COMPARABILITY OF COVARIANCE STRUCTURES AND ACCURACY OF INFORMATION CRITERIA IN MIXED MODEL METHODS FOR LONGITUDINAL DATA ANALYSIS

by

DANIEL J. YANOSKY II

(Under the Direction of Stephen Olejnik)

ABSTRACT

Modern mixed model methods for analyzing longitudinal data require researchers to select a covariance structure for the data to fully specify the model and obtain statistical tests of the fixed effects. The current study is a Monte Carlo simulation with primary purposes to 1) identify surrogate covariance structures for seven known models and estimate the severity of committing an error in covariance specification in terms of empirical Type I error rates and statistical power, 2) estimate accuracy rates of five information criteria in selecting appropriate covariance structures, and 3) estimate the empirical Type I error rates and power for models chosen by each information criterion.

Data were generated corresponding to a single group repeated measures design with N = 10, 30, or 60 subjects and a quantitative response variable measured over t = 3 or 6 occasions. Other salient variables included the magnitude of serial correlation, presence of non-constant variance, and so forth. Data were generated under 72 conditions with 10,000 replications per condition. Statistical Analysis System (SAS) version 9.1 and R version 2.4.0 were used to generate and analyze the data.

A preliminary investigation demonstrated that the Kenward-Roger degrees of freedom approximation yields *F*-tests for the mixed models with superior Type I error control compared to the Between/Within method, Satterthwaite approximation, and the sandwich estimator.

Results corresponding to the primary research questions demonstrated 1) seven covariance structures were found to be acceptable approximations of a given true model in 14 instances, 2) rates of selecting appropriate covariance structures for information criteria were found to be substantially influenced by accounting for surrogate structures with a rate of 69% for both AIC and BIC, 3) empirical Type I error rates were found to be slightly conservative and therefore well controlled and power estimates comparable when models were selected by AIC, AICC, HQIC, BIC, and CAIC.

Secondary investigations compared the performance of mixed models with classical methods and evaluated the empirical Type I error control of the Group x Time interaction test.

The implications of these findings are discussed and heuristics for applied researchers working with this variety of data are suggested.

INDEX WORDS: Longitudinal data analysis, Mixed effects models, Covariance modeling, Information criteria, Type I error rates, Statistical Power

COMPARABILITY OF COVARIANCE STRUCTURES AND ACCURACY OF INFORMATION CRITERIA IN MIXED MODEL METHODS FOR LONGITUDINAL DATA ANALYSIS

by

DANIEL J. YANOSKY II

B.A., Emory University, 1996

M.S., The University of Georgia, 2005

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

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2007

COMPARABILITY OF COVARIANCE STRUCTURES AND ACCURACY OF INFORMATION CRITERIA IN MIXED MODEL METHODS FOR LONGITUDINAL DATA ANALYSIS

by

DANIEL J. YANOSKY II

Major Professor: Stephen Olejnik

Committee: Deborah Bandalos
Carl Huberty
Dan Hall

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2007

DEDICATION

This dissertation is dedicated to my family: the memory of my father, Daniel J. Yanosky; my mother, Karen L. Yanosky; and my brother, Jeffrey D. Yanosky. Their unconditional love and support has had a profound impact in my life and made it possible for me to achieve this goal.

ACKNOWLEDGEMENTS

I would like to sincerely thank Dr. Stephen Olejnik for his guidance, mentorship, and support during this project and my doctoral program of study in general. Furthermore, I thank my committee members: Dr. Deborah Bandalos, Dr. Carl Huberty, and Dr. Dan Hall for their exemplary instruction, advisement, and their guidance through my dissertation.

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION 1

Background 1

The Problem 5

Purposes of the Study 8

Preliminary Research Question 9

Primary Research Questions 10

Secondary Research Questions 12

Significance of the Study 15

Summary 17

CHAPTER II: LITERATURE REVIEW 19

Information Criteria 20

The Accuracy of Information Criteria 27

Empirical Type I Error Rates, Statistical Power Estimates, and the Accuracy of Information Criteria 36

Comparing Mixed Model and Classical Methods 44

Summary 46

CHAPTER III: METHODS 51

The Classical Linear Model 52

The Mixed Model 55

Explication of Experimental Factors 64

Additional Factor for Empirical Power Estimation 72

Additional Factor for the Comparison of Between-Subjects Groups 73

Data Generation and Model Fitting Procedures Overview: Phase I 74

Data Generation and Model Fitting Procedures Technical Notes: Phase I 76

Data Generation Overview: Phase II 82

Data Generation Technical Notes: Phase II 83

Generation of Correlation and Covariance Matrices 85

Analysis & Results 87

The Evaluation of Empirical Type I Error Rates 87

The Evaluation of Empirical Power Estimates 88

Limitations 89

Summary 91

CHAPTER IV: RESULTS 102

Nonconvergence Rates 104

Preliminary Research Question i 106

Primary Research Question 1 109

Primary Research Question 2 114

Primary Research Question 3 117

Secondary Research Question 4 120

Secondary Research Question 5 129

Summary 131

CHAPTER V: DISCUSSION 155

Synopsis 155

Discussion 159

Preliminary Research Question 159

Primary Research Questions 161

Secondary Research Questions 170

Limitations and Suggestions for Future Research 173

Recommendations for Applied Researchers 175

REFERENCES 177

APPENDIX 185

LIST OF TABLES

Table 3.1: *Number of Experimental Conditions* 94

Table 3.2: *Structures of Parameterized Covariance Matrices* 95

Table 3.3: *Variance Multipliers for Data Transformation* 98

Table 3.4: *Mean Configuration Values for Power Analysis* 99

Table 3.5: *Population Correlation Matrices for the Toeplitz Pattern* 100

Table 3.6: *Population Correlation Matrices for the Unstructured Pattern* 101

Table 4.i.1: *Empirical Type I Error Rates by Test Statistic Option and Marginal Conditions* 133

Table 4.i.2: *Empirical Type I Error Rates (b) by Test Statistic Option and Each Individual Condition* 134

Table 4.1.1: *Empirical Type I Error Rates and Power Estimates for All True by Potential Surrogate Combinations* 135

Table 4.1.2: *Surrogate Covariance Structures* 136

Table 4.1.3: *Mean Multivariate Distance (u) between the Model-Estimated and Sample-Based Covariance Matrices* 137

Table 4.1.4: *Empirical Type I Error Rates for Mixed Models Aggregated Across the Seven True Models* 138

Table 4.1.5: *Empirical Power Rates for Mixed Models Aggregated Across the Seven True Models* 139

Table 4.2.1: *Information Criteria Selection Rates for Correct and Surrogate Models* .. 140

Table 4.2.2: *Information Criteria Selection Rates for Incorrect Models* 141

Table 4.3.1: *Empirical Type I Error Rates by Information Criteria* 142

Table 4.3.2: *Empirical Power Estimates by Information Criteria* 143

Table 4.3.3: *Empirical Type I Error Rates for Fitted Models Selected by AIC by True Model* 144

Table 4.3.4: *Empirical Type I Error Rates for Fitted Models Selected by BIC by True Model* 145

Table 4.4.1: *ε Values for Population Covariance Matrices* 146

Table 4.4.2: *Empirical Type I Error Rates and Statistical Power: Correct Model Fit Only* 147

Table 4.4.3: *Empirical Type I Error Rates for Comparing Mixed Models with Classical Methods under Marginal Conditions* 148

Table 4.4.4: *Empirical Power Estimates for Comparing Mixed Models with Classical Methods under Marginal Conditions* 149

Table 4.4.5: *Empirical Power Estimates for Comparing Mixed Models with Classical Methods under Marginal Conditions* 150

Table 4.4.6: *Empirical Type I Error Rates for Extremely Non-Spherical Data* 151

Table 4.4.7: *Empirical Type I Error Rates for Extremely Non-Spherical Data: Expanded View of UN* 152

Table 4.4.8: *Empirical Power Estimates for Models Fit to Extremely Non-Spherical Data* 153

Table 4.5.1: *Empirical Type I Error Rates for the Test of Interaction (Aggregated across all seven true models)* 154

Table A1: *Example of Permanent SAS Dataset* 186

Table A2: *Population Covariance Matrices for both Independence & Variance Components Structures* 190

Table A3: *Population Covariance Matrices for the Compound Symmetry Structure* 191

Table A4: *Population Covariance Structure for the Heterogeneous Compound Symmetry Structure* 192

Table A5: *Population Covariance Matrices for the Heterogeneous Autoregressive Structure* 193

Table A6: *Population Covariance Matrices for the Heterogeneous Toeplitz Structure* 194

Table A7: *Population Covariance Structures for the Unstructured Pattern* 195

Table A8: *Frequencies of Nonconvergence* 196

Table A9: *Empirical Type I Error Rates (a) by Test statistic Option and Each Individual Condition* 197

Table A10: *Empirical Type I Error Rates for the Independence True Model* 198

Table A11: *Empirical Type I Error Rates for the Variance Components True Model* .. 199

Table A12: *Empirical Type I Error Rates for the Compound Symmetry True Model* 200

Table A13: *Empirical Type I Error Rates for the Heterogeneous Compound Symmetry True Model* 201

Table A14: *Empirical Type I Error Rates for the Heterogeneous Autoregressive True Model* 202
Table A15: *Empirical Type I Error Rates for the Heterogeneous Toeplitz True Model* 203
Table A16: *Empirical Type I Error Rates for the Unstructured True Model* 204
Table A17: *Empirical Power Estimates for the Independence True Model* 205
Table A18: *Empirical Power Estimates for the Variance Components True Model* 206
Table A19: *Empirical Power Estimates for the Compound Symmetry True Model* 207
Table A20: *Empirical Power Estimates for the Heterogeneous Compound Symmetry True Model* 208
Table A21: *Empirical Power Estimates for the Heterogeneous Autoregressive True Model* 209
Table A22: *Empirical Power Estimates for the Heterogeneous Toeplitz True Model* ... 210
Table A23: *Empirical Power Estimates for the Unstructured True Model* 211
Table A24: *Population Covariance Matrices for both ARH & UN Structures for Follow-up Analysis* 212
Table A25: *Empirical Type I Error Rates for the Independence True Model: Test of the Interaction* 213
Table A26: *Empirical Type I Error Rates for the Variance Components True Model: Test of the Interaction* 214
Table A27: *Empirical Type I Error Rates for the Compound Symmetry True Model: Test of the Interaction* 215
Table A28: *Empirical Type I Error Rates for the Heterogeneous Compound Symmetry True Model: Test of the Interaction* 216

Table A29: *Empirical Type I Error Rates for the Heterogeneous Autoregressive True Model: Test of the Interaction* 217

Table A30: *Empirical Type I Error Rates for the Heterogeneous Toeplitz True Model: Test of the Interaction* 218

Table A31: *Empirical Type I Error Rates for the Unstructured True Model: Test of the Interaction* 219

CHAPTER I: INTRODUCTION

Background

Historically, applied researchers have collected and analyzed longitudinal data for two main purposes. *First*, researchers are often interested in individual changes over time. *Second*, researchers wish to increase the statistical power of their analyses. One way of achieving the second goal is by having each subject serve as his/her own control. For these purposes, multiple observations are collected from each subject through time[1]. Classical approaches to analyzing data of this sort originated out of the analysis of split-plot designs, becoming known as the Repeated Measures Analysis of Variance (RM ANOVA) as well as a multivariate approach based on the Multivariate Analysis of Variance (MANOVA).

While the univariate RM ANOVA approach retains the statistical assumptions of the general linear model (including normality, constant population variances, etc.), it also requires one additional assumption, known as sphericity. The sphericity assumption requires that the variances of differences of all possible pairs of measurements be constant. That is, in the context of longitudinal data, the sphericity assumption requires the relationships of measurements spaced further apart in time to equal those spaced

[1] By definition, longitudinal data are collected through time; however, more general sampling designs may collect multiple observations from each subject through a different dimension, such as space.

closer in time. Unfortunately, this assumption is rarely if ever met in applied longitudinal data analysis (Everit, 2001, p.209; Rogan, Keselman, & Mendoza, 1979). Furthermore, it has been shown that if the assumption of sphericity is violated, RM ANOVA test of the null hypothesis of no Time effect is positively biased, or, that is, liberal – the test tends to reject the null hypothesis at a greater rate than that specified by the nominal significance level (α_n) (Everitt, pp. 139-141; Rogan et al.). This is a serious dilemma that calls into question the validity of the statistical inferences drawn from a RM ANOVA via classical methods.

In attempts to ameliorate this problem, statisticians such as Box, Greenhouse and Geisser, Huynh and Feldt, etc. developed correction procedures that entail adjusting the degrees of freedom of the affected *F*-tests by a quantity that measures the extent of the sphericity violation, known as ε (epsilon) (Keppel, 1991, pp. 351-353). While it has been demonstrated that such corrections are able to hold the empirical Type I error rates (α_e's) close to nominal levels (Rogan et al., 1979), reviews of the applied literature in psychology, education, and other social sciences show that researchers often fail to utilize these procedures (Keselman, Huberty, et al., 1998; Kowalchuk, Lix, & Keselman, 1996). Furthermore, the entire approach is somewhat flawed. Basically, the method is to fit a model to the data known to be unrealistic, or even incorrect, and then adjust for the extent of "incorrectness" post hoc in hopes to neutralize the effects of misspecifying the model in the first place.

The other classical approach to longitudinal data analysis is the multivariate or MANOVA approach. Here the t measurement occasions are treated as separate outcome variables and expressed as a vector of t-1 contrasts (where t = the number of measurement occasions). It is this vector of mean contrasts that the MANOVA approach analyzes. The advantage of this approach is that it does not require the restrictive sphericity assumption. On the contrary, unlike the RM ANOVA, the multivariate approach makes no assumptions concerning the relationships of measurements across time. Thus, the multivariate approach may over-generalize the situation. It assumes no structure to the covariances of the measurement occasions, and, as a result, estimates each of the t variances and each of the $t(t-1)/2$ covariances separately. While this degree of flexibility may be appropriate in some cases, it is often not needed and may have a serious disadvantage. The number of variance components to be estimated in the MANOVA approach increases markedly with the number of measurement occasions. The estimation of a large number of variance components requires substantial degrees of freedom and consequently reduces the statistical power of the analysis. If a large sample size has been obtained for the analysis, this may not be problematic. However, repeated measures designs are often used because of their economical use of small or moderate sample sizes and, if this is the case, then the MANOVA analysis may subvert the main advantage of the design.

