highest power estimates (between 7.7% and 11.5%) while GG provided the lowest power estimates with power estimates ranging from 0% to 0.5%.

When *n*=30, ML/SB1 became the most powerful methods when ε=1 and .48 with an estimate of 20.2% and 24.8% respectively and all the other RMM methods had been removed from analysis. GG provided the lowest power estimates across all conditions and the power estimates ranged from 0.1% to 1.8%.

When *n*=60, GG provided the smallest power estimates (between 0.8% and 5.7%) while YBRES was the most powerful across all conditions (between 37% and 51.3%). RMM methods tended to be more powerful than ANOVA-based methods. The power estimates for RMM methods ranged from 20.5% to 51.3% while those for ANOVA-based methods ranged from 0.8% to 37.7%. When ε=1 and .96, YBRES, ML and SB1 were three most powerful methods and the order of power strength was obtained as YBRES > ML/SB1 > FADF/FM > FRES > *F* > HF, with power estimates ranging from 0.8% to 37% and from 1% to 46.6% respectively. When ε=.75, the order of power strength yielded YBRES > *F* > ML/SB1 > HF > BOX > FM/FADF > FRES > YBADF > GG with power estimates falling between 3.3% and 36.2%. When ε=.48, RMM methods were more powerful than all ANOVA-based methods and the order of strength of power came out to be YBRES > ML/SB1 > FM/FADF > FRES > YBADF > HF > BOX > GG and power estimates ranged from 5.7% to 51.3%. All RMM methods provided power estimates greater than 30% while all ANOVA-based methods except FM provided power estimates lower than 30%.

When *n*=100, GG provided the lowest power estimates among all methods across all conditions (between 3.7% and 14.2%). YBRES was the most powerful across all conditions when ε=1 and .96 with power estimates of 50.4% and 62.9% respectively. All methods except GG provided power estimates greater than 30%. The order of power strength yielded YBRES > ML/SB1/RES > FADF > FM > FRES > *F* > HF > BOX > GG and the power estimates fell between 3.7% and 50.4% for ε=1 and between 5.3% and 62.9% for ε=.96. When ε=.75, *F* yielded the greatest power estimate (47.7%) and the order of magnitude of power estimates became *F* > YBRES > HF > ML/SB1/RES > BOX > FM/FADF > FRES > YBADF > GG. When ε=.48, RMM methods were more powerful than ANOVA-based methods. The power estimates for RMM methods were all greater than 55% and the order of strength of power came out to be ADF > YBRES > ML/SB1 > FADF > FM > FRES > YBADF > HF > BOX > GG and power estimates ranged from 14.2% to 67.8%.

When *n*=200, GG provided the lowest power estimates among all methods across all conditions (between 19.8% and 45.5%). Generally speaking, RMM methods tended to be more powerful than ANOVA-based methods except when ε=.75. Among all methods, ADF provided the greatest power estimates across all conditions except when ε=.75 (between 75.7% and 93.3%). When ε=1, .96 and 0.48, the power estimates for all methods except GG were greater than 74% and the order of the magnitudes of power estimates became ADF > YBRES > ML/SB1/RES > FADF/FM > FRES > YBADF > *F* > HF > BOX > GG. When ε=.75, the power estimates for all methods except GG were greater than 70% and the order of power strength came out to be *F* > HF > BOX > ADF > YBRES > ML/SB1/RES > FADF/FM > FRES > YBADF > GG with power estimates ranging from 30.8% and 81.2%.

Table 19: Empirical Power (%) for Moderately Non-normal Distribution with *k*=8

	ε	*F*	Box	HF	GG	FM	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.00	7.8	2.5	4.8	0.2	7.2								
	0.96	8.9	3.0	6.4	0.0	10.0								
	0.75	12.0	4.2	6.8	0.3	6.8								
	0.48		5.1	6.9	1.0									
n=30	1.00	13.6	8.2	10.9	0.3	14.2								
	0.96	13.5	8.2	10.8	0.2	17.0								
	0.75	17.3	8.8	11.0	0.5	10.3								
	0.48		8.6	9.6	0.9									
n=60	1.00	25.4	20.7	22.8	0.8	28.9	32.7		22.3	29.0	32.7		42.0	27.5
	0.96	25.8	20.4	22.6	1.1	31.1	35.7		25.7	31.3	35.7		45.4	30.3
	0.75	29.6	21.5	23.4	2.7	18.2			14.9	18.3			31.3	17.8
	0.48		20.9	21.8	4.4				31.9				49.5	35.9
n=100	1.00	42.5	38.9	40.2	4.0	44.1	47.5		40.3	44.2	47.5		53.1	43.3
	0.96	45.7	41.2	42.6	3.6	56.5	59.2		51.5	56.5	59.2	59.2	66.1	56.0
	0.75	49.2	41.3	42.4	7.6	34.2	37.4		30.3	34.3	37.4		41.3	33.3
	0.48		40.7	41.7	11.4				54.9				66.5	
n=200	1.00	77.4	76.2	76.7	20.9	81.0	81.9	84.2	79.8	81.0	81.9	81.9	84.0	80.7
	0.96	82.8	79.8	80.5	21.2	92.4	92.7	93.4	91.6	92.4	92.7	92.7	93.4	92.0
	0.75	79.2	73.0	73.4	30.5	71.7	72.5	76.3	70.1	71.7	72.5	72.5	75.6	71.1
	0.48		76.8	77.6	42.1		92.1	93.4	91.0	91.4	92.1	92.1	93.2	91.4

Note. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted *F* test. FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled chi-square test. RES =

Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

As is shown in Table 19 for Moderately Non-normal Distribution for *k*=8, the results for some cells were removed from the power analysis because these cells provided Type I error rates larger than 7.5%. Therefore, the results for the traditional *F* test and FM across all sample sizes when the sphericity assumption was severely violated (ε =.48) and those for all RMM methods when sample size equaled 15 and 30 were removed from the final analysis. All cells for ADF when *n*=60 and 100, for RES when *n*=60, for RES when n=100 except when ε = .96 were also removed from the analysis. The results for ML and SB1 (*n*=60, ε = .75 and .48, *n*=100, ε = .48) and those for FRES and FADF (*n*=100, ε = .48) were also eliminated for the same reason.