An alternative approach to modeling longitudinal data within a univariate conceptualization using modern mixed model (MM) methods is somewhat different than RM ANOVA. Here, the necessity to meet the sphericity assumption is obviated by modeling the covariance structure of the data with respect to time instead of assuming the form of that structure. That is, the idea is to fit a more correct model to the data initially, by using the data or some a priori knowledge of the data collection mechanism to estimate an appropriate covariance structure. Inferences are then made directly from this model, thereby precluding the need to make any corrections for initial implausible assumptions.

Moreover, the modern MM approach uses a number of parameterized structures for common covariance patterns found in longitudinal data. The use of these parameterized structures reduces the number of estimates needed to approximate the overall covariance matrix (sometimes to a substantial degree) and thereby circumvents the loss of statistical power often encountered when using the MANOVA approach (Wolfinger, 1996).

Additionally, if the univariate assumption of sphericity has been met and the covariance of the data is compound symmetric[2] (CS), this covariance structure can be specified in the modern MM approach for longitudinal data. That is, the MM for longitudinal data is a generalization of the split-plot model, which, in turn is an extension of the classical linear model (CLM). Therefore, the typical CLM assumptions of independence of observations and homogeneity of population variances are relaxed in the MM for longitudinal data and the variables may be both fixed and random (Fitzmaurice, Laird, & Ware, 2004, pp. 187-197; Rencher, 2000, pp. 426-429; Vallejo & Livacic-Rojas, 2005). As a result, the modern MM may be specified so that the MM *F*-test reduces to the typical CLM *F*-test[3] (Vallejo & Livacic-Rojas). Furthermore, when the modern MM is specified with the CS covariance structure, the MM *F*-test[4] reduces to the RM ANOVA conventional *F*-test. (Schaalje, McBride, & Fellingham, 2002; Wright & Wolfinger,

---

[2] Technically, compound symmetry is a subset of sphericity; however, the two conditions are often used interchangeably.
[3] MM conditions under which this is true include the specification of an independence model for the covariance of the response, unadjusted degrees of freedom for the *F*-test, and the exclusion of any random effects other than the usual error term.
[4] This is true when using unadjusted degrees of freedom for the *F*-test [i.e., the exact degrees of freedom method (Between/Within) implemented in SAS].

1997, p. 150). Likewise, if the variance components are suspected to be substantially different and not conforming to any known pattern, the modern MM can be specified with an unstructured covariance model with respect to time and therefore estimate each variance component separately, like the MANOVA approach. In this situation, the MM *F*-test is related to the Lawley-Hotelling multivariate statistic (Wright & Wolfinger). As a result, the modern MM approach not only provides modeling alternatives between the two extremes of the univariate and multivariate ANOVA approaches, it also subsumes both (Wolfinger, 1993). In addition, the modern MM approach can be used in situations where the data are not balanced over measurement occasions and where missing values and time-varying covariates are present.

The Problem

While the modern MM approach may constitute a more modern and flexible alternative to modeling longitudinal data than the classical methods, it is not a panacea. Instead of concerning oneself with correcting for non-sphericity or with the ramifications of insufficient statistical power, one is now faced with the challenge of selecting an appropriate covariance structure for the modern MM. While modeling the covariance structure is not typically of central interest in its own right, it is considered to be an important feature in obtaining valid tests and inferences for the fixed effects in the model (Littell, Milliken, Stroup, & Wolfinger, 1996, p. 171).

Methodologists have suggested a few methods for selecting an appropriate covariance structure for a given dataset (Davis, 2002, pp. 130-156; Diggle, 1988; Ferron, Dailey, & Yi, 2002; Tonidandel, Overall, & Smith, 2004; Verbeke & Molenbergs, 2000,

p. 74-76). The overwhelmingly preferred option, however, entails choosing an appropriate covariance structure from a set of plausible alternatives based on information criteria (Davis; Lindsey, 1999, p. 44; Pinheiro & Bates, 2000, pp. 253-256; Verbeke & Molenbergs, pp. 74-76).

Information criteria are a group of quantitative indicators that combine a measure of the overall model fit (some function of the maximized likelihood function) and a penalization based on model complexity. This penalization is usually some function of the number of parameters to be estimated in the model and other characteristics of the data (e.g., sample size). In this way, information criteria summarize both fit and parsimony for any given fitted model. These criteria are especially useful in choosing appropriate covariance structures in the context of modern MMs because candidate structures are often not nested and therefore Likelihood Ratio Tests (LRTs) are not applicable. However, information criteria for candidate or competing non-nested models are comparable.

The most common information criteria are the Akaike Information Criterion (AIC) and the Baseyian Information Criterion (BIC; or Schwarz Baseyian Criterion, SBC); however, many other formulations have been proposed (Akaike, 1974; Bozdogan, 1987; Burnham & Anderson, 2002; Hannan & Quin, 1979; Schwarz, 1978). AIC and BIC differ in the penalization with AIC only penalizing for increasing complexity of the model (increasing explanatory variables) and BIC incorporating information concerning the size of the sample in the amount to be penalized.

There are two main reasons why information criteria are not a perfect solution to the problem of covariance structure modeling. *First*, it is possible to obtain different preferred covariance models for the same data depending on which information criterion is used (that is, depending on the type of penalization that is applied to the likelihood function, which is a subjective choice left up to the researcher). This is especially disconcerting because there is no consensus in the research literature as to which criterion is most appropriate in covariance modeling situations (Ferron et al., 2002; Gomez, Schaalje, & Fellingham, 2005; Keselman, Algina, Kowalchuk, & Wolfinger, 1998; Vallejo & Livacic-Rojas, 2005).

Second, Monte Carlo simulations studying the performance of information criteria report widely varying accuracy of these criteria in selecting the correct covariance structure (Ferron et al., 2002; Gomez et al., 2005; Keselman, Algina, et al., 1998; Vallejo & Livacic-Rojas, 2005). Keselman, Algina, et al. (1998) evaluated the accuracy of AIC and BIC selecting the correct covariance structure among 15 candidate structures for data that had been generated from six population or correct structures. The authors reported relatively low estimates of accuracy for both criteria: AIC selected the correct covariance structure only 47% of the time and BIC only 35% of the time. Ferron et al. (2002) reported accuracy rates of 79% and 66% for AIC and BIC, respectively, when only two candidate models were considered. Like Keselman, Algina, et al., Gomez et al. (2005) investigated the accuracy of AIC and BIC under many different simulated conditions (including varying samples sizes, varying degrees of design imbalance, etc.); however, the authors did not report marginal accuracy rates across those conditions. They did note that accuracy rates ranged anywhere from 3% to 79% for particular conditions and rates were most influenced by sample size and covariance structure (15 candidate models were evaluated). Moreover, these authors found that AIC outperformed BIC for modeling more complex structures whereas BIC outperformed AIC for simpler structures. Finally, as a side note to their comparison of α_e's between the Brown-Forsythe test and the MM approach, Vallejo and Livacic-Rojas (2005) noted that AIC accuracy rates ranged from 23% to 87% depending on the complexity of the structure being modeled.

This has been a brief presentation of the issues involved in the use of information criteria, covariance structure modeling, and the use of modern MM methods in the analysis of longitudinal data. The current study was designed to address many of these issues. The specific purposes of the current study are outlined in detail next.

Purposes of the Study

The current study was multifaceted with preliminary, primary, and secondary purposes that relate to the analysis of longitudinal data using modern MMs and classical methods. Briefly, the preliminary purpose was to evaluate four test statistic options within the modern MM framework with respect to α_e's. Next, the three primary purposes were to 1) identify comparable or surrogate covariance structures, 2) estimate information criteria selection rates of appropriate covariance structures, and 3) estimate the α_e's of modern MM test statistics whose covariance models were selected by a given information criterion. Finally, two secondary purposes were to 1) compare between modern MM methods and classical methods in analyzing longitudinal data with respect to α_e's and statistical power estimates, and 2) investigate the α_e's of the interaction test when both within-subjects and between-subjects factors are present in the design. These purposes and the problems they address are delineated in greater detail below.

Preliminary Research Question

The CLM *F*-statistics follow the *F*-distribution exactly and their degrees of freedom are well defined and can be obtained via straightforward calculations. In contrast, MM Wald-type *F*-statistics are available for inference; however, these statistics are only *approximately F*-distributed. Furthermore, in longitudinal data analysis, multiple observations from the same experimental unit are often correlated. This association among the observations complicates the estimation of the number of independent pieces of information available for statistical inference. As a result, the calculation of the degrees of freedom for test statistics becomes more complicated. Further complications arise when the data exhibit multiple random effects and/or are not balanced (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006, p. 152).

A number of methods for obtaining valid test statistics under these conditions have been developed. Some of these methods estimate the degrees of freedom for test statistics under these conditions. Other methods rely on the estimation of empirical or robust variances for use in the computation of statistical tests.

A comparison of several of these methods is a preliminary interest in the current study. More specifically, known problems exist with the default manner (the Between/Within method[5]) in which SAS's PROC MIXED computes degrees of freedom for test statistics for MMs (Littell et al., 2006, p. 188)[6]. Statisticians have suggested the use of either the Satterthwaite or the Kenward/Roger (KR) approximations in order to address this problem (Fitzmaurice et al., 2004, pp. 98-99; Gomez et al., 2005; Keselman

[5] In SAS version 9.1, the Between/Within method is the default for calculating degrees of freedom for models specified with the REPEATED statement, as was the case in the current study. See Chapter 3 for more details.
[6] See page 47 for further explanation.

et al., 1999; Littell et al., p. 188). Furthermore, Fitzmaurice et al. (p.177) have suggested the use of the "sandwich" estimator in order to obtain standard errors for parameter estimates and test statistics for the fixed effects. The use of the sandwich estimator addresses problems of covariance model misspecification in general. Therefore, the preliminary research question is:

> *i) How do test statistics for the fixed effects of the mixed model compare with respect to* α_e's *when the SAS PROC MIXED default (the Between/Within method), the Satterthwaite or KR approximations, or the sandwich estimator options are used?*

Primary Research Questions

As mentioned above, accuracy rates for information criteria estimated in previous investigations were found to be substantially low and highly variable. It is suspected that these unfortunate qualities can be attributed to two main features of these studies. *First*, studies like Ferron et al. (2002) that considered only a modest number of candidate models were found to report higher accuracy rates than those studies that considered many candidate models, such as Keselman, Algina, et al. (1998). Therefore, the number of candidate models that one is willing to entertain influences the accuracy of information criteria to a substantial degree. *Second*, some statisticians have mentioned that accuracy rates may be biased due to the fact that under certain conditions some "incorrect" covariance structures may in fact serve as surrogates or acceptable approximations of the correct structures (Gomez et al., 2005; Keselman, Algina, et al). While a specific definition of a surrogate was not provided by these authors, it is obvious from their usage of the term that a surrogate model is one that is comparable to the correct model in obtaining relatively good fit to the data, and consequently providing statistical tests that closely approximate those of the correct model. Thus, in the context of information criteria accuracy, even though a given information criterion did not select the correct covariance structure, there is the possibility that it did select a comparable covariance structure that in fact obtains results that closely approximate those of the correct model. By not accounting for the possibility of surrogate covariance structures, previous estimates of the accuracy rates of these information criteria may drastically underestimate their usefulness in covariance modeling. These results may discourage the use of the modern MM approach by applied researchers.

In an attempt to investigate and neutralize the effect of this bias in the performance of information criteria, the current study generated data via a Monte Carlo simulation to identify covariance structures that may serve as surrogates for a correct structure. Moreover, the accuracy rates of information criteria are reported more precisely with these structures taken into account. That is, rates of selecting *appropriate* covariance models are reported, where appropriate covariance models are defined as a set of models including the correct model and any surrogate models that have been identified for the correct model. The primary research questions in the current study are:

1) Do surrogate covariance structures exist? If so, which structures serve as acceptable approximations for a given population or correct structure and under what conditions?

2) What are the selection rates of a particular information criterion with respect to selecting a) the correct model, b) a surrogate model, and c) an appropriate model? What are the selection rates with respect to a) underfitting or b) overfitting the data?

3) Will the analysis be statistically valid if one uses a particular information criterion to select a covariance model? That is, under what conditions are the α_e's *controlled for models selected by a given information criterion?*

Secondary Research Questions

Review of the current research literature in the areas of longitudinal data analysis in psychology, education, and the social sciences in general, has demonstrated the continued wide-spread use of the classical methods of analysis (Keselman, Huberty, et al., 1998; Kowalchuk et al., 1996). Therefore, a comparison of α_e's and power estimates among classical and modern MM methods was a secondary interest of the current study. More specifically, the RM ANOVA conventional *F*-test, the Greenhouse-Geisser (G-G) and Huynh-Feldt (H-F) corrected *F*-tests, and the MANOVA Wilks' Λ test statistic were compared to the modern MM methods with respect to α_e's and power estimates.

Furthermore, the robustness of the interaction test was also of interest when a between-subjects factor was present in the design. Therefore, a two level between-subjects factor (Group) was added to the design in order to facilitate investigation into this question. These two secondary research questions are:

4) How does the mixed model approach compare to the classical methods of repeated measures analysis in the context of covariance model misspecification? More specifically, how does the mixed model Wald-type F-statistic compare to the RM ANOVA conventional F-statistic, the G-G or H-F corrections, or the MANOVA Wilks' Λ test statistic with respect to α_e's*?*

5) What are the mixed model α_e's *for the test of the interaction in repeated measures data with a between-subjects factor?*

For these purposes, data were generated and analyzed under three distinct phases of the current study using a Monte Carlo simulation. The purpose of phase I was to obtain α_e's for the test of the Time main effect for various modern MMs[7] as well as the classical methods of analysis. In contrast, the purpose of phase II was to obtain estimates of statistical power for the test of the Time main effect. Therefore, data for both phases I and II were generated from the single-group repeated measures design with one within-subjects factor (Time)[8]. Finally the purpose of phase III was to obtain α_e's for the Group x Time interaction test for mixed and classical models when a between-subjects factor (Group) was added to the design.

---

[7] Mixed models with various covariance models specified.
[8] No between-subjects factors were present in this design.

In the first two phases, data were generated for one homogenous group of *N* subjects with a quantitative response variable measured over multiple occasions. Five design features were manipulated: 1) population (or correct) covariance structure (with seven levels[9]), 2) the number of measurement occasions (with two levels: $t = 3$ & 6), 3) the magnitude of serial correlation (with two levels: $r = .3$ & .5), 4) the presence of non-constant variances over time (with two levels: either constant or non-constant variances), and 5) sample size (with three levels: $N = 10$, 30, & 60). Furthermore, the number of replications was set to 10,000 for both of these phases.

Phase I and II differed in the following ways. As mentioned above, the purpose of phase I was to evaluate α_e's. Consequently, data in phase I were generated so that the null hypothesis of no Time effect was known to true. As a result, data were generated under 72 conditions[10] for phase I. In contrast, the purpose of phase II was to estimate statistical power. Therefore, data in phase II were generated so that the alternative hypothesis of a Time effect was known to be true. In order to coerce the data appropriately for the alternative hypothesis to be true, an additional experimental factor was added in the phase II data generation process: Mean Effect. Mean Effect was specified with three levels corresponding to small, medium, and large effect sizes as defined by Cohen (1977, pp. 284-288). Thus, data were generated under the 72 conditions delineated by phase I for each of the three levels of Mean Effect. That is, data were generated under 216 conditions[11] in phase II. Other than these differences, phases I and II were identical.

---

[9] See section *Explication of Experimental Factors*, p. 62, for details.
[10] See Table 3.1 for a summary of the number of conditions.
[11] While data were generated in all 216 conditions, power estimates were only interpreted for those conditions where acceptable levels of α_e's were found.

In contrast, the purpose of phase III was to investigate α_e's for the interaction test of Group x Time for both mixed and classical models. As a result, phase III involved data generation where one between-subjects factor was added to the single-group repeated measures design of phases I and II. Similar to phase I, however, data were generated so that the null hypothesis was known to be true. Under these data generation conditions, all experimental factors remained the same as phase I except sample size. Levels of sample size for phase III were n_j = 5, 15, 30. For example, under the first level of the sample size factor, data were generated for groups 1 and 2 with five subjects each. Thus, group sample sizes were always held equal in the current study. Finally, due to time constraints, only 5,000 replications were performed for phase III. Otherwise, phase III procedures followed those of phase I closely.

This has been a brief discussion of the design of the current simulation study and how it relates to the research questions of interest. In summary, data were generated under three distinct phases: 1) phase I, a single-group repeated measures design with the null hypothesis known to be true, 2) phase II, a single-group repeated measures design with the alternative hypothesis known to be true, and 3) phase III, a design with one within-subjects and one between-subjects factors with the null hypothesis known to be true. A discussion of the significance of the current study follows.

Significance of the Study

The current study is judged to be significant because it addresses a number of limitations in the existing research literature. To date, the current author is unaware of any reported studies that take into account the potential effect of covariance model comparability on the accuracy rates of information criteria. In order to address the issue of surrogate influence on the accuracy rates of information criteria, rates of selecting *appropriate* covariance models are reported.

A second limitation in the existing literature is the influence of the number of candidate covariance models in information criteria accuracy rate estimation. Previous studies have used widely varying number of candidate models ranging from two to fifteen, and have reported highly variable marginal accuracy rates ranging from 35% to 79% (Ferron et al., 2002; Gomez et al., 2005; Keselman, Algina, et al., 1998; Vallejo and Livacic-Rojas, 2005). While the effect of this extraneous factor can not be neutralized completely, an attempt to minimize its influence was made in the current study. Based on a review of this literature (Davis, 2002, pp. 130-156; Diggle, 1988; Ferron et al., 2002; Littell et al., 1996, pp. 177-187; Tonidandel, Overall, & Smith, 2004; Verbeke & Molenbergs, 2000, pp. 98-101), nine covariance structures were identified as being those most commonly espoused by methodologists and used by applied researchers. These structures were targeted for use in the current study[12] to minimize the effect of being overly restrictive or overly nondiscriminatory in the number of candidate models one might realistically consider in a real-world modeling situation.

Finally, the existing research literature comparing accuracy of information criteria has highly favored comparisons of AIC and BIC. As mentioned previously, there are a number of other information criteria that have been proposed by methodologists that have not received equal treatment in the literature base. For example, Hurvich and Tsai (1989) have developed a finite sample version of the AIC statistic known as Akaike Information

[12] While the data were generated under only seven of these covariance structures, all nine structures were used as the candidate model set.

Criterion - Corrected (AICC). In order to arrive at the best possible estimate of the accuracy of information criteria in general, five information criteria were included for comparison in the current study: AIC, AICC, Hannan and Quin Information Criterion (HQIC), BIC, and Consistent Akaike Information Criterion (CAIC) (Akaike 1974; Hannan & Quin, 1979; Hurvich & Tsai; Schwarz; Bozdogan, 1987). Therefore, the current investigation is considered significant because it meets and addresses these limitations of the previous research.

Summary

In conclusion, classical statistical modeling of longitudinal data is often problematic because either 1) the RM ANOVA assumption of sphericity is untenable, or 2) the MANOVA approach suffers from insufficient power. As a modeling alternative, many methodologists suggest using modern MM methods. These methods obviate many of the problems of the classical approaches; however, new challenges arise with their use. Specifically, the researcher is faced with the task of selecting an appropriate covariance structure for the model in order to obtain valid tests of the fixed effects. While many methodologists suggest the use of information criteria to guide the selection process, problems with the accuracy of these criteria have been documented. Therefore, the current study was primarily concerned with 1) identifying comparable or surrogate covariance structures, 2) estimating the selection rates of *appropriate* covariance structures, and 3) estimating the α_e's of statistical tests whose covariance models were selected by a given information criterion.

The next chapter provides a review of research studies and published works that address information criteria development and use, the accuracy of information criteria, and investigations into α_e's and/or power estimates of methods for analyzing longitudinal data. Next, Chapter III outlines the specific methods employed in the current study to generate and analyze the data as well as criteria and methods for evaluating results and answering the research questions of interest. Finally, Chapter IV reports the results and Chapter V provides a discussion of those results and their implications in longitudinal data analysis in applied social science research.

CHAPTER II: LITERATURE REVIEW

As discussed in the previous chapter, modern mixed model (MM) methods have been suggested by many methodologists as a modern and flexible approach to analyzing longitudinal data. These methods obviate some of the potential pitfalls of the classical approaches to modeling data of this variety and therefore may appear attractive to applied researchers. However, new issues arise.

One of these issues is the selection of a covariance model. A covariance model must be chosen in order to obtain valid statistical tests of the fixed effects. The modern MM approach allows the researcher to use information from the data to guide the choice of the covariance model. This information includes the sample covariance/correlation matrix with respect to time, various graphical tools, and indicators of model fitness.

Researchers and methodologists alike generally favor the use of a specific class of model fit indicators: information criteria. However, recent literature has brought into question the accuracy of these indicators. It is the primary purpose of the current study to investigate the comparability of certain covariance models, the accuracy of a specific set of information criteria in choosing the correct and surrogate covariance models, and the empirical Type I error rates (α_e's) and statistical power estimates of models selected by a given information criterion.

For these purposes, literature covering the following four topics was reviewed: the use and development of information criteria; the accuracy of information criteria; empirical Type I error rates and statistical power estimates; and the comparison of modern MM and classical methods. These four topics provide a foundation for what is currently known regarding covariance modeling for MMs within the research community. Therefore, the current survey distinguishes the present boundaries of the literature, and further demonstrates the relevance of the current study.

The literature for this review was identified by using the ERIC, PyschInfo, and Web of Science databases as well as analysis of references from relevant published works. Keywords used in these searches included: repeated measures, longitudinal studies, ANOVA, MANOVA, mixed models, mixed effects model, information criteria, AIC, BIC, AICC, CAIC, HQIC, Type I error rates, statistical power, covariance misspecification, covariance surrogate, Huynh-Feldt, Greenhouse-Geisser, etc.

Information Criteria

This section reviews the use of information criteria in general and, specifically, the five different criteria that have been evaluated in the current study. This is intended to serve as an introduction to the history, theory, use, and differences among the five criteria.

As mentioned in the previous chapter, the general form of the information criteria is some function of the maximum likelihood function (usually negative two times the maximum likelihood) plus some penalty. The penalty term, in general, is a measure of model complexity; however, the actual realization of that measure is different among the differing information criteria. The penalty typically includes the number of parameters required to estimate the model and many times incorporates some measure of the sample size.

The general use of information criteria for model selection purposes begins with the fitting of multiple candidate models using a likelihood-based estimation method. It is important at this step of the process to systematically and logically vary only one aspect of the model (usually either the mean or covariance model) while keeping other dimensions of the model constant. This is important because only by varying one aspect of the model at a time is the researcher able to isolate the effects of a particular change in the model to model fitness.[13] The next step in the model selection process is to obtain the values of a preferred information criterion for each candidate model fit. Finally, these values are compared and the model with the most favorable value[14] is selected as the best model in terms of balancing model fit and parsimony. The five information criteria intended for inclusion in the current study are briefly presented below in historical order.

In the early seventies, Akaike (1974) introduced AIC which came to be known as Akaike's Information Criterion. The AIC was originally proposed to aid in model selection in the context of time series analysis (Akaike). The original formulation of this statistic is as follows:

$$\text{AIC} = -2(\text{maximum likelihood function}) + 2k, \qquad (2.1)$$

where k = the number of parameters in the model.

---

[13] Because the current study is primarily interested in the specification of the covariance model, the model for the mean was held constant across all model fits so that changes can be attributed to only the covariance model.

[14] "Smaller is better" parameterizations are currently used in SAS, therefore the candidate model with the smallest value is selected as the best model (Littell et al., 2006, pp. 183-184).

Equation (2.1) is the “smaller-is-better” parameterization, meaning that models that yield smaller values for this statistic are considered to provide better fit to the data; however, other parameterizations are in use (Littell et al., 2006, pp. 183-184). As one can see from (2.1), AIC is a function of the maximum likelihood and the number of parameters estimated in the proposed model.

In the 1974 article, Akaike references the sensitivity of the likelihood function in regard to deviations of model parameters from the true values in order to justify the use of the likelihood function to this end. Akaike further presented AIC as a mathematical realization of the principle of parsimony. He explicitly points out that if two models with an identical likelihood function were evaluated using this method, AIC would select the model with fewer parameters. That is, of two models with identical fit to the data, AIC selects the more parsimonious of the two.

Akaike demonstrated the use and performance of the statistic through several simulations, showing how the method selected between autoregressive (AR) and moving average models of various orders. Throughout the article and in closing, Akaike mentions the AIC method can be applied without subjective judgment. It is ironic that with the development of information criteria procedures (including, most notably, the development of other information criterion themselves), that the method has become more subjective. That is, with the development of more information criteria, researchers are forced to make a rather subjective choice as to which one to use in a given situation. Additionally, as was mentioned earlier, the performance of information criteria (specifically, the probability of selecting the correct model) is suspected to be affected by the number of candidate models initially entertained by the researcher. This adds another dimension of subjectivity to the process because the choice of the number and the identity of the candidate models to include is ultimately a subjective one left up to the individual researcher.

Schwarz (1978) presented a second information criterion, based on Bayesian statistics, as an alternative to AIC. This criterion has come to be known as the Bayesian Information Criterion (BIC) [or Schwarz's Bayesian Criterion (SBC)]. BIC is presented in the original article as a method to choose the appropriate dimensionality of a model to fit a given set of collected data. Schwarz derives BIC by finding the Bayes solution for a model in generality and then evaluating the maximum likelihood estimator (the leading term of the Bayes solution) in terms of its asymptotic properties. Because BIC is ultimately not dependent on the a priori distribution, Schwarz states that the resulting criterion is applicable in large-sample situations beyond a Bayesian context.

In the original article, Schwarz (1978) presents BIC in a "larger is better" parameterization. Because SAS[15] implements all information criteria in the "smaller is better" parameterizations, this formulation is presented below:

$$\text{BIC} = -2(\text{maximum likelihood}) + k\log(N)^{16} \tag{2.2}$$

where k = the number of parameters in the model and

N = the sample size.

Schwarz mentions that BIC differs from AIC only in the penalty term. As can be seen from above, the BIC penalty is not as great in terms of the number of parameters

[15] Version 9.1.
[16] All references to the log() function in the current document refer to the natural logarithm unless otherwise specified.

estimated in the model; however, the penalty also incorporates a new term: the natural logarithm of the sample size. Schwarz states that BIC tends to select lower dimensional models (or, in other words, more parsimonious models) than AIC when *N*, the sample size, is greater than 8. Along the same lines, differences in model selection using AIC and BIC become greater with large sample sizes.

Later, Hannan and Quinn (1979) introduced their version of an information criterion which came to be known as the Hannan-Quinn Information Criterion (HQIC). This criterion was developed in the context of estimating the order (or the dimensionality) of an AR model. According to the authors, the impetus behind developing HQIC was to address the established less than ideal asymptotic properties of AIC and BIC. In particular, AIC was found to be inconsistent and to overestimate the dimension of models in "large" samples. In contrast, BIC had been established as being highly consistent at that time, but Hannan and Quinn claimed that the BIC penalty term did not decrease fast enough as the sample size increased, thereby underestimating the dimension of models in large samples.

Hannan and Quinn's (1979) solution to this problem revolved around developing a penalty term in the equation that asymptotically decreased at a maximal rate. The SAS "smaller is better" parameterization appears below:

$$\text{HQIC} = -2(\text{maximum likelihood}) + 2k\log(\log(N))^{17}. \qquad (2.3)$$

The authors conclude their derivation of this statistic by comparing the performance of HQIC with that of AIC in simulation studies. These studies compared

---

[17] Both *k* and *N* are defined in 2.2.