Generally speaking, when sample sizes increased, all the methods became more powerful. For *n*=15, power estimates provided by all the methods except YBRES were less than 15% while the all the methods except GG provided power estimates greater than 70% when *n*=200.

For ANOVA based methods, except for ε=0.48, the order of the strength of power estimates came out with *F* > HF > BOX > GG across all sample sizes. FM, however, provided larger power estimates than *F* in numerous conditions (*n*=15, ε = .96; *n*=30, ε =1 and .96; *n*=60, ε =1 and .96; *n*=100, ε =1 and .96; *n*=200, ε =1 and .96). When ε=0.48, with *F* being removed from comparison, FM provided strongest power and the order of the strength of power estimates came out with FM >

HF > BOX > GG.

For RMM methods, ML and SB1 yielded the same power estimates across all conditions. FADF was consistently more powerful than YBADF and the power estimates provided by YBRES were consistently higher than those by FADF across all conditions. ML/SB1 provided comparatively higher power estimates than FADF among all the cells available for analysis. Therefore, the order of the magnitude of power estimates could come out with YBRES > ML/SB1 > FADF > FRES > YBADF across all the conditions that entered in to analysis.

When *n*=15, no RMM methods entered into analysis. *F* provided the highest power estimates except when ε=.96 (ranging from 7.8% to 12%) while GG provided the lowest power estimates with power estimates falling between 0.2% and 1%. When ε=.96, FM became the most powerful method with a power estimate of 10%.

When *n*=30, no RMM methods entered into analysis. GG provided the lowest power estimates (between 0.2% and 0.9%). FM provided the highest power estimates when ε=1 and .96 with power estimates being 14.2% and 17% respectively and *F* was the most powerful method when ε=.75 with an power estimate of 17.3%.

When *n*=60, GG provided the smallest power estimates across all conditions (between 0.8% and 4.4%) while YBRES was the most powerful across all conditions (between 31.3% and 49.5%). When ε=1 and ε=.96, RMM methods tended to more powerful than ANOVA-based methods and the power estimates provided by all methods except GG were greater than 20%. The order of magnitude of power estimates came out to be YBRES > SB1/ML > FADF > FM > FRES > *F* > HF > BOX > GG with power estimates falling between 0.8% and 42% for ε=1 and between 1.1% and 45.4% for ε=.96. When ε=.75, the order of magnitude of power estimates became YBRES > *F* > HF > BOX > FADF > FM > FRES > GG with power estimates ranging from 2.7% to 31.3%. When ε=.48, RMM methods tended to be more powerful than ANOVA-based methods and the order of strength of power came out to be YBRES > FRES >YBADF > HF > BOX > GG and power estimates ranged from 4.4% to 49.5%.

When *n*=100, GG provided the lowest power estimates among all methods across all conditions (ranging from 1% to 11.4%). YBRES was the most powerful across all conditions except when ε=.75 (ranging from 41.3% to 66.5%). All methods except GG provided power estimates greater than 30%. When ε=1 and .96, RMM methods tended to be more powerful than ANOVA-based methods. The power estimates for RMM methods ranged from 43.3% to 66.1% and those for ANOVA-based methods ranged from 38.9% to 45.7%. The order of magnitude of power estimates turned out to be YBRES > SB1/ML/RES > FADF > FM > FRES > *F* > HF > BOX > GG. When ε=.75, *F* yielded the greatest power estimate (49.2%) and ANOVA-based methods tended to be more powerful than RMM methods and the order of magnitude of power estimates became *F* > HF > BOX/YBRES > ML/SB1 >

FADF > FM > FRES > YBADF > GG. When ε=.48, RMM methods tended to be more powerful than ANOVA-based methods. The power estimates for RMM methods were all greater than 50% and the order of strength of power came out to be YBRES > YBADF > HF > BOX > GG and power estimates ranged from 11.4% to 66.5%.

When n=200, GG provided the lowest power estimates among all methods across all conditions (between 20.9% and 42.1%). Generally speaking, RMM methods tended to be more powerful than ANOVA-based methods when ε=1, .96 and .48 and ANOVA-based methods tended to be more powerful than RMM methods when ε =.75. Among all methods, ADF provided the greatest power estimates across all conditions except when ε=.75 and the power estimates fell between 76.3% and 93.4%. When ε=1 and .96, the power estimates for all methods except GG were greater than 75% and the order of power strength came out with ADF > YBRES > ML/SB1/RES > FADF/FM > YBADF > FRES > F > HF > BOX > GG and the power estimates fell between 20.9% and 84.2% for ε=1 and between 21.2% and 93.4% for ε=.96 respectively. When ε=.48, the power estimates for all methods except GG were greater than 90% and the order of magnitude of power estimates was obtained as ADF > YBRES > ML/SB1/RES > FADF > YBADF > FRES > HF > BOX > GG. When ε=.75, the power estimates for all methods except GG were greater than 65% and the order of power strength came out to be F > ADF > YBRES > HF > BOX > ML/SB1/RES > FADF/FM > FRES > YBADF > GG and power estimates ranged from 30.5% to 79.2%.