HQIC and AIC across varying samples sizes (N = 50, 100, 200, 500, & 1,000). Results were as expected by the authors: HQIC underestimated model dimension in smaller samples relative to AIC, but outperformed AIC in large sample situations.

Bozdogan (1987) introduced the Consistent Akaike's Information Criterion (CAIC). This was an attempt to make AIC asymptotically consistent by adding a further adjustment to penalize more heavily for overparameterization and thereby reduce the probability of overfitting models. Bozdogan states explicitly the major issue with the information criteria approach: finding the optimal balance between risks of underfitting and overfitting. According to Bozdogan, the probability of underfitting or overfitting the correct model goes to zero for CAIC as the sample size approaches infinity. The SAS "smaller-is-better" parameterization appears below:

$$\text{CAIC} = -2(\text{maximum likelihood}) + k(\log(N) + 1). \qquad (2.4)$$

Bozdogan reports results from two simulations which show that CAIC is more accurate than AIC in sample sizes ranging from 50 to 200 with 100 replications. Specifically, the first simulation demonstrated that AIC overfit the model 14% to 20% of the time across the three conditions while CAIC did so only 1% to 3% of the time. In the second simulation, the error variance was increased uniformly and held constant across the conditions of differing sample sizes (N = 50 to 200). Here, accuracy rates ranged from 81% to 94% and 73% to 88% across the differing sample size conditions for CAIC and AIC, respectively.

Hurvich and Tsai (1989) introduced a finite sample version of the AIC criterion for regression and autoregression models. According to these authors, the trend of AIC overfitting models is a result of a substantial negative bias encountered when the dimension or complexity of a given candidate model increases in relation to *N*, the sample size. That is, this bias causes AIC to select a more complex model than the correct model on average. To address this issue, the authors introduced AIC corrected (AICC). AICC is defined as the sum of AIC and an additional non-stochastic penalty term incorporating sample size. According to these authors, this formulation minimizes the aforementioned bias and results in less overfitting. The SAS “smaller-is-better” parameterization appears below:

$$\text{AICC} = -2(\text{maximum likelihood}) + 2kN/(N-k-1). \qquad (2.5)$$

Hurvich and Tsai (1989) concluded by reporting the results of a few simulations they performed demonstrating the superiority of AICC in selecting the correct model in small sample situations. For example, the first simulation was designed to evaluate AICC performance in selecting the correct dimension of a regression model (that is, the correct number of predictors). The authors set the true model dimension to three and varied sample sizes between $N = 10$ & 20. Other information criteria such as AIC, HQIC, & BIC as well as criteria used in traditional regression model selection procedures such as Mallow’s Cp and the PRESS statistic were used to select models based on seven candidate variables[18]. One hundred replications were conducted. Results demonstrated that AICC selected the correct model 96 and 88 times out of 100 opportunities based on

[18] Therefore, the dimensions of the candidate models ranged from $m = 2$ to 7.

sample sizes of 10 and 20, respectively. Other criteria tended to overfit the data: AIC selected the correct model 36 and 64 times out of 100; BIC, 41 and 84; and HQIC, 24 and 70. The authors performed similar simulations for AR and moving average models. Based on these results, the authors conclude that AICC should be used routinely over AIC in regression and autoregression contexts.

This has been a brief review of five of the most common information criteria in current use in longitudinal data analysis (Davis, 2002, pp. 130-156;; Lindsey, 1999, p. 44; Pinheiro & Bates, 2000, pp. 253-256; Verbeke & Molenbergs, 2000, pp. 74-76). In summary, BIC, HQIC, and CAIC are all attempts to lessen the tendency of AIC to overfit models in large sample situations. Additionally, AICC was formulated to lessen overfitting by AIC in small sample situations.

These five information criteria were considered important in the current study for three main reasons. *First*, social science research involving longitudinal data often involve widely varying sample sizes (Keselman, Huberty et al., 1998) and the effects of using these different criteria have not been studied in this particular context. *Second*, AIC and BIC have been thoroughly studied, however, not along with HQIC, CAIC, and AICC. *Finally*, all five criteria are easily accessible in SAS version 9.1. Therefore, these five criteria were evaluated in terms of selection rates and α_e's in the current study. A review the literature specifically addressing the accuracy of these criteria follows.

The Accuracy of Information Criteria

There is a substantial literature base addressing the development, use, and accuracy of information criteria in general. However, there are far fewer published works that address the accuracy of information criteria strictly within the context of covariance modeling for modern MM methods and their use in the analysis of longitudinal data.

Keselman, Algina, Kowalchuk, and Wolfinger (1998) published results from a Monte Carlo simulation where they compared the performance of AIC and BIC in selecting covariance structures. The authors investigated the accuracy of these information criteria by simulating a repeated-measures-type study with three between-subjects groups and one within-subjects factor (Time). Data were generated using six population (i.e., correct) covariance structures: unstructured (UN), autoregressive with heterogeneity present with respect to time (ARH), random coefficients with heterogeneity with respect to time (RCH), unstructured with heterogeneity with respect of group (UNj), autoregressive with heterogeneity with respect to both time and group (ARHj), and random coefficient with heterogeneity with respect to both time and group (RCHj). AIC and BIC were used to select a preferred model from a set of eleven candidate covariance structures. Four variables operative in the data generation and overall analysis of the simulation were: varying sample sizes, equal and unequal group sizes, positive and negative pairings of covariance matrices and group sizes, and normal and non-normal data. These variables defined 26 distinct conditions under which performance of the AIC and BIC information criteria were evaluated. One thousand replications were executed under each of the 26 experimental conditions.

The authors report that neither criterion performed adequately in selecting the correct covariance structure. Results showed that on average, across the 26 conditions of the study, AIC selected the correct covariance structure only 47% of the time. Moreover, AIC performed notably worse in identifying complex covariance structures (i.e., UN).

However, AIC performed remarkably better when the correct structure was either ARHj or RCHj and the data were normally distributed as well as when the correct structure was either UNj or ARHj and the data were log-normally distributed. Results for BIC showed that this criterion only selected the correct structure 35% of the time and failed to select the correct structure at all in 14 out of the 26 conditions. The authors note that the BIC criterion more frequently selected an incorrect structure over the correct structure in all but a few cases.

Keselman, Algina, et al. (1998) and other authors (Gomez et al., 2005) suggest that at least one reason for the poor performance of these information criteria may be due to certain incorrect structures serving as adequate approximations or surrogates for the correct structure. If this were the case, then the accuracy rates of AIC and BIC reported may substantially underestimate the performance of these criteria.

Two unpublished works dealing with the effect of covariance model misspecification and the accuracy of information criteria in modern MM methods for longitudinal data analysis were also identified during the review of the literature. The first is an unpublished paper that was presented at the 12th annual conference on Applied Statistics in Agriculture by Guerin and Stroup (2000). The second is an unpublished master thesis completed in 1996 by J. M. Robertson at Brigham Young University. Both of these works are similar in purpose, design, and methods to both the Keselman, Algina, et al. (1998) article and the current study.

Guerin and Stroup (2000) were mainly interested in the behavior of information criteria, the effect of covariance misspecification on statistical inference, the comparison of the various degrees of freedom options available in SAS's PROC MIXED, and the frequency of convergence problems among differing covariance structures for modern MMs. To these ends, the authors employed a Monte Carlo simulation to generate data for a design with one between-subjects factor (Treatment) with two levels and one within-subjects factor (Time) with six levels and six subjects per group ($n_j = 6$).

The data were generated for three correct covariance models: AR, heterogeneous autoregressive (ARH), and antedependence (ANTE). These covariance models represented serial correlation in the generated data. In addition, the authors also included different levels of between-subjects random effects (see Methods section for more detail). For both the AR and ARH models, the AR parameter ρ was set to either .25 or .75. In addition, the between-subjects random effects were set to either 0.25, 1.00, or 4.00. As a result, there were six possible permutations for each of these structures. In contrast, only one permutation was specified for the ANTE model. Therefore, data were generated under a total of 13 conditions (six permutations of the AR model, six permutations of the ARH model, and one permutation of the ANTE model).

The authors evaluated the accuracy of AIC, BIC, HQIC, and CAIC in selecting the correct model from eight candidate models: CS, CSH, AR, ARH, TOEP, TOEPH, ANTE, and UN. Moreover, three different degrees of freedom options were evaluated: the containment method (the SAS default for these models), the Satterthwaite approximation, and the Kenward-Roger (KR) approximation. As a result, 24 analyses were performed on each dataset.

Similar to the models fit in the current study[19], Guerin and Stroup (2000) fit models that estimated unconstrained means at each measurement occasion. That is, they fit ANOVA-type models (the profile analysis approach) that did not assume linearity in

[19] Discussed in the third chapter.

the mean response over time. Further technical notes include the use of confidence intervals to evaluate the acceptability of α_e's. The authors evaluated robustness of the MM test statistics based upon α_e's for both Treatment and Time main effect tests and the interaction test. They also evaluated simple effect tests of Time at Treatment 1 and simple effect tests contrasting Treatment 1 and Treatment 2 at a given time point. The authors ran 500 replications for each condition due to time constraints.

With regard to the performance information criteria, the authors did not report marginal accuracy rates for each information criterion, but they did report rates with respect to the correct model. For the AR data, the authors stated that all criteria tended to select the CS model when the autocorrelation coefficient was low (.25); however, when the autocorrelation was high, the correct model was chosen most often. Furthermore, accuracy rates were generally low for all information criteria when the correct model was ARH and there was a high level of between-subjects random effect present ($\sigma^2_s = 4.00$). At lower levels of the between-subjects random effect, the correct model was selected most often by all information criteria. According to the authors, the ARH was most often chosen when the correct model was ANTE. Finally, BIC and CAIC were found to choose simpler models and AIC and HQIC more complex ones, as expected.

With regard to the degrees of freedom options, the authors found that the KR approximation was superior or, at the least, equal to the other options in terms of controlling α_e's. They also reported that fitting more complex models without the KR approximation resulted in substantially inflated α_e's. They did note, however, that even the KR approximation did not effectively control α_e's when the covariance model was misspecified.

In summary, consumers of this research should remain mindful that these results were obtained under only small sample conditions ($n_j = 6$). With this limitation in mind, the authors made several general statements. *First*, they suggest using the KR approximation in all similar applications of MMs. *Second*, they found α_e's to be severely out of control when models assuming constant variance were fit to data with heterogeneous variances. Therefore, they call for careful inspection of the sample covariance matrix with attention to possible heterogeneous variances. If such evidence exists, they suggest fitting a model that allows this heterogeneity. Even further, the authors suggest to err on the side of fitting more complex models, or overfitting, in order to minimize the chance of a liberal test. *Finally*, they point out that if one fails to model a between-subjects effect when one is in fact present, this may lead to inflated estimates of the covariance matrix. In order to minimize the impact of this problem, the authors suggest retaining any between-subjects random effects in the model that may be present and use both a covariance matrix for the random effects (**G**) and a covariance matrix for the association through Time (**R**) in order to model the covariance of **y**[20].

As with the Guerin and Stroup (2000) paper, the Robertson (1996) thesis consisted of a Monte Carlo simulation that closely parallels the Keselman, Algina, et al. (1998) article and the current study. The purpose of the thesis was to ascertain the effect of an optimal (correct) model for the covariance structure of MMs with respect to α_e of tests for the fixed effects, the distributions of those test statistics, and the effect of covariance model misspecification on α_e when the model was chosen by information criteria. Data were generated under a design with one between-subjects factor

[20] See Chapter 3 for a full discussion of these modeling issues.

(Treatment) with four levels and one within-subjects factor (Time) with five levels and with ten subjects per group ($n_j = 10$).

Additionally, data were generated with three correct covariance models: CS, autoregressive plus common covariance (AR+CC), and UN. The author states that parameter values for these covariance models were chosen so that the determinants of each covariance matrix would be “similar”. For the CS structure, values of five and ten were chosen for σ^2 and σ, respectively. The AR+CC model is a combination that introduces serial correlation in the data with respect to time and a between-subjects random effect that is the same for each subject. In order to fully specify this structure, values of 9.15, 0.60, and 10 were chosen for σ^2, ρ (the AR parameter), and σ_1 (the common covariance), respectively. Procedures for choosing the values of the UN structure were not discussed other than this matrix’s determinant was similar to the others. Finally, data were generated 10,000 times for each of these three conditions.

The study evaluated AIC, HQIC, BIC, and CAIC in choosing potential covariance models for these data. The set of candidate models was the same as the set of correct models (CS, AR+CC, & UN). Like the Guerin and Stroup (2000) study, 95% confidence intervals were constructed and used for evaluation purposes. Finally, for the CS and UN fitted models, the containment method[21] was used in order to calculate the denominator degrees of freedom of the F-tests. However, for the AR+CC fitted models, the

[21] Although, not explicitly stated, it appears that the author used SAS version 6.11. The containment method was the default for these models in this release of SAS.

Between/Within method was used in order to achieve the minimum denominator degrees of freedom for these test statistics[22].

The author reported that AIC performed well for all correct models with accuracy rates of .97, .88, and .75 for CS, AR+CC, UN structures, respectively. In addition, the author reported that performance of HQIC, BIC, and CAIC was excellent for simpler models. Accuracy rates for HQIC were .99, .94, and .14 for CS, AR+CC, and UN structures, respectively. BIC obtained rates of .99 and .92 for CS and AR+CC structures, but only .09 for UN. Similarly, CAIC obtained rates of .99 and .90 for CS and AR+CC, while never correctly choosing the UN structure.

In a similar fashion, the author reported the α_e of the fixed effects when the covariance model was chosen by a particular information criterion. Overall, AIC obtained a α_e value of .0816. In the same manner, HQIC, BIC, and CAIC obtained values of .0750, .0704, and .0694, respectively. For each fitted model (regardless of the correct model), AIC obtained values of .0524, .1299, and .1493 when CS, AR+CC, and UN models were chosen. Similarly, HQIC obtained values of .0485, .1308, and .4143 for CS, AR+CC, and UN chosen models. BIC obtained .0501, .0623, and .3333 for the same chosen models. In contrast, CAIC obtained values of .0495 and .0618 when CS and AR+CC were chosen, CAIC never chose the UN structure.