Table 20: Empirical Power (%) for Severely Non-normal Distribution with *k*=8

	ε	F	Box	HF	GG	FM	ML	ADF	YBADF	FADF	SB1	RES	YBRES	FRES
n=15	1.00	7.5	2.4	3.8	0.1	9.9								
	0.96	7.8	1.8	3.2	0.1	9.9								
	0.75	9.9	2.9	4.5	0.2	7.6								
	0.48		4.0	6.4	0.5	17.3								
n=30	1.00	13.2	7.2	8.9	0.1	16.3	22.3				22.3			
	0.96	15.3	7.0	9.7	0.1	17.6	27.9				27.9			
	0.75	16.8	8.5	9.6	1.1	10.3								
	0.48		7.3	8.7	0.8	20.8								
n=60	1.00	23.4	17.0	19.2	1.0	28.9	33.5		25.2	30.9	33.5		42.2	28.7
	0.96	24.7	17.4	19.2	0.6	37.9	43.4		33.2	39.0	43.4		54.0	36.9
	0.75	26.8	17.3	19.1	2.6	18.8			14.0				29.9	17.8
	0.48		20.3	21.7	3.5	37.3			31.5				52.2	
n=100	1.00	44.0	38.7	39.3	2.6	50.3	49.9		44.3	47.9	49.9		56.1	46.6
	0.96	43.4	38.2	39.6	3.8	58.3	61.1		56.0	59.2	61.1		66.7	57.3
	0.75	45.4	36.6	37.9	7.7	36.2			33.4	36.8			46.1	35.1
	0.48		38.3	39.9	8.3	53.5							61.5	
n=200	1.00	73.6	70.3	71.4	17.8	81.4	82.7	85.2	79.9	81.1	82.7	82.7	85.0	80.6
	0.96	78.7	74.1	75.0	18.7	90.2	90.6	92.7	89.8	90.3	90.6	90.6	92.3	90.0
	0.75	76.4	71.0	71.5	28.1	67.2	68.3		66.3	67.5	68.3		71.6	66.7
	0.48		77.2	77.5	39.7	88.1	88.5		87.3	88.1	88.5		90.6	88.0

Note. *F* = traditional *F* test. Box = Box's adjusted *F* test. HF = Huynh-Feldt adjusted *F* test. GG = Geisser-Greenhouse lower bound adjusted *F* test. FM = one-sample multivariate T^2 test. ML = the maximum likelihood method. ADF = Browne's asymptotic distribution-free test. YBADF = Yuan and Bentler adjusted ADF I test. FADF = Yuan and Bentler adjusted ADF II test. SB1 = Satorra-Bentler scaled chi-square test. RES =

Residual-based ADF test. YBRES = Yuan and Bentler adjusted RES I test. FRES = Yuan and Bentler adjusted RES II test.

As is shown in Table 20 for Severely Non-normal Distribution for k=8, the results for some cells were removed from the power analysis because these cells provided Type I error rates larger than 7.5%. Therefore, the results for the traditional F test across all sample sizes when the sphericity assumption was severely violated (ε =.48) and those for all RMM methods when sample size equaled 15 and those for all RMM methods when sample size equaled 30 across all sphericity levels except for ML and SB1 methods ε =1 and .96 were removed from the final analysis. All cells for ADF and RES across all sample sizes except when n=200, ε = 1 and .96, were also removed from the analysis. The results for ML and SB1 (n=60, ε = .75 and .48; n=100, ε = .75 and .48) and those for FADF (n=60, ε = .75 and .48; n=100, ε = .48) , for YBADF (n=100, ε = .48), and for FRES (n=60, ε = .48; n=100, ε = .48) were also eliminated for the same reason.

Generally speaking, when sample sizes increased, all the methods became more powerful. For n=15, power estimates provided by all the methods except YBRES were less than 10% while the all the methods except GG provided power estimates greater than 65% when n=200.

For ANOVA based methods, except for ε=0.48, the order of the strength of power estimates came out with F > HF > BOX > GG across all sample sizes. FM, however, provided larger power estimates than F in numerous conditions (ε = 1 and .96 across all sample sizes). When ε=0.48, with F being removed from comparison, FM provided strongest power and the order of the strength of power estimates came out with FM > HF > BOX > GG.

For RMM methods, ML and SB1 yielded the same power estimates across all conditions. FADF was consistently more powerful than YBADF and the power estimates provided by YBRES were consistently higher than those by FADF across all conditions. ML/SB1 provided comparatively higher power estimates than FADF across all sample sizes. Therefore, the order of the magnitude of power estimates came out with YBRES > ML/SB1 > FADF > FRES > YBADF across all the conditions entering into final analysis.

When *n*=15, no RMM methods entered into analysis. FM provided the highest power estimates (ranging from 7.6% to 17.3%) except when ε=.75 where *F* became the most powerful method with a power estimate of 9.9%, while GG provided the lowest power estimates with power estimates ranged from 0.1% to 0.5%.

When *n*=30, ML/SB1 became the most powerful methods when ε=1 and .96 with an estimate of 22.3% and 27.9% respectively and all the other RMM methods had been removed from analysis. GG provided the lowest power estimates across all conditions and the power estimates ranged from 0.1% to 1.1%. When ε=.75, *F* was the most power method with an estimate of 16.8%. When ε=.48, with *F* being removed from analysis, FM became the most powerful method with an estimate of 20.8%.

When *n*=60, GG provided the smallest power estimates (between 0.6% and 3.5%) and YBRES was the most powerful (between 29.9% and 54%) across all conditions. RMM methods tended to be more powerful than ANOVA-based methods. When ε=1 and ε=.96, the order of magnitude of power estimates came out to be YBRES > SB1/ML > FADF > FM > FRES > YBADF > *F* > HF > BOX > GG. The power estimates for RMM methods ranged from 25.2% to 54% while those for ANOVA-based methods ranged from 17% to 37.9%. When ε=.75, the order of magnitude of power estimates became YBRES > *F* > HF > FM > FRES > BOX > YBADF > GG with power estimates ranging from 2.6% to 292.9%. When ε=.48, the order of strength of power came out to be YBRES > FM > YBADF > HF > BOX > GG and power estimates ranged from 3.5% to 52.2%.

When *n*=100, GG provided the lowest power estimates (between 2.6% and 8.3%) and YBRES was the most powerful among all methods (between 46.1% and 61.5%) across all conditions. All methods except GG provided power estimates greater than 30%. When ε=1 and ε=.96, RMM methods tended to be more powerful than ANOVA-based methods and the order of power strength yielded YBRES > ML/SB1 > FADF > FRES > YBADF > *F* > HF > BOX > GG and the power estimates fell between 2.6% and 56.1% for ε=1 and between 3.8% and 66.7% for ε=.96. When ε=.75, the order of magnitude of power estimates became YBRES > *F* > HF > FADF > BOX > FM > FRES > YBADF > GG and the power estimates ranged from 7.7% to 46.1%. When ε=.48, the order of strength of power came out to be YBRES > FM > HF > BOX > GG with power estimates ranging from 8.3% to 61.5%.

When *n*=200, GG provided the lowest power estimates (between 17.8% and 39.7%) among all methods at all combinations of distributional forms and sphericity levels. Generally speaking, RMM methods tended to be more powerful than ANOVA-based methods when ε=1, .96, and .48 and ANOVA-based methods tended to be more powerful than RMM methods when ε=.75. Among all methods, ADF provided the greatest power estimates when ε=1 and .96 with power estimates of 85.2% and 92.7% respectively. When ε=1 and .96, the power estimates for all methods except GG were greater than 70% and the order of power strength came out with ADF > YBRES > ML/SB1/RES > FM > FRES > *F* > HF > BOX > GG with power estimates falling between 17.8% and 85.2% for ε=1 and between 18.7% and 92.3% for ε=.96. When ε=.75, *F* became the most powerful method and the power estimates for all methods except GG were greater than 65%. The order of power strength came out to be *F* > YBRES > HF > BOX > ML/SB1 > FADF > FM > FRES > GG with power estimates falling between 28.1% and 76.4%. When ε=.48, RMM methods tended to be more powerful than ANOVA-based methods. The power estimates for RMM methods were greater than 87% and the order of power strength came out to be YBRES>ML/SB1>FADF/FM>FRES>HF>BOX>GG with power estimates ranging from 39.7% to 90.6%.

Chapter 5: Discussion and Conclusion

The current study was focusd on comparing the ANOVA-based methods and RMM methods to determine the best methods for repeated measures designs. The results of the comparision were summarized first, followed by discussion. The conclusion was then drawn and recommendations were made.

None-convergence Summary

In the current study, ANOVA-based methods examined were not encountered with any non-convergence issue. Among RMM methods, only ADF and RES as well as the methods derived based on them including YBADF, FADF, YBRES, and FRES yielded non-convergence. The non-convergence rates tended to get greater when the violation of normality assumption became more seriously. When sample size got bigger, the non-convergence rates became smaller. RES-based methods (RES, YBRES, and FRES) performed better than ADF-based methods (ADF, YBADF, and FADF) in terms of convergence rates. The reason why these methods encountered non-convergence should be tied to the fact that these methods needed to employ the inverse of the fourth-order moments of the measured variables to compute parameter estimates, standard errors, as well as test statistics (Satorra & Bentler, 1994), which put smaller samples under very challenged situation.

Type I Error Rates Results Summary

When *k*=4, among all the ANOVA-based methods, generally speaking, the Type

I error rates provided by Geisser-Greenhouse lower bound adjusted *F* test (GG) tended to be below the lower bound of Bradley's liberal criterion (2.5%). On the contrary, the β -trimmed method using β = 0.2 (TR) tended to provide inflated Type I error rates that were beyond the upper bound of Bradley's liberal criterion (7.5%). For *F* test, as the degree of deviation from sphericity increased, the Type I error rates had a tendency of becoming increasingly inflated. When the sphericity assumption was seriously violated (ε=.48), *F* test was not robust any longer. However, under most circumstances, across different sample sizes and different sphericity levels and distributional shapes, the Box's adjusted *F* test (BOX) and Huynh-Feldt adjusted *F* test could provide robust Type I error rates with only a few exceptions where the Type I error rates were below the lower bound of Bradley's liberal criterion.

The majority of the RMM methods, encouragingly, did provide comparatively robust Type I error rates across different sample sizes and different sphericity levels and distributional shapes especially when the sample size was over 15, with the notable exceptions of Browne's asymptotic distribution-free test (ADF) and residual-based ADF test (RES) which started to perform well when the sample size was over 60. The maximum likelihood method (ML) provided robust Type I error rates when the distribution was normal, or elliptically nonnormal with skewness and kurtosis of (0,7), or moderately nonnormal with skewness and kurtosis of (2, 7).

When the distribution was severely nonnormal with skewness and kurtosis of (3, 21), ML started to yield inflated Type I error rates when the sphericity assumption was seriously violated (ε=.48) (Note that non-normality can come from discreteness, too, such as Likert scales.) Derived from ML, Satorra-Bentler scaled χ^2 test (SB1) yielded similar Type I error rates and mirrored ML closely. Yuan and Bentler adjusted ADF I test (YBADF) and Yuan and Bentler adjusted RES I test (YBRES) could provide robust Type I error rates when the sample size was over 15 under most circumstances with only a few exceptions where the Type I error rates were below the lower bound of Bradley's liberal criterion. Yuan and Bentler adjusted ADF II test (FADF) and Yuan and Bentler adjusted RES II test (FRES), however, could provide robust Type I error rates when the sample size was over 15 under most circumstances with only a few exceptions where the Type I error rates were inflated and beyond the upper bound of Bradley's liberal criterion. Both ADF and RES provided inflated Type I error rates when sample size equaled 15 and 30 and started to perform well when sample size was over 60.

When *k*=8, among all the ANOVA-based methods, the Type I error rates provided by GG were close to 0. On the contrary, TR tended to provide greatly inflated Type I error rates that were beyond the upper bound of Bradley's liberal criterion. For *F* test, as the degree of deviation from sphericity increased, the Type I error rates tended to become increasingly inflated. When the sphericity assumption was seriously violated (ε=.48), *F* test provided inflated Type I error rates that were off the range of robustness. Under most circumstances across different sample sizes and different sphericity levels and distributional shapes, both BOX and HF were robust except when the distribution was severely nonnormal with skewness and kurtosis of (3, 21) and sample size was small (*n*=15, 30) where their Type I error rates were lower than the lower bound of Bradley's liberal criterion.

The majority of the RMM methods, encouragingly, did provide comparatively robust Type I error rates across different sample sizes and different sphericity levels and distributional shapes especially when the sample size was over 60, except ADF and RES which were robust only when the sample size was 200, which was in line with previous findings. ML, SB1, FADF, and FRES provided robust Type I error rates when the distribution was normal, or elliptically nonnormal with skewness and kurtosis of (0, 7) or when the distribution was moderately nonnormal with skewness and kurtosis of (2, 7) or severely nonnormal with skewness and kurtosis of (3, 21) when the sphericity assumption was met or slightly violated (ε =.96). When the sphericity assumption was seriously violated (ε=.48) and the distribution was moderately nonnormal or severely nonnormal, these methods started to yield inflated Type I error rates. All these methods were robust when the sample size was 200. YBADF and YBRES could provide robust Type I error rates when the sample size was over 60 across different sample sizes and different sphericity levels and distributional shapes.