As an extension of these results, the author states that the CS model approximated more complicated structures well. Furthermore, inflated α_e were found for models fitted with AR+CC and UN even when these were the correct models. The author speculated

[22] The author notes that the DDFM option was used because, unlike the CS and UN model fits, the AR+CC fit utilized both RANDOM and REPEATED statements in the call to PROC MIXED in order to model both the serial correlation and the between-subjects random effects that were present in the data.

that this may be due to the containment method not computing the degrees of freedom correctly for these models.

Because the α_e rates were found to be markedly inflated in relation to the nominal α level for the AR+CC and UN fitted models (regardless of the correct model), the author conducted a brief investigation into the distribution of these test statistics. This investigation is largely tangential to the purpose of the current study; however, it is worth noting. Specifically, the author used two methods in an effort to identify the denominator degrees of freedom that would best specify the distribution of these *F*-tests. One method involved the Anderson-Darling goodness of fit statistic, the other, the method of moments. In both cases, however, Robertson obtained denominator degrees of freedom for AR+CC and UN that were considered unreasonably small. Moreover, these adjustments did not improve the rejection rates of the test statistics.

In conclusion, the author suggested that the different parameters for the structures be investigated and more correct structures included in future research. Additionally, the author suggested including a power analysis. Finally, the inclusion of the effect of missing data on α_e and information criteria performance was suggested as well.

Similar to the current study, the Keselman, Algina, et al. (1998), Guerin and Stroup (2000), and Robertson (1996) research were all primarily concerned with covariance model misspecification, α_e's, and the accuracy of information criteria in MMs used to analyze longitudinal data. Other articles that report the performance of information criteria in selecting covariance models for the modern MM approach often do so as a secondary point of interest. A brief review of this research is provided in the next section.

Empirical Type I Error Rates, Statistical Power Estimates, and the Accuracy of Information Criteria

The following published works are principally concerned with comparing statistical methods for analyzing longitudinal data in terms of robustness to assumption violations (as measured by α_e's and power estimates). However, secondary objectives included the accuracy of information criteria. Consequently, these articles are considered especially relevant in both their principal endeavors as well as their secondary treatment of information criteria accuracy to the current study.

Gomez, Schaalje, and Fellingham (2005) investigated the α_e's of the MM Wald-type *F*-tests for the fixed effects with the KR approximation. Of secondary interest, the authors evaluated the performance of AIC and BIC in selecting the correct covariance structure.

These authors used a Monte Carlo simulation to generate data for an experimental design with one between-subjects factor (Treatment) with three levels, and one within-subjects factor (Time) with three or five levels. Additionally, the data were generated with varying experimental conditions including correct covariance structure (with 15 correct structures), sample size (three or five subjects per Treatment), number of measurement occasions (three or five), equal and unequal variances between Treatment groups (with unequal variances being specified as three and five times greater than an arbitrary base group), degree of group size imbalance, and positive or negative pairings of covariance matrices and group sizes[23]. In all, the combinations of these experimental factors resulted in 60 conditions under which α_e's and information criteria accuracy were studied.

The authors investigated the two statistical tests for the main effects of Treatment and Time, but did not evaluate the interaction test. The 15 correct covariance structures under which the data were generated also served as the candidate models for AIC and BIC selection. The authors fit the correct model and those models selected by AIC and BIC only. Empirical Type I error rates were evaluated using a variation on the Bradley liberal criterion: if the empirical values fell within the interval [.022 - .080], they were deemed acceptable (Bradley, 1978).

Type I error rates for the correct model were close to the nominal α level for simple structures [compound symmetry (CS), Toeplitz (TOEP), and random coefficients (RC) (error rates ranging from .03 to .05)]. Error rates were reported to be fairly close to the nominal α level for more complex structures (mostly structures involving variance heterogeneity either with respect to Treatment or Time). However, error rates for more complicated structures such as the UN pattern and those that involved variance heterogeneity for both Treatment and Time were found to be much greater than the nominal level. Finally, the authors stated that error rates were best controlled when the correct model was fit and the correct structure was either CS or RC.

Results with regard to the performance of AIC and BIC were reported as follows. The authors state that accuracy rates for both criteria were generally low with rates as low as 3% to 30% and as high as 73.9%. In accordance with the findings of Keselman,

[23] A positive pairing occurs when the covariance matrix with the largest elements is associated with the largest group size. A negative pairing occurs when the covariance matrix with the largest elements is associated with the smallest group size.

Algina, et al. (1998), AIC was more accurate in the selection of complicated structures whereas the opposite was true for BIC. Moreover, α_e's were reported to be always higher than the nominal α value for both AIC and BIC; however, models selected by BIC obtained empirical error rates closer to the nominal value. Finally, the authors stated that accuracy of these information criteria depended mainly on sample size and the complexity of the correct covariance structure. Higher accuracy rates were obtained under conditions with larger sample sizes and simpler covariance structures.

The authors also reported consistency in regard to a given correct structure. A structure was considered consistent if the correct structure was chosen more often than an incorrect one. The authors noted that consistency occurred more often with larger sample sizes and that the TOEP structures were not consistent.

Finally, Gomez et al. (2005) reported convergence rates of the fitted models. As may be expected, convergence rates were found to be dependent upon the complexity of the true covariance structure. There was no mention of the influence of sample size on convergence rates; however, it should be noted that this study only dealt with small samples with sample size varying from $N = 9$ or 15 across the three between-subjects groups. The authors closed by mentioning the finite sample version of AIC (AICC) and suggested a similar investigation into the performance of this information criterion.

Ferron, Dailey, and Yin (2002) investigated the effects of covariance misspecification in the first-level error structure[24] for mixed-effects models. In the process, the authors also studied the accuracy of AIC and BIC as well as Likelihood Ratio Tests (i.e., LRTs) in identifying the correct structure. The primary interests of the

[24] The first-level error structure corresponds to the covariance matrix (**R**) for the random errors in the mixed-model formulation offered in Chapter 3.

investigators were the ability of the information criteria to distinguish between the correct and an incorrect model, the effect of error misspecification on the variance parameter estimates, estimates of the fixed effects, and tests of the fixed effects.

The authors generated data with one correct covariance structure: an AR structure. Furthermore, the authors only considered two candidate models: the AR model (correct) and the IN model (incorrect): $\sigma^2\mathbf{I}$. Data were generated under the following three experimental conditions: magnitude of autocorrelation (with two levels: ρ = .30 & .60), number of measure occasions (what the authors refer to as "series length") (with five levels: 3, 4, 6, 8, & 12), and sample size (with three levels: 30, 100, & 500). As a result, 10,000 replications were executed under the 30 different experimental conditions. Finally, the authors report using the Between/Within method for estimating the denominator degrees of freedom for the Wald-type *F*-statistic for the MM (the simulation was performed in SAS PROC MIXED with the option DDFW=BW).

Ferron et al. (2002) reported relatively high estimates of the accuracy of AIC and BIC compared to the other existing literature (Gomez et al., 2005; Keselman, Algina, et al., 1998, Vallejo and Rivica-Rojas, 2005). The mean accuracy rate for AIC across all conditions was 79% (ranging from 3% to 100%). Mean accuracy rates for BIC and LRTs were 66% and 71%, respectively. Longer series lengths and larger sample sizes along with higher amounts of autocorrelation were associated with higher accuracy rates. The authors used eta-squared (η^2) to partition the variability in these results and reported that series length was the most salient effect (η^2_{AIC} = .71, η^2_{BIC} = .77, & η^2_{LRT} = .73). In like manner, accuracy rate results also varied with sample size and a non-negligible series length by sample size interaction was reported with sample size having a greater effect when series length was short. These results coincided with the results from Singer (1998) where it was found that correctly specifying the covariance structure for the random errors is more difficult when fewer measurement occasions are available.

In regard to their second primary interest, Ferron et al. found substantial bias in the estimates of the elements of the covariance matrix for the random effects when the first level error covariance matrix was misspecified. Specifically, the diagonal elements of the covariance matrix for the random effects (the variances of the random effects) were overestimated and the off-diagonal element was underestimated. This bias was found to depend on the amount of autocorrelation and the series length. In addition, σ^2, the variance of the random errors, was found to be substantially underestimated when the first level error covariance was misspecified.

The authors found that estimates of the fixed effects were robust to misspecification of the error covariance matrix. The authors evaluated α_e's by constructing interval estimates for each condition and then checking whether or not the nominal α value of .05 was contained with these intervals. Type I error rates were found to be within acceptable ranges from .05 in all but two conditions and, unlike Gomez et al. (2005), Ferron et al. concluded that the statistical tests of the fixed effects were robust to error covariance misspecification. However, these studies used different correct and candidate models as well as widely varying sample sizes.

Keselman, Algina, Kowalchuk, and Wolfinger (1999) published research investigating the performance of the MM approach and the Welch-James-type multivariate test (WJ) in analyzing repeated measures data. Performance was compared based on Type I error rates and power analyses of both methods. In this simulation, the authors generated data with both a three-level between-subjects factor (Group) and a four-level within-subjects factor (Time). Additionally, data were generated with three different population (or correct) covariance structures: ARH[25], RCH, and UN. Other conditions under which the data were generated included sample size (with values of N = 30, 45, & 60); equal and unequal between-group covariance matrices (with values of either 1, 1, & 1; or 1, 1/3, & 1/5), etc.).

For the investigation of Type I error rates, AIC was used exclusively to select the covariance structure from eleven candidate structures. The authors referenced the findings of Keselman, Algina, et. al (1998) in regard to superior performance of AIC over BIC to support their exclusive use of AIC. Satterthwaite F-tests were used exclusively for the MM tests of the fixed effects. The Satterthwaite approximation is appropriate because the conventional F-test is known to be liberal in the presence of non-spherical data. The Satterthwaite method approximates the denominator degrees of freedom so that the test statistic's reference distribution more closely conforms to the F-distribution. Finally, to evaluate robustness, the authors adopted the Bradley liberal criterion by specifying the interval $0.5\alpha \leq \alpha \leq 1.5\alpha$ (Bradley, 1978). Thus, a given method was considered robust under a certain set of assumption violations if its α_e over 10,000 replications fell within this interval.

The authors reported that the MM approach effectively controlled Type I error rates across conditions when AIC selected the correct covariance structure. They continued by stating that applied researchers should not expect AIC to select the correct structure. Results also demonstrated that the MM approach controlled Type I error rates less adequately when the data were generated with complicated covariance structures

[25] In both of these case, heterogeneous indicates variance heterogeneity with respect to time.

(i.e., UN). Furthermore, the authors reported that the heterogeneous AR and random coefficient models for the covariance structure controlled the error rates adequately, regardless of the true covariance structure. Consequently, the authors suggested that a possible alternative to relying on AIC would be to always fit one of these two models.

For the power analysis, the authors chose a maximum range configuration for the vector of non-null main effect means. A maximum range configuration specifies two of the elements of the vector of means to be of magnitude μ; however, one is positive, the other negative. Moreover, all other elements are specified to be zero. The authors then used six permutations of this configuration that were possible with the four measurement occasions used in the study. Values of μ were selected so that an a priori target power value for the WJ test would be .50.

For the sake of brevity, the authors only investigated the power of the MM tests when the correct covariance structure was specified. In addition, they report excessive execution times for the power simulation. As a result, they executed the simulation over only enough conditions to confidently determine the outcome. Results demonstrated that the MM Satterthwaite *F*-tests were more powerful than the WJ test; however, the differences in power never exceeded 6% and were on average as small as 3.1%. The authors concluded that the power advantage of the MM approach is small and therefore the WJ test is in contention as a viable alternative.

In like manner, Vallejo and Livacic-Rojas (2005) performed a Monte Carlo simulation designed to compare the MM approach using the KR approximation and a multivariate extension of the modified Brown-Forsythe (BF) test. Similar to the Satterthwaite technique, the KR approximation is another method to adjust the distribution of MM test statistics to the *F*-distribution (Vallejo & Livacic-Rojas). More specifically, because the EGLS solution is derived using an estimate of the dispersion matrices of the random effects in the model, extra variability is introduced into the EGLS estimator. The KR approximation is used to correct the influence of this bias and account for the extra variability in the estimator and to provide denominator degrees of freedom for the resulting test statistics.

In many ways, the design of this study follows that of Keselman et al (1999). Vallejo and Livacic-Rojas generated comparable data with three between-subjects groups and a four-level within-subjects factor (Time), three population covariance structures (UN, AR, and random coefficients), and differing distributional shapes, etc. However, the authors used both AIC and BIC to select the covariance structure (at least initially) from a set of candidate models that included 15 alternatives as opposed to 11. Two hundred sixteen conditions were specified overall with 1,000 replications per condition. Unlike Keselman et al., who used the Bradley liberal criterion for evaluating robustness, Vallejo and Livacic-Rojas constructed and used a 95% confidence interval around the nominal alpha.

According to the authors, the results of the Type I error rate analysis demonstrated that MMs specified using AIC were occasionally liberal (especially for negative pairings of group size and covariance matrix magnitude and small sample sizes) and models selected by BIC were incapable of controlling Type I errors.

In terms of the performance of the information criteria, the authors reported the following. *First*, AIC accuracy rates ranged from 23% to 87% depending on the complexity of the structure being modeled. AIC correctly selected the random coefficients structure 87% of the time, 56% of the time when a heterogeneous AR structure was the correct structure, and 23% of the time for the UN pattern. *Second*, models based on both AIC and BIC produced liberal tests of the fixed effects under the negative pairings condition where the between-subjects group with the smallest sample size was paired with the covariance matrix with the largest elements (the most variability). However, this liberalness became less severe as the sample size increased and the magnitude of group inequality lessened.