Empirical Power Results Summary

Generally speaking, RMM methods were more powerful than ANOVA-based methods with only a few exceptions. Since those cases where the Type I error rates were greater than the upper boundary of Bradley's liberal criterion were removed from power analysis, the tables 13-20 didn't show all the results of the empirical power estimates. If those cases were shown in the tables, it can be seen that RMM provided the largest power estimates in most cases except that TR did exceed RMM methods in numerous cases. But with those cased being removed, RMM still took the lead in terms of magnitude of power estimates.

When k=4, except that most cells for ADF when sample sizes were small (n=15 and 30) were removed due to inflated Type I error rates, ADF yielded the greatest power estimates under most circumstances where ADF entered into analysis. When ADF was removed, YBRES became the most powerful method in the majority of conditions. One of the exceptions was that F tended to be the most powerful method when ε=.75 when sample size got larger (n=60, 100, 200). Under the majority of the conditions, ADF, YBRES, ML/SB1, FADF, and RES provided largest power estimates among all methods.

When k=8, all asymptotic distribution-free tests including ADF, YBADF, FADF, RES, FRES, and YBRES didn't converge when sample sizes were small (n=15 and

30). Meanwhile, the majority of the cells of RMM methods were removed when sample sizes were 15 and 30 due to inflated Type I error rates. ADF yielded the greatest power estimates under most circumstances where ADF entered into analysis. When ADF was removed, YBRES became the most powerful method across the majority of conditions. One of the exceptions was that *F* tended to be the most powerful method when ε=.75 when sample size got larger (*n*=60, 100, 200). Under the majority of the conditions when sample sizes were larger than 60, ADF, YBRES, ML/SB1, FADF, and RES provided largest power estimates among all methods.

Discussion and Conclusion

Some of the key findings of the current study include the following:

- Among the three distribution-free methods, ADF performed very well when the sample size was large, but if the sample size was small to moderate and also if the model was complex, it was not able to control Type I error very well and tended to over-reject the correct models. When *k*=4, ADF tended to be the most powerful method when sample size was 60 or above and was able to control Type I error rates well. But when *k*=8 where the model became more complex, ADF became the most powerful method only when sample size was 200 across majority of distribution forms and sphericity levels . As discussed earlier, when the sample size was larger than 60, the distribution-free methods did not encounter non-convergence issues. But when the sample sizes were not as large as 200, ADF tended to provide inflated Type I error rates and thus was removed from power analysis in the majority of conditions. On the other hand, the two adjusted methods FADF and YBADF proposed by Yuan and Bentler (1997, 1999) were more conservative and performed much better than ADF in terms of controlling Type I error rates. YBADF tended to over correct the inflation of ADF and yielded the smallest Type I error rates among the three methods. As sample size increased, the Type I error rates provided by the three methods got closer. At the same time, YBADF and FADF yielded adequate power but FADF outperformed YBADF over a range of conditions. This result matched up with findings from previous studies (Yuan & Bentler, 1997, 1999; Nevitt & Hancock, 2004). Therefore, ADF was not recommended when the sample size was small and model was complex.

- Among the three residual-based methods, similar to ADF, RES performed comparably well when the sample size was large but if the sample size was small to moderate, it provided inflated Type I error rates, thus rejecting correct models far too frequently. On the other hand, the two adjusted methods FRES and YBRES proposed by Bentler and Yuan (1999) were more conservative and able to correct the over-rejection of RES for correct models in finite samples. YBRES tended to over-correct the inflation of ADF and yielded the smallest Type I error rates among the three methods. As sample size increased, the Type I error rates provided by the three methods got closer. This finding was also in line with previous studies (Yuan & Bentler, 1998; Bentler and Yuan, 1999; Nevitt & Hancock, 2004). However, among the three methods, YBRES behaved most stably across different conditions and provided the largest power estimates, thus making YBRES very recommendable. This finding was different from the results of Yuan and Bentler (1998) and Bentler and Yuan (1999), which indicated that FRES performed better than YBRES and was recommended as the first choice by Bentler and Yuan (1999). However, the current study also showed that YBRES performed better than FRES under most circumstances except some rare cases, which makes YBRES even more attractive. At the same time, YBRES also consistently outperformed FADF, which distinguished YBRES as the most recommendable method among all the RMM methods.

- Some previous studies (e.g., Wilcox, 1993, 1997, 1998) concluded that the trimmed mean method (TR) was more powerful than the traditional *F* test. This finding was also supported by the current study. As a matter of fact, TR was more powerful than all the other ANOVA-based methods that entered into analysis. However, TR was not able to control Type I error well and yielded inflated Type I error rates under the majority of conditions, which was different from the finding from Berkovits et al. (2000) that the trimmed mean method was able to control Type I error rate well. And considering the fact that TR actually tested the equality of the population trimmed means instead of population means, and thus modified the null hypothesis, this method is not recommended based on the findings of this study.

- Thanks to the simplicity in calculation, the GG lower bound adjustment method has been used widely. However, the current study indicated that GG was very conservative in controlling Type I error rates and provided the smallest power estimates among all methods across almost all conditions. GG performed especially poorly when the number of the levels was large (*k*=8), providing power estimates that were almost 1/30 the magnitude the power estimates other methods provided. Therefore, GG is not recommended.
- ML's performance was affected by the increasing departure from multivariate normality, decreasing sample size, and the increasing number of repeat measures as well as the degree of violation of sphericity assumption. ML tended to produce inflated Type I error when distribution

form deviated from normality with small sample size and when there was severe violation of sphericity. When the number of repeated measures increased to 8 (*k*=8), ML started to provide inflated Type I error rates much more frequently when sample size was small and when the sphericity assumption was seriously violated. The same also held true to ML's adjusted method, SB1. Therefore, ML and SB1 are only recommended when there is no or little violation of sphericity assumption but not recommended when the model is complex (*k*=8) and the sample size is small (15, 30).