For the power analysis, the authors used two permutations of the maximum range configuration for the non-null mean vectors. The parameter for these vectors, $\boldsymbol{\mu}$, was chosen so that the power for the Scheffe univariate MM would equal .70. The authors only evaluated the BF test and MMs selected by AIC due to the poor performance of BIC in regard to Type I error rate control. Results demonstrated that neither procedure was uniformly more powerful. Moreover, power estimates were found to be dependent on the mean configuration (permutation of the mean vector). Finally, the authors identified circumstances when the MM approach did not perform optimally: when the correct model requires many parameters to be estimated (i.e., an UN covariance structure) and small sample sizes. In these situations, the authors suggest the use of the BF test.

Comparing Mixed Model and Classical Methods

The review of the literature produced only one reported study that explicitly compared modern MM and classical methods. Wright and Wolfinger (1997) compared the RM ANOVA conventional *F*-test, the G-G and H-F corrections, the MANOVA, and the modern MM when all models were estimated within SAS's PROC MIXED. That is, RM ANOVA conventional model was specified in MIXED by fitting the CS covariance model and using the unadjusted degrees of freedom. Similarly, the G-G and H-F[26] corrected tests were obtained by fitting the CS model with the proper degrees of freedom adjustment, respectively. The MANOVA model was specified by fitting the UN structure with the unadjusted degrees of freedom. Finally, the modern MM was specified by fitting a number of correct covariance structures (e.g., CS, AR, TOEP, & UN) with the unadjusted degrees of freedom. Therefore, the modern MMs fit in this comparison were always specified with the correct model.

The authors compared these models across three estimation methods (i.e., REML, SSCP, & MIVQUE0) and three nominal α levels (α_n = .01, .05, & .10). Sample size (N = 15 or 60) and number of missing values (M = 0 or 3) were also systematically varied. Furthermore, the authors evaluated α_e's with respect to the test a Time main effect and the Group x Time interaction test; however, the authors state that results were similar and therefore only reported results for the test for Time. Nonetheless, the data were generated under a two-way repeated measures design with three groups and four measurement occasions.

Results showed the modern MM unadjusted F-test was liberal in small samples situations. More interesting, the authors found both the G-G and H-F adjustments to perform well under all estimation method by α_n level conditions. The authors do note, however, that the G-G corrected tests were found to be slightly conservative and the H-F corrected test slightly liberal in small sample situations.

[26] This was the Lecoutre corrected H-F adjustment.

While Wright and Wolfinger (1997) provided this initial comparison, which included some consideration of missing values[27], further research is needed. Specifically, this study only compared the modern MM approach with the classical methods when the correct covariance model was used in the modern MM. The effect of covariance model misspecification in modern MMs was not taken into account here. Moreover, this study did not evaluate the performance of the modern MM *F*-tests with the degrees of freedom approximations (i.e., Satterthwaite, Kenward-Roger) that are currently in common use in applied longitudinal data analysis. Because the modern MM *F*-tests are known to be liberal in small sample situations (Schaalje et al., 2002), evaluation of corrected MM *F*-tests is needed. Furthermore, these authors close by stating that a study of power characteristics of these modeling alternatives is needed.

Summary

This has been a brief review of the literature and introduction to the concepts central to the development and use of information criteria, the accuracy of information criteria in covariance model selection, and α_e's and power estimates for modern MM methods. This chapter began with a brief discussion of the general form of information criteria and their use with modern MMs in the longitudinal data context. Then, several research articles were reviewed. Many of these introduced their respective information criteria (AIC, AICC, HQIC, BIC, or CAIC) for the first time.

Next, three pieces of research[28] were reviewed that addressed the accuracy of information criteria, covariance model misspecification, and α_e's. Keselman, Algina, et

[27] The missingness conditions were either no missing values or 3 out of 15 x 4 = 60 observations.
[28] One published, two unpublished.

al. (1998) found AIC to be accurate in selecting the correct covariance model only 47% of the time when averaged across 26 experimental conditions. Furthermore, these authors found BIC to be accurate only 35% of the time, while not selecting the correct model at all in 14 out of 26 conditions. Finally, Keselman, Algina, et al. suggest these low rates may be affected by the existence of surrogate covariance models.

In a similar investigation, Guerin and Stroup (2000) evaluated information criteria accuracy, α_e's, and degrees of freedom options available for computing MM test statistics. The authors reported that the KR degrees of freedom approximation provided superior Type I error control over the Containment and Satterthwaite methods; however, the KR approximation did not perform well when the covariance model was misspecified. Furthermore, the authors reported that Type I error control was especially poor when models assuming constant variances were fit to data that exhibited non-constant variances. Finally, the authors suggest to overfit the data (i.e., fit a more complex model than what is perceived as being needed) to minimize the chance of a liberal test.

In the Robertson (1996) thesis, high estimates of information criteria accuracy were reported, with estimates for simpler models as high as .97 for AIC and .99 for HQIC, BIC, and CAIC. However, these estimates are based on a selection situation where only three candidate models were used. Other results showed that α_e's were inflated when information criteria were used to select the covariance model. Values of α_e ranged from.069 to .082, being most inflated when AIC was used to select the model. Finally, α_e's were inflated when the UN covariance model was fit to the data regardless of the true structure and the CS model was found to be a good fit to the data regardless of the true structure.

While comparing statistical methods in the context of longitudinal data, the following authors also investigated covariance model selection and information criteria accuracy. Gomez et al. (2005) found α_e's to always be higher than the nominal value when models were selected with either AIC or BIC. Furthermore, these authors report accuracy rates ranging from 3% to 79% for particular experimental conditions with accuracy greater under conditions where the sample size was large and the correct covariance model was simple.

Ferron et al (2002) reported accuracy rates of 79% and 66% for AIC and BIC, respectively, when averaged across all experimental conditions. Once again, however, the data were generated with only a small set of candidate models: the correct model (AR) and the IN model.

Unlike Robertson (1996), Keselman et al. (1999) found α_e's to be controlled when AIC was used to select the covariance model. However, these studies used different methods for computing the degrees of freedom for these test statistics: Robertson used the Containment method, while Keselman et al. used the Satterthwaite method. Even more importantly, these studies used different criteria for defining acceptable levels of α_e's: Robertson used the 95% confidence interval method (based on $N = 10{,}000$ replications), while Keselman used the Bradley liberal criterion. In their analysis of statistical power, Keselman et al. found that MMs were uniformly more powerful than the Welch-James-type multivariate test (WJ); however, only by 3.1% on average and never more than 6%.

Vallejo and Livacic-Rojas (2005) found neither the MM nor the modified Brown-Forsythe test to be uniformly more powerful. Furthermore, these authors note that the MM approach did not perform well in controlling Type I error rates when the sample size was small and the population covariance structure was complex.

Finally, Wright and Wolfinger (1997) compared modern MM and CLASSICAL methods. These authors found the G-G and H-F corrected tests to perform adequately when compared to modern MMs specified with only the true covariance structure and unadjusted degrees of freedom for the modern MM *F*-test. Further research is needed in order to compare these methods when: 1) the modern MM is specified without prior knowledge of the true covariance structure, 2) the degrees of freedom for the modern MM *F*-test have been adjusted using methods currently in common use, 3) more extreme cases of missing values are encountered, and 4) when statistical power is considered.

In light of this review, three issues concerning the use of information criteria and modern MMs in longitudinal data analysis are still unresolved. *First*, only the Guerin and Stroup (2000) paper evaluated degrees of freedom options (Containment, Satterthwaite, and Kenward/Roger) for the modern MM test statistics. Furthermore, their analysis was only performed in small sample situations ($N = 12$; $n_j = 6$). Thus, test statistic options including three degrees of freedom options (Between/Within, Satterthwaite, and Kenward/Roger) as well as one option for estimating empirical or robust error variances for these statistics (i.e., the sandwich estimator) were compared in the current study under small, moderate, and large sample size conditions. *Next*, while several authors made casual statements concerning the relative good fit of certain models to data in general, no systematic investigation into the existence of surrogate covariance models has been performed. Therefore, the current study systematically investigated the existence of surrogate covariance models and the accuracy of information criteria while accounting for these models. *Finally*, several authors have compared other statistical models with the modern MM approach; however, no simulation studies were identified that compared the modern MM methods with the classical univariate and multivariate approaches.

This has been a review of the literature and a discussion concerning how the current study augments and expands the existing literature base. The next chapter outlines the methods that were used in the current study.

CHAPTER III: METHODS

The current study is concerned with 1) comparing differing methods for computing test statistics for modern mixed models (MMs) with respect to empirical Type I error rates (α_e's), 2) identifying surrogate covariance structures, 3) estimating the rates of selecting appropriate covariance structures for five information criteria, 4) estimating the α_e's if a given information criterion is used to select covariance models, 5) comparing modern MM methods with classical methods in the analysis of longitudinal data, and 6) evaluating the α_e's for the interaction test of a within-subjects/between-subjects design.

In order to address these research objectives, the current study was designed as a Monte Carlo computer simulation. A Monte Carlo study is one that involves generating data with a stochastic component adhering to a specified distribution to simulate the characteristics and/or behavior of a particular phenomenon under specified conditions (Hutchinson & Bandalos, 1997). In the current study, this equates to the investigation of the behavior of candidate covariance models in relation to the correct model and the accuracy of information criteria under conditions of varying sample sizes, magnitude of serial correlation, etc.

Simulations of this type usually employ computer software that uses a pseudo-random number generator. In the current study, data were generated and analyzed using a combination of software packages including Statistical Analysis System (SAS) version 9.1 and the statistical computing environment known simply as R version 2.4.0[29].

Furthermore, data were generated under three distinct phases. The first phase was designed to evaluate α_e's of test statistics for data with one within-subjects factor, Time. In contrast, the second phase was designed to evaluate empirical statistical power of the same test statistics and experimental design evaluated in phase I. Finally, the third phase of the current study was designed to evaluate the α_e's of the interaction test for data with one repeated measures factor (Time) and one between-subjects factor (Group). Specific details concerning the design, statistical methods under evaluation, and the generation of data for these different phases are provided in this chapter.

However, before describing the procedures to generate the data, obtain the outcome variables, and answer the research questions; a brief introduction to the classical linear model (CLM), the MM, and a discussion of their similarities and differences is warranted. Next, a full treatment of the experimental conditions in the current study is provided. Finally, specific details concerning the data generation and analysis procedures in all three phases are presented.

The Classical Linear Model

The CLM has seen wide-spread use in statistical modeling in innumerable disciplines of science. The CLM is typically used to model the variation of a normally distributed continuous response or outcome variable (**y**) with respect to a single or

[29] R is a freeware version of the computing language S-plus, originally developed by Bell Laboratories.

multiple explanatory variable(s) (**X**). The vector-valued formulation of the CLM appears below for the univariate case (Christensen, 1987, pp. 1-3):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.1}$$

where: $\mathbf{y}$ = a vector of observed responses,

$\mathbf{X}$ = a known model/design matrix ,

$\boldsymbol{\beta}$ = a vector of unknown model parameters, and

$\boldsymbol{\varepsilon}$ = a vector of unobserved random errors,

with the following mean, variance, and distributional assumptions:

$$\boldsymbol{\varepsilon} \sim \text{iid N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

where iid = independent, identically-distributed,

N = Normal,

$\mathbf{0}$ = a mean vector of zero's,

σ^2 = a constant variance estimated from the data, and

$\mathbf{I}_n$ = the identity matrix of order (n x n).

Briefly, **y** is a column vector of response values, one response for each experimental unit. The matrix **X** is a known design matrix that assumes a different formulation depending on whether the regression or the analysis of variance approaches are adopted. In the regression case, **X** is composed of a column vector of 1's and *p* column vectors of values for each of the p explanatory variables. In the ANOVA case, **X** is composed of p column vectors[30] of coded variables indicating group membership. In either case, $\boldsymbol{\beta}$ is a column vector of length p, containing unknown population-valued coefficients. In the CLM, there is one random effect: $\boldsymbol{\varepsilon}$, a vector of random errors. As mentioned earlier, the CLM is usually used to model data where the response variable is continuous and normally-distributed. Typically, however, these assumptions are expressed in terms of the distribution of the errors, $\boldsymbol{\varepsilon}$. In the usual situation, the errors are assumed to be independent. That is, the errors are assumed to show no correlation or association among themselves. In addition, the errors are assumed to be identically-distributed with a normal distribution. Finally, they are assumed to have a mean of zero and an unknown constant variance (that is estimated from the data).

This brief review of the CLM serves as a convenient referent with which to juxtapose the modern MM. The modern MM is a generalization of the split-plot model, which, in turn, is an extension of the CLM (Fitzmaurice, Laird, & Ware, 2004, pp. 187-197; Rencher, 2000, pp. 426-429; Vallejo & Livacic-Rojas, 2005). The extension from the CLM to the split-plot model allows for the inclusion of more than one random effect in the split-plot model which facilitates the analysis of clustered or hierarchical (nested) data. Furthermore, certain assumptions concerning the independence of the observations are relaxed in such a way that observations are allowed to covary. The generalization from the split-plot model to modern MM allows for further flexibility in the manner in which observations are allowed to covary and also allows for the modeling of nonconstant variances. As is shown in the next section, the modern MM is able to account for these factors through the explicit modeling of the covariance matrix of the

[30] Where $p = J - 1$; J = the number of groups.

errors (and also the covariance matrix of the random effects). Finally, it is shown that the increased flexibility of the MM in dealing with these factors is especially relevant and useful in the analysis of longitudinal data.