- Both BOX and HF provided a small number of attenuated Type I error rates when sample size was small and distribution was non-normal. Though they were not as powerful as RMM methods in most cases, they were able to provide robust Type I error rates under most conditions including when the sphericity assumption and normality assumption were severely violated. FM, on the other hand, provided inflated Type I error rates with non-normal distribution and severe violation of sphericity assumption and it tended to be more powerful than *F* when the distribution deviated more from normality except when ε=.75. Therefore, HF should be preferred among ANOVA-based methods if the distribution was severely non-normal and sphericity assumption was severely violated since HF was consistently more powerful than BOX.

Generally speaking, this study suggests that RMM methods tended to be more powerful than ANOVA-based methods. It is worth noting that sphericity levels, sample sizes, as well as the number of repeated measures, could be used as the baseline for practitioners to decide which method to choose when they encounter real data. Table 21 shows the best methods to be recommended under each condition. As mentioned previously, the cells that produced inflated Type I error rates had been removed from power analysis. Therefore, the methods recommended for each distribution were identified based on the power estimates each method provided which entered into the final analysis. Tables 21-24 present the recommended methods for each of the four distribution conditions investigaged in the current study: normal distribution, elliptical distribution, moderately non-normal distribution, and severely non-normal distribution. In these tables, sample sizes were combined into two categories (15, 30) and (60, 100, 200) as RMM methods were encountered with some non-convergence issues when the sample size was smaller than 60. For each condition, five methods with the greatest power estimates were identified and then the common methods were picked based on the sample size category. The methods shown in the tables were ordered based on the magnitudes of power estimates; that is, the first method was the one with highest power estimate among all the methods listed for each condition.

Table 21

Recommended Methods for Normal Distribution

Degree of Non-sphericity	Number of Levels	Sample Size	Most Powerful/Recommended Methods
1	*K*=4	15, 30	YBRES, ML/SB1, *F*
		60, 100, 200	ADF, YBRES, ML/SB1/RES
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	YBRES, ML/SB1, *F*, HF
0.96	*K*=4	15, 30	YBRES, ML/SB1, *F*
		60, 100, 200	ADF, YBRES, ML/SB1/RES
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	YBRES, ML/SB1
0.75	*K*=4	15, 30	YBRES, *F*, HF
		60, 100, 200	*F*, YBRES
	K=8	15, 30	*F*, FM, HF
		60, 100, 200	*F*, HF, BOX, YBRES
0.48	*K*=4	15, 30	YBRES, ML/SB1, FM
		60, 100, 200	YBRES, ADF, ML/SB1/RES
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	YBRES, ML/SB1, FADF/FM

Table 21 shows that the majority of the methods recommended were RMM methods. Except for k = 8 with small sample sizes (15, 30), YBRES was recommended for all conditions. *F*, however, was mostly likely to be recommended when sample size was small. When the sphericity assumption was seriously violated (ε =.48), *F* was not recommended. This finding is interesting and of great significance as it has long been accepted that it is safe to use *F* when normality assumption is not violated. This study, however, shows that *F* is not a good choice for normal data when the sphericity assumption is seriously violated.

Table 22

Recommended Methods for Elliptical Distribution

Degree of Non-sphericity	Number of Levels	Sample Size	Most Powerful/Recommended Methods
1	*K*=4	15, 30	YBRES, ML/SB1
		60, 100, 200	ADF, YBRES, ML/SB1/RES
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	YBRES, ML/SB1
0.96	*K*=4	15, 30	YBRES, ML/SB1, *F*, HF
		60, 100, 200	ADF, YBRES
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	YBRES, ML/SB1
0.75	*K*=4	15, 30	YBRES, F, HF
		60, 100, 200	*F,* HF, BOX, ADF, YBRES
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	*F*, YBRES
0.48	*K*=4	15, 30	YBRES, ML/SB1, FADF, FM
		60, 100, 200	ADF, YBRES, ML/SB1/RES
	K=8	15, 30	FM, HF, BOX
		60, 100, 200	YBRES, ML/SB1

Table 22 shows that the majority of the methods recommended were RMM methods. Except for k = 8 with small sample sizes (15, 30), YBRES was recommended for all conditions. *F*, however, was mostly recommended when sample size was small and k = 8 except for when the sphericity assumption was seriously violated (ε =.48). When ε =.75, *F* seemed to be a favored method. ADF was recommended when the sample size got large (60, 100, 200) with k =4.

Table 23

Recommended Methods for Moderatly Non-normal Distribution

Degree of Non-sphericity	Number of Levels	Sample Size	Most Powerful/Recommended Methods
1	*K*=4	15, 30	YBRES, ML/SB1
		60, 100, 200	ADF, YBRES, ML/SB1/RES
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	YBRES, ML/SB1
0.96	*K*=4	15, 30	YBRES, ML/SB1, *F,* HF
		60, 100, 200	ADF, YBRES, ML/SB1/RES
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	YBRES, ML/SB1
0.75	*K*=4	15, 30	YBRES, *F*, HF, ML/SB1
		60, 100, 200	*F,* HF, BOX, ADF, YBRES
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	*F*, YBRES, HF, BOX
0.48	*K*=4	15, 30	YBRES
		60, 100, 200	YBRES, ML/SB1/RES
	K=8	15, 30	HF, BOX
		60, 100, 200	YBRES

Table 23 shows that the majority of the methods recommended were RMM methods. Except for $k = 8$ with small sample sizes (15, 30), YBRES was recommended for all conditions. *F*, however, was mostly recommended when sample size was small and $k = 8$ except for when the sphericity assumption was seriously violated (ε =.48). When ε =.75, *F* and HF seemed to be the favored methods.

Table 24

Recommended Methods for Severely Non-normal Distribution

Degree of Non-sphericity	Number of Levels	Sample Size	Most Powerful/Recommended Methods
1	*K*=4	15, 30	YBRES, ML/SB1
		60, 100, 200	ADF, YBRES
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	YBRES, ML/SB1
0.96	*K*=4	15, 30	YBRES, ML/SB1, FADF
		60, 100, 200	ADF, YBRES, ML/SB1/RES
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	YBRES, ML/SB1
0.75	*K*=4	15, 30	YBRES, FM
		60, 100, 200	*F*, YBRES, HF, BOX
	K=8	15, 30	*F,* FM, HF
		60, 100, 200	YBRES, *F*. HF
0.48	*K*=4	15, 30	YBRES, YBADF, HF, BOX
		60, 100, 200	YBRES, FADF, FM, FRES, YBADF
	K=8	15, 30	FM, HF, BOX
		60, 100, 200	YBRES, FM

Table 24 shows that the majority of the methods recommended were RMM methods. Except for k = 8 with small sample sizes (15, 30), YBRES was recommended for all conditions. *F*, however, was mostly recommended when sample size was small and k = 8 except for when the sphericity assumption was seriously violated (ε =.48). When ε =.75, *F* and HF seemed to be the favored methods.

Then the four tables were combined to come up with a summary table for recommended methods for all distribution conditions as is shown in Table 25.

Table 25

Recommended Methods for Conditions across Distributions

Degree of Non-sphericity	Number of Levels	Sample Size	Most Powerful/Recommended Methods
1	*K*=4	15, 30	YBRES, ML/SB1
		60, 100, 200	ADF, YBRES
	K=8	15, 30	*F*, HF, FM
		60, 100, 200	YBRES, ML/SB1
0.96	*K*=4	15, 30	YBRES, ML/SB1
		60, 100, 200	ADF, YBRES
	K=8	15, 30	*F*, HF, FM
		60, 100, 200	YBRES, ML/SB1
0.75	*K*=4	15, 30	YBRES
		60, 100, 200	*F*, YBRES
	K=8	15, 30	*F*, HF, FM
		60, 100, 200	*F*, YBRES
0.48	*K*=4	15, 30	YBRES
		60, 100, 200	YBRES
	K=8	15, 30	HF
		60, 100, 200	YBRES

Table 25 shows that the majority of the methods recommended were RMM methods. Except for *k* =8 with small sample sizes (15, 30), YBRES was recommended for all conditions. *F*, however, was mostly recommended when sample size was small and *k* = 8 except for when the sphericity assumption was seriously violated (ε =.48). Of course, it is clear that no one method performs consistently well under all situations and outperforms all the other methods. But the current study carried out on combinations of 5 sample size conditions, 4 distributions, 4 levels of sphericity conditions, and 2 levels of repeated measures provided strong support for applying RMM methods to one-way repeated measure design. Based on the previous analysis, RMM methods were recommended under almost all conditions except when the model was complex (*k*=8) and sample size was small (15, 30), and among all the RMM methods, YBRES outperformed all the other methods. When there were 8 levels of repeated measures and the sample size was small (15, 30), ANOVA-based methods were recommended, among which HF became a safe choice as it performed most stably among all the ANOVA-based methods.

It is worth noting that when the sphericity assumption was seriously violated (ε=.48), the traditional *F* test tended to over-reject the correct model and thus got inflated Type I error rates, despite the distribution form. Thus, *F* test was not recommended for ε =.48 even when the data were normal. This leads to the suggestion that the tests for both normality and sphericity should precede the analysis of data collected for the one-way repeated measure design. If both normality and sphericity assumptions are met, of course *F* can be used for the sake of simplicity. If the sphericity assumption is violated, Tables 21-25 can be employed to facilitate the choice of the methods to be used. However, a potential issue with the guideline provided in Tables 21-25 could be that ε is the population sphericity parameter and is usually unknown. Even though it can be estimated from a sample, practitioners need to take some cautions especially when sample size is small.

The findings obtained in this study echoed those in Fan and Hancock (2012) which compared somewhat different sets of ANOVA-based methods (*F* test, Welch's test, the Brown-Forsythe test, James' second-order test, and the Alexander-Govern test) and RMM methods (ADF, SB1,YBADF, FADF, and Bartlett's correction to the ML) applied for between-subjects designs. The findings for the between-subjects designs indicated that RMM methods not only provided robust Type I error rates across different conditions specified, but also were more powerful than AVOVA-based methods. Among all the RMM methods, YBADF and FADF outperformed the other RMM methods. This study, however, added RES, YBRES, and FRES to comparison and found that YBRES became the most recommended method but FADF also performed comparably well among all the RMM methods. Fan and Hancock called for the study to be extended to repeated mesures, which was the focus of this study.

Though this study has gained ground for RMM methods to be applied in one-way repeated measure design, the findings from this study warrant many future investigations. First, this study used only two levels of repeated measure, 4 and 8. There seemed to be some drastic differences in performance for RMM methods between these two levels. Therefore, some additional levels (e.g., 3, 6) may be added in future studies to investigate the ideal conditions of applying RMM models to analysis. In addition, more distribution forms and more sample sizes can also be included in the future studies to gain a more comprehensive picture of RMM’s performance. Second, because this study only examined the one-way repeated measure design, in order to generalize the conclusion to a broader scenario, a between-subject factor can also be included in the study so that the performance of RMM methods can be investigated in the balanced and unbalanced designs containing one between-subjects and one within-subject factor. Third, this study can also be extended to perform with more complex models like multivariate repeated measures designs.

Appendix

<u>*Sample EQS Codes*</u>

```
/SPECIFICATIONS

VAR=3; cases=n; ME=AGLS; MA=RAW; DATA='Path1’;

ANAL=MOMENT;

/EQUATIONS

V1 = *V999 + 1.000 E1;

V2 = *V999 + 1.000 E2;

V3 = *V999 + 1.000 E3;

/VAR

E1 to E3= *;

/CONSTRAINT

(V1,V999)=(V2,V999)=(V3,V999)= )=(V4,V999);

/PRINT

FIT=ALL; COV=YES;

/END
```

References

Algina, J., & Keselman, H. J. (1997). Testing repeated measures hypotheses when the covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James Test. *Multivariate Behavioral Research*, *32*, 255-274.