The Mixed Model

Mixed model is a general term that refers to an entire class of statistical models that are typically characterized by the inclusion of both fixed and random effects. These models are used in many disciplines and, as a result of this versality, are known under many different names, including mixed effects models and covariance components models in statistics, mixed linear models in biomedical research, random coefficient models in econometrics, and hierarchical linear models and multilevel models in psychology, sociology, education, and the social sciences in general (Bryk & Raudenbush, 2002, pp. 3-4; Littell et. al, 2006, p. 161). As mentioned above, the MM is an extension of the classical or general linear model that does not require standard assumptions concerning independence and homogeneity of error variances (Vallejo & Livacic-Rojas, 2005). As mentioned above, the model typically is composed of both fixed and random effects; that is, random effects beyond the usual stochastic error term in the CLM. The usual parameterization of the MM appears in vector-valued form below.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \tag{3.2}$$

where: $\mathbf{y}$ = a vector of observed responses,

$\mathbf{X}$ = a known model/design matrix for the fixed effects,

β	=	a vector of unknown model parameters,
Z	=	a known model/design matrix for the random effects,
b	=	a vector of unknown random effects, and
ε	=	a vector of unobserved random errors,

with the following mean, variance, and distributional assumptions:

1) $\boldsymbol{\varepsilon} \sim \text{iid } N(\mathbf{0}, \mathbf{R})$,
2) $E(\varepsilon_{ij}\, \varepsilon_{kl}) = \text{cov}(\varepsilon_{ij}, \varepsilon_{kl}) = 0$ for $i \neq k, j \neq l$,
3) $\mathbf{b} \sim \text{iid } N(\mathbf{0}, \mathbf{G})$, and
4) $E(b_i \varepsilon_{ij}) = \text{cov}(b_i, \varepsilon_{ij}) = 0$ for all i & j.

Briefly, the marginal or population mean of **y** is modeled by **Xβ** (Fitzmaurice et al., 2004, pp. 187-192). This is analogous to the CLM treatment of the mean for **y**. However, the MM also allows cluster or subject-specific random effects above and beyond the usual error term of the CLM to be incorporated in the model. This is achieved through the specification of the vector **b** and the design matrix **Z**. In this formulation, **Z** is a subset of **X** which links the vector of random effects to the conditional or subject-specific mean of **y** (Fitzmaurice et al., pp. 192-197).

Thus, **β** is a vector of unknown regression coefficients for the population of interest and is considered to be true for all individuals within that population when averaged across individual differences. On the contrary, the vector **b** is composed of cluster or subject-specific deviations from $\boldsymbol{\beta}$. Therefore, it follows that the conditional or subject-specific regression weights for the i^{th} subject are $\boldsymbol{\beta} + \mathbf{b}_i$ and the conditional or subject-specific mean for the i^{th} subject is $\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i$ (Fitzmaurice et al., 2004, pp. 192-197).

Moreover, the mean, variance, and distributional assumptions of the MM state the following. *First*, the errors are normally distributed with mean zero and variance that is modeled in a covariance matrix **R**, which is not required to take the form of a constant variance term times the identity matrix ($\sigma^2\mathbf{I}$). *Second*, the errors among experimental units are assumed to be independent. That is, the errors among units are assumed to be uncorrelated; however, the errors of multiple measurements on the same unit are allowed to covary. *Third*, the vector of random effects are distributed normally with mean zero and variance that is modeled in the covariance matrix **G**, which, like **R**, is flexible and able to take different forms. *Finally*, the random effects specified in **b** are statistically independent of the random error term, $\boldsymbol{\varepsilon}$.

In the MM, the variance of $\mathbf{y_i}$ is modeled through $\mathbf{Z_i}$, **G**, and $\mathbf{R_i}$ (Fitzmaurice et al., 2004; Wolfinger, 1993):

$$\text{cov}(\mathbf{y_i}) = \boldsymbol{\Sigma}_\mathbf{i} = \mathbf{Z_i G Z_i}' + \mathbf{R_i}. \tag{3.3}$$

The possibilities present in the modeling of the variance of **y** through **Z**, **G**, and **R** constitute a substantial extension of the general linear model (Wolfinger, 1993). While these options provide remarkable flexibility and may be especially important when modeling data with complex clustered structures and repeated measurements over Time, this flexibility is not always needed, may sometimes be redundant, and may also contribute to complications.

Guerin and Stroup (2000) state that identifiability issues often cause nonconvergence when both **G** and **R** are used to model the covariance of **y.** This is especially true for situations when one uses **G** to model the covariance of the between-subjects random effects and simultaneously specifies **R** to be UN. This is analogous to over-parameterizing the model. In order to avoid this problem, Guerin and Stroup state that researchers often drop the between-subjects random effects and use only **R** to model the covariance in **y**. However, they do mention that this may have an undesirable impact on the estimates of standard errors if there is a substantial between-subjects random effect present in the data.

In similar fashion, Wolfinger (1993) mentions the Lindstrom and Bates (1988) approach where these authors set $\mathbf{R} = \sigma^2\mathbf{I}$ (the general linear model specification of the variance of the error term) and use only random effects to model variability. In similar fashion, Wolfinger uses Liang and Zeger (1986) as an example of authors who chose to model all variability through **R**. In this case, **R** becomes block-diagonal with blocks of $\mathbf{R}_i$ corresponding to measurements from the same subject (Wolfinger).

Fitzmaurice et al. (2004, p. 195) discuss two complications of using both **G** and **R** to model the covariance of **y**. *First*, when $\boldsymbol{\varepsilon}$ is distributed with an independence covariance model (IN) (either $\mathbf{R} = \sigma^2\mathbf{I}$ or diag($\sigma^2_1, \sigma^2_2, \ldots, \sigma^2_t$)), $\boldsymbol{\varepsilon}$ has the straightforward interpretation as being simply the sampling error. However, when the cov($\boldsymbol{\varepsilon}$) takes forms that allow dependence, the implication is that $\boldsymbol{\varepsilon}$ incorporates a certain amount of model misspecification. Consequently, the interpretation of $\boldsymbol{\varepsilon}$ is altered (Fitzmaurice et al., p.

195). *Second*, Fitzmaurice et al. (p. 195) mention the model identification issues that often arise when **R** takes a non-diagonal form[31]. Therefore, like Lindstrom and Bates (1988), Fitzmaurice et al. (p. 195) model the covariance of **y** through **G** exclusively and set $\mathbf{R} = \sigma^2\mathbf{I}$.

Because the data generated in the current study are of a single-group repeated measures format (with no other clustering and/or hierarchical structure present), the variability of **y** was modeled strictly through the **R** matrix (see Methods section Data Generation and Model Fitting Procedures Technical Notes: Phase I for more detail). Therefore, the term *mixed model* in the current document refers to a particular model within the mixed model methodology that accounts for the variance of **y** exclusively through the **R** matrix. Unlike the typical mixed model formulation, this model does not contain the random effect terms **Z** or **b**. The model is presented as a subset of (3.1) below:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.4}$$

where $\boldsymbol{\varepsilon} \sim \text{iid } N(\mathbf{0}, \mathbf{R})$.

The capacity of MM methods to account for correlated data exemplifies the increased flexibility of these methods over that of the CLM and makes them especially suitable for longitudinal data analysis.

Model Estimation. There are various methods available for estimating covariance parameters ($\boldsymbol{\Sigma}$) like those associated with MMs. However, Littell et al. (2006, p. 747) identified Restricted (or Residual) Maximum Likelihood Estimation (REML) as the most

[31] In other words, when **R** accounts for some dependence among the data.

important technique of MM estimation. REML is a likelihood-based method that attempts to address the deficiencies of regular maximum likelihood (ML) estimation.

It is well known that ML estimators have good asymptotic properties. It is equally well known that ML estimators for σ^2 are biased in finite samples. This bias is a function of using estimates for the fixed effects in the estimation of the covariance parameters without taking into account the extra variability (Fitzmaurice et al., 2004, pp. 99-102; Littell et al., 2006, pp. 746-750).

REML addresses this problem by minimizing an objective function (known as the restricted, residual, or modified log-likelihood) where the fixed effects have been removed from the equation. Fitzmaurice et al. (2004, p. 99-102) state that this is analogous to estimating $\mathbf{\Sigma}$ using only information from the data that are independent of the fixed effects. Therefore, REML estimators lessen bias of $\mathbf{\Sigma}$.

For these reasons, REML estimates of $\mathbf{\Sigma}$ are commonly favored in MM estimation. However, REML does not provide for estimation of the fixed effects ($\boldsymbol{\beta}$) in the model (Vallejo & Livacic-Rojas, 2005). For this reason, Generalized Least Squares (GLS) estimation is often used with the REML estimate of $\mathbf{\Sigma}$ substituted in for $\mathbf{V}$ matrix in the GLS equation, yielding Estimated Generalized Least Squares (EGLS) estimators (Vallejo & Livacic-Rojas). Previous authors have reported that EGLS parameter estimators of the fixed effects are unbiased with large sample sizes, even when the covariance structure of the data is misspecified (Fitzmaurice et al., 2004, pp. 90-91; Vallejo & Livacic-Rojas). These are the estimation methods implemented by SAS's PROC MIXED (Littell et al., 2006, pp. 747-750), and therefore the ones used in the current study.

Model Based Inference. Wald-type *F*-distributed statistics for testing hypotheses concerning the fixed effects (**β**) under the null hypothesis are available for MMs. Littell et al. (2006, p. 25) state, however, that the default option in SAS's PROC MIXED (i.e., the Between/Within method) is to compute denominator degrees of freedom for these tests based on the traditional analysis of variance assumptions using an IN model for the error covariance structure[32]. Additionally, PROC MIXED computes naïve standard errors by using computations for deriving these standard errors that assume estimated covariance parameters are known. Kackar and Harville (1984) have shown that using estimated covariance parameters as if they were known quantities biases standard error estimates and test statistics. Specifically, these authors found estimates of standard errors to be underestimated and test statistics involving these results to be overestimated.

In order to address these issues, many authors promote the use of the Satterthwaite degrees of freedom approximation (Fitzmaurice et. al., 2004, pp. 98-99; Keselman et al., 1999; Verbeke & Molenbergs, 2000, p. 57). In both repeated measures and split-plot models, the error variance estimator of the *F*-statistic is a linear combination of mean squares (Kutner, Nachtsheim, Neter, & Li, 2005, pp. 1130-1134). As a result, these statistics do not exactly follow the *F*-distribution. The Satterthwaite method is a technique to approximate the denominator degrees of freedom of these *F*-statistics to adjust their reference distribution so as to more closely approximate the *F*-distribution.

As an alternative to the Satterthwaite method, many authors suggest the use of the Kenward-Roger (KR) approximation (Fitzmaurice et. al., 2004, pp. 98-99; Gomez et al., 2005; Guerin & Stroup, 2000; Littell et al., 2006, pp. 188). Similar to the Satterthwaite

[32] I.e., assuming independence of observations.

method, the KR approximation is a technique to adjust MM Wald-type *F*-statistics to more closely follow the *F*-distribution. The KR approximation achieves this by 1) providing improved standard errors for parameter estimates that account for the extra variability in their estimation and 2) providing denominator degrees of freedom so that the distributions of test statistics more closely conform to the *F*-distribution. Littell et al. strongly recommend the use of the KR approximation in the analysis of data with repeated measurements.

Fitzmaurice et al. (2004, p. 177) also discuss the use of the sandwich estimator with MMs and its effects on model based inference. Although the sandwich estimator is more often used in the analysis of discrete longitudinal data with generalized estimating equations (GEEs), it can be applied to MMs for continuous data (Fitzmaurice et al., p. 177). The sandwich estimator is an attractive option for computing test statistics because it is widely known to produce estimates that are robust to covariance model misspecification (Fitzmaurice et al., p.177; Manor & Zucker, 2004). That is, the sandwich estimator produces asymptotically consistent estimates of the covariance matrix without distributional assumptions and when the covariance model is either misspecified or not specified at all (Kauermann & Carroll, 2001). However, some authors have reported poor performance of the sandwich estimator in small sample situations (Diggle et al., 2002, p. 74; Manor & Zucker).

To date, the current author is aware of no reported study that compares the use of the Between/Within method (the PROC MIXED default setting), the Satterthwaite and KR approximations, and the sandwich estimator in obtaining test statistics for the Time factor in repeated measures analysis[33]. Given the importance of these options in the context of computing test statistics and forming statistical inferences based on MMs, the current study investigates possible differences among these with respect to α_e's.

Model Fitting Procedures. Many authors advocate similar procedures for fitting MMs using three or four distinct steps (Diggle, 1988; Fitzmaurice et al., 2004, pp. 173-177; Littell et al., 2006, p. 161; Wolfinger, 1993). These steps typically are as follows.

1) Fit a saturated or maximal model in terms of the mean structure to the data. That is, fit a model that is most flexible in the mean structure, even to the point of over-parameterization.
2) Model the covariance of **y** using the methods described in the current document based on information criteria and/or those alternative methods.
3) Revisit modeling the mean structure once a suitable model for the covariance of **y** has been selected. It is now appropriate to more parsimoniously model the mean structure, if desired.
4) Estimate the final model and perform statistical inferences based on these model parameter estimates.

Fitzmaurice et al. (2004, pp. 173-177) stress the importance of initially fitting a maximal model for the mean structure of the data. According to these authors, misspecifying the mean model can result in introduction of spurious covariance among the residuals of the model. Because the covariance model attempts to account for the

[33] In certain situations, these options provide identical results. For example, all three methods produce identical results when applied to data with a CS population covariance model. Furthermore, the Satterthwaite and KR approximations produce the same results when applied to IN data.

association among and variability of these residuals, initial mean model misspecification may inflate the covariance and negatively impact the covariance modeling process through the introduction of this bias (Fitzmaurice et al.). Consequently, all potential candidate model fits in the current study incorporate a saturated or maximal model for the mean (see Methods section Data Generation and Model Fitting Procedures Technical Notes: Phase I for more details).

This section has been a brief overview of the usual parameterization and distributional assumptions, estimation, inference, and model fitting procedures for the MM. Detailed descriptions of the experimental conditions and data generation procedures in the current study follow.

Explication of Experimental Factors

There are five main experimental factors under which data were generated in the current study: 1) population covariance structure (with seven levels), 2) number of measurement occasions (two levels), 3) magnitude of serial correlation (two levels), 4) presence of non-constant variance (two levels), and 5) sample size (three levels). All permissible combinations of these factors[34] determine 72 experimental conditions under which data were generated (see Table 3.1, p. 94). Further description of each factor and its levels follow.