Anderson, J. C., & Gerbing, D. W. (1984). The effects of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*, 155-173.

Bandalos, D. L. (2006). The use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 385-426). Charlotte, NC: Information Age Publishing, Inc.

Bentler, P. M. (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology*, *47*, 563-592.

Bentler, P. M. (2006). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*, 181-197.

Berkovits, I., Hancock, G. R., & Nevitt, J. (2000). Bootstrap resampling approaches for repeated measure designs: Relative robustness to sphericity and normality violations. *Educational and Psychological Measurement*, *60*, 877-892.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. *Annals of Mathematical Statistics*, *25*, 290-302.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *34*, 144-152.

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

Browne, M. W., & Arminger, G. (1995). Specification and estimation of mean-and covariance-structure models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, 185-249. Boston, MA: Springer.

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

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

Fan, W., & Hancock, G. R. (2012). Robust means modeling: An alternative to hypothesis testing of independent means under variance heterogeneity and non-normality. *Journal of Educational and Behavioral Statistics*, *37*, 137-156.

Fernándes, P., Vallejo, G., Livacic-Rojas, P., Herrero, J., & Cuesta M. (2009). Comparison of the power of four statics in repeated measures design in the absence of sphericity with and without serial autocorrelation. *Review of Psychology*, *16*, 65-76.

Finney, S. J., & DiStefano, C. (2013). Nonnormal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd eds.) (pp. 437-490). Charlotte, NC: Information Age Publishing, Inc.

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

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate non-normality. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*, 356-410.

Geisser, S., & Greenhouse, S.W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics*, *29*, 885-891.

Hancock, G. R. (1997). Structural equation modeling methods of hypothesis testing of latent variable means. *Measurement and Evaluation in Counseling and Development*, *20*, 91-105.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*, 373-388.

Harwell, M. R., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101-125.

Harwell, M. R., & Serlin, R. C. (1977). An empirical study of five multivariate tests for the single factor repeated measures model. *Communications in Statistics-Simulation and Computation*, *26*, 605-618.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329-367.

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

Huynh, J. W., & Feldt, L. S. (1970).Conditions under which mean square ratios in repeated measurements designs have exact *F*-distribution. *Journal of American Statistical Association*, *65*, 1582-1589.

Huynh, J. W., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics*, *1*, 69-82.

Jensen, D. R. (1982). Efficiency and robustness in the use of repeated measurements. *Biometrics*, *38*, 813-825.

Keiffer, K. M. (2002). On analyzing repeated measures designs with both univariate and multivariate methods: A primer with examples. *Multiple Linear Regression Viewpoints*, *28*, 1-17.

Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology*, *54*, 1-20.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, *52*, 63-78.

Keselman, H. J., Kowalchuk, R. K., Algina, J. Lix, L. M. & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical and Statistical Psychology*, *53*, 175-191.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous. *Journal of Educational Statistic*s, *18*, 305-319.

Keselman, H. J., & Keselman, J. C. (1988). Comparing repeated measures means in factorial designs. *Psychophysiology*, *25*, 612-618.

Keselman, J. C., & Keselman, H. J. (1990). Analyzing unbalanced repeated measurements: A quantitative research synthesis. *British Journal of Mathematical and Statistical Psychology, 43*, 265-282.

Keselman, H. J., Keselman, J., C., & Shaffer, J. P. (1991). Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity. *Psychological Bulletin*, *110*, 162-170.

Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L., & Wilcox, R. R. (2000). Testing treatment effects in repeated measures design: Trimmed means and bootstrapping. *British Journal of Mathematical and Statistical Psychology*, *53*, 175-191.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

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

Lix, L. M., Algina, J., & Keselman, H. J. (2003). Analyzing multivariate repeated measures designs: A comparison of two approximate degrees of freedom procedures. *Multivariate Behavioral Research*, *38*, 403-431.

Lix, L. M., & Keselman, H. J. (2010). Analysis of variance: Repeated measures designs. In Hancock & Muller, *The reviewer's guide to quantitative methods in* *the social sciences* (pp. 15-27). New York: Routledge.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance *F* Test. *Review of Educational Research*, *66*, 579-619.

McCall, R. B., & Appelbaum, M. I. (1973). Bias in the analysis of repeated-measures designs: some alternative approaches. *Child Development*, *44*, 40-415.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data*. Belmont, CA: Wadsworth.

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

Muthén, B. (1989). Multiple-group structural modeling with non-normal continuous variables. *British Journal of Mathematical and Statistical Psychology*, *42*, 55-62.

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

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 353-377.

Nevitt, J., & Hancock, G. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*, 439-478.

Powell D. A., & Schafer, W. D. (2001). The robustness of the likelihood ratio chi-square test for structural equation models: A Meta-analysis. *Journal of Educational and Behavioral Statistics*, *26*, 105-132.

Quintana, S. M., & Maxwell, S. E. (1994). A comparison of six estimates of epsilon in repeated measures designs with small samples sizes. *Journal of Educational Statistics*, *19*, 57-71.

SAS Institute, Inc. (1990). *SAS/IML software: Usage and reference, Version 6.* Cary, NC: Author.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality & Quantity*, *24*, 367-386.

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociological Methodology*, *22*, 249-278.

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

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent Variables Analysis: Applications for Developmental Research* (pp. 399-419). Thousand Oaks, CA: Sage.

Thompson, M. S., & Green, S. B. (2006). Evaluating between group differences in latent variable means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 119-169). Charlotte, NC: Information Age Publishing, Inc.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*, 465-471.

Wilcox, R. R. (1993). Analyzing repeated measures or randomized block designs using trimmed means. *British Journal of Mathematical and Statistical Psychology*, *46*, 63-76.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research*, *65*, 51–77.

Wilcox, R. R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*. San Diego, CA: Academic Press.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*, 300-314.

Yuan, K. H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical*

Association, *92*, 767-774.

Yuan, K. H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, *51*, 289-309.

Yuan, K. H., & Bentler, P. M. (1999). *F* tests for mean and covariance structure analysis. *Journal of Education and Behavioral Statistics*, *3*, 225-243.