1) Population Covariance Structure

The first experimental factor under which the data were generated is the population covariance structure. A survey of the applied and methodological literature in

[34] The population covariance structure and non-constant variance factors can not be completely crossed. See Table 3.1 and the end of this section for explanation.

the area of longitudinal data analysis identified a number of structures that are commonly employed in the context of applied data analysis and are highly favored by methodologists (Davis, 2002, pp. 130-134; Fitzmaurice et al., 2004, pp. 166-173; Verbeke & Molenbergs, 2000, pp.98-101). For example, Wolfinger (1993) identified the diagonal (i.e., IN), compound symmetry (CS), UN, and AR structures as the most common in use. However, in order to focus on longitudinal data analysis where the variances are most often found to increase with time, covariance structures that allowed heterogeneity among the variances were identified as the most appropriate to include for data generation purposes.

Therefore, the current study investigates the possibility of surrogate structures for seven population or correct covariance structures: independence (IN) (i.e., simple), variance components (VC), compound symmetry (CS), compound symmetry with heterogeneity present with respect to time (CSH), autoregressive with heterogeneity with respect to time (ARH), Toeplitz with heterogeneity with respect to time (TOEPH), and UN. The IN and VC structures were identified because they represent the situation where the data show little or no correlation (i.e., independent observations). The CS structure was identified because it represents the ideal situation for the classical univariate methods where the assumption of sphericity is met. The ARH and TOEPH structures were chosen because they represent situations that are especially typical of longitudinal data: 1) decreasing correlation among the data with increasing separation in time, and 2) increasing variances with time. Finally, the UN structure was selected because it represents the situation where the classical multivariate and the MM approaches may perform optimally.

Therefore, the data in the current study were generated with these seven covariance structures. However, it was of interest to evaluate the consequences of fitting models that assumed homogeneous variances to data that exhibit variance heterogeneity. Therefore, nine covariance structures were used as candidate models in the current study: the seven previously discussed plus the AR and TOEP structures. A brief description of each of these nine covariance structures follows. These structures are displayed pictorially and in generality in Table 3.2 (pp. 95 – 97).

Independence (IN). This pattern is the covariance structure assumed by the general linear model with a constant variance value on the main diagonal elements of the matrix, and all off-diagonal elements set to zero.

Variance components (VC). This is a generalization of the IN pattern that allows for unequal variances on the main diagonal with all off-diagonal elements still set to zero.

Compound symmetry (CS). The CS structure is a covariance pattern where each main diagonal element is decomposed into $\sigma^2 + \sigma_1$ and all off-diagonal elements are set to the value of σ_1. This covariance structure was originally adapted to the analysis of repeated measurements from the split-plot design analysis. It requires that the variances of differences of all possible pairs of measurements within clustered structures be constant. This assumption is generally considered tenable in the split-plot design where levels of the sub-plot are randomly assigned to units within main plots and the measurements are taken simultaneously. However, in the analysis of longitudinal data, this is equivalent to requiring that the relationships of measurements spaced further apart in time are equal to those spaced closer in time. This assumption is typically not considered tenable in applied longitudinal data analysis. It should be noted that the CS pattern is generally synonymous with the RM ANOVA assumption of sphericity.

Heterogeneous Compound Symmetry (CSH). This pattern is a simple generalization of the CS structure that allows for non-constant variances on the main diagonal. While the assumption of sphericity is met in the CS structure, sphericity is not met in CSH due to the introduction of variance heterogeneity.

Autoregressive (AR). The AR structure was adapted from time series analysis where it is often used to account for the internal structure or autocorrelation of observations over time. It models the relationships of the measurement occasions parsimoniously by imposing a decreasing exponential trend between time lags. That is, adjacent measurement occasions (time lag 1) are specified to have the same covariance. Occasions that are two time points apart (time lag 2) are also all specified to have the same covariance; however, they are modeled as having exponentially decreased from the covariance at time lag 1. Thus, the AR structure is only appropriate for those instances where measurements are collected at equally spaced occasions. The AR structure has two parameters ($k = 2$): one for constant variance and ρ, a parameter governing the rate of exponential decrease between covariation at successive time lags.

Heterogeneous autoregressive (ARH). This is a simple generalization of the AR pattern that allows for non-constant variances with respect to measurement occasions.

Toeplitz (TOEP). The TOEP structure is akin to the AR structure in that the covariance between any two measurement occasions is constant for all at the same time lag. Therefore, all covariances at time lag 1 are equivalent; all covariances at time lag 2 are equivalent, etc. This manifests itself in the physical appearance of the matrix so that all diagonal elements are the same (see Table 3.2, pp. 95 - 97). In this respect, the TOEP structure shares the AR constraint of only being appropriate when the measurements are collected at equally spaced intervals in time. The TOEP structure is unlike the AR structure, however, in that it is less restrictive and allows the different covariances on each diagonal of the covariance matrix to be freely estimated instead of imposing the constraint that they decrease exponentially. As a result, the TOEP structure may fit data better when the covariances over time approach zero more slowly; however, it requires multiple parameters to be estimated. The number of parameter estimates needed (k) is equal to the number of measurement occasions (t) in the dataset (i.e., $k = t$).

Heterogeneous Toeplitz (TOEPH). Once again, this pattern is a generalization that allows for non-constant variances with respect to measurement occasions.

Unstructured (UN). The UN pattern imposes no constraints on the form of the covariance matrix and, as a result, requires a parameter estimate for each element of the matrix. Therefore, a total of $k = [t(t+1)/2]$ parameters must be estimated requiring the same number of degrees of freedom. This resulting structure is the most flexible, and allows for the best possible fit, but at the expense of expending degrees of freedom and consequently loosing statistical power. This is the covariance pattern assumed by the MANOVA approach to the analysis of repeated measurements.

This has been a brief review of some of the most commonly used covariance structures in longitudinal data analysis. Their inclusion in the current study is a direct result of their prevalence in both applied research contexts and methodological works dealing with the statistical analysis of repeated measurements. All population covariance matrices that were used to generate data in the current study are displayed in tables A2 – A7 (pp. 190 – 195) in the Appendix.

2) Number of Measurement Occasions

The second experimental factor in the current study is the number of measurement occasions with two levels: three and six occasions. These values were chosen to reflect current trends in the analysis of repeated measurements data in the social sciences. In these situations only a few measurement occasions are typically available from multiple subjects. This is a distinct situation from time series data; for example, where only one independent observation is available with a large number of measurement occasions. Furthermore, in eight out of ten Monte Carlo studies reviewed dealing with α_e's and power estimates in longitudinal data, four to eight measurement occasions were selected for investigation by the authors. While these values were certainly in contention for use in the current study, it has been the experience of the author that much applied research in psychology involves repeated measurements with as few as three measurement occasions. For these reasons, the values of three and six were chosen.

3) Magnitude of Serial Correlation

The magnitude of the serial correlation present in the data is the third experimental factor with two levels: .3 and .5. Although, of the ten simulation studies reviewed in this area, only one incorporated any measure of the amount of serial correlation present in the data, this factor is considered especially important in the current investigation. Specifically, it was suspected that certain candidate covariance structures would closely approximate the correct structure when the amount of serial correlation was low. Moreover, while the true amount of serial correlation present in the population will always remain unknown, an applied researcher may estimate this value from their observed data as well as a rough estimate of the structure of the covariance matrix. Therefore, knowing which covariance structures behave alike under these conditions would be of interest. Ferron et al. (2002) generated data with an AR covariance structure while varying ρ, the autoregression parameter[35] between .3 and .6. Values of .3 and .5 were chosen in the current study because of their close proximity to the values used by Ferron et al. Above and beyond this precedent; however, it is well known that low correlations among variables suspected to be related are often encountered in the social sciences when human subjects are involved. Consequently, values of .3, and .5 were considered to realistically and adequately represent this situation.

4) Presence of Non-Constant Variance

The fourth experimental factor is the presence of non-constant variance present in the data with two levels: one and six. These values correspond to the presence of the variance at the final measurement occasion in proportion to the variance at the initial measurement occasion. For example, under the second level of this factor, the multiplier would be six. In this case, the variance at the last measurement occasion is proportional to six times the variance at the first measurement occasion, with the variances of the occasions in between increasing linearly from the value at time one to the last measurement occasion. Thus, a value of one for this factor implies constant variance;

[35] The interpretation of the autoregressive parameter is the amount of correlation present at time lag 1, or, in other words, the amount of correlation present between any two responses collected at adjacent measurement occasions.

however, a value of six implies linearly increasing variances with a maximum variance at the last measurement occasion equal to six times the variance at the first time point. See Table 3.3 (p. 98) for a list of the multipliers used to achieve these increasing patterns of variances over time.

5) Sample Size

The fifth and final experimental factor in the current study is sample size with three levels: 10, 30, & 60. Keselman et al. (1999) and Vallejo and Livacic-Rojas (2005) have found that sample size influences the α_e's of MMs, and it is well known that sample size influences the power of statistical methods. Values less than 60 were chosen here based on a survey of repeated measurement analyses by Kowalchuk et al. (1996). In that survey, the authors found that out of 226 published studies in applied psychology and educational journals, the majority of those studies used 60 subjects or less. At the other end of the spectrum, Gomez et al. (2005) evaluated samples sizes as small as $N = 9$ with $n_i = 3$ for each of three between-subjects groups ($i = 1$ to 3 groups). Additionally, ten was chosen as the smallest value for sample size based on the author's own experience working in a statistical consulting center at a Research I university.

In summary, for the investigation of Type I error rates, the current study generated data with regard to five experimental factors with seven, two, two, two, and three levels, respectively. However, all experimental factors were not completely crossed. For example, generating data for the IN structure precludes that the variance be constant. That is, crossing this true structure with the non-constant variance specification would not be possible because, by definition, the IN structure consists of constant variance. As a result, all experimental factors were crossed except where a contradiction between true structure, magnitude of serial correlation, or nonconstant variance was encountered. This design resulted in a total of 72 experimental conditions under which the data were generated.

Additional Factor for Empirical Power Estimation

For the estimation of empirical statistical power it is necessary to add a sixth factor to the design: Mean Effect. The Mean Effect factor was specified with three levels corresponding to effect sizes measured by ω^2 = .01, .06, .14. These effect sizes relate to small, medium, and large effects as designated by Cohen (1977, pp. 284-288). Because an equally-spaced mean configuration was used in the current study so that the means were increasing linearly with time, mean effect is the only additional factor necessary to fully specify this factor.

The addition of this sixth factor would potentially increase the number of experimental conditions beyond a manageable number. Authors of previous research such as Keselman et al. (1999) and Vallejo and Livacic-Rojas (2005) have encountered similar situations, and chose to investigate empirical power for only those situations where α_e's were robust. Therefore, in order to keep the number of experimental conditions within a manageable range, this approach was adopted in the current study.

Additional Factor for the Comparison of Between-Subjects Groups

As mentioned in Chapter I, a secondary interest of the current study was to assess the Type I error control of the interaction test when a between-subjects factor was present in the design. Therefore, data were generated for a third phase where an additional factor was added to the single-group repeated measures design used in phase I: a between-subjects variable (Group) with two levels (i.e., $j = 1, 2$). For these purposes, the two group data were generated with sample sizes $n_j = 5, 15, \text{or } 30$. Also, group population covariance matrices were specified to be equal. Finally, data were generated where the null hypotheses for all factors (interaction, Group, Time) were true. That is, no interaction effect or main effects of Group or Time were specified.

Other aspects of the experimental design remained constant across phases including: the number of population covariance structures (seven), the number of measurement occasions ($t = 3$ or 6), the magnitude of serial correlation ($r = .30$ or .05), and the presence of non-constant variance (constant or non-constant).

Also, where 10,000 replications were obtained for both phases I and II, only 5,000 replications were performed for this secondary investigation due to time constraints. Data were also generated for only those degrees of freedom options in SAS that proved to be robust in the single-group repeated measures design (phase I; see Chapter IV for these results). Other than these exceptions, the data generation, model fitting, and analysis of these data followed the procedures for phase I.

Data Generation and Model Fitting Procedures Overview: Phase I

Phase I procedures correspond to data generation when the null hypothesis for the test of Time is known to be true in order to obtain α_e's. For this first phase of the study, the SAS macro generated correlated normally-distributed random variables under a given experimental condition (with particular values for the sample size, number of measurement occasions, magnitude of the serial correlation, etc.). At this phase, the data are generated so that the null hypothesis of no Time main effect is known to be true. The macro then fit eight candidate covariance models and the correct model to the same dataset. From these model fits, the five information criteria, likelihood function, convergence status, and information concerning the statistical tests for the main effect of Time were extracted for each fit and saved to a permanent SAS dataset for final analysis (see Appendix, Table A1, p. 186 - 189 for an example of a SAS dataset).

As mentioned earlier in this chapter, a saturated or maximal mean model was specified in all model fits. This is analogous to an ANOVA treatment-effects parameterization of the model (or profile analysis approach) where indicator variables are used to specify separate means at each time point as opposed to modeling Time continuously with a growth curve approach. Once again, this is important to minimize the chance of introducing spurious covariance in the data during the covariance modeling process due to mean model misspecification Fitzmaurice et al. (2004, pp. 173-177).

The diagram provided below displays the general form of the design matrix of these models based on SAS PROC MIXED[36] fitting procedures. In this example, there

[36] The SAS code is provided later in this section.

are ten subjects (i = 1 to N; N = 10) with six measurement occasions (j = 1 to t; t = 6). Therefore, the following model holds for y_{ij}.

$$\mathbf{y} = \mathbf{X} * \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{16} \\ y_{21} \\ y_{22} \\ \cdot \\ \cdot \\ \cdot \\ y_{95} \\ y_{96} \\ y_{101} \\ y_{102} \\ y_{103} \\ y_{104} \\ y_{105} \\ y_{106} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \\ e_{21} \\ e_{22} \\ \cdot \\ \cdot \\ \cdot \\ e_{95} \\ e_{96} \\ e_{101} \\ e_{102} \\ e_{103} \\ e_{104} \\ e_{105} \\ e_{106} \end{bmatrix}$$

A model with a grand mean and six deviation parameters from that grand mean for each of the six measurement occasions is over-parameterized. In order to remedy this situation, SAS automatically sets the estimate of the final measurement occasion to zero. Therefore, the 'intercept term' (β_0) in the model displayed above takes on the value of the final measurement occasion (the 6^{th}, in this example) and $\beta_1 - \beta_5$ are deviations from that mean. This is often called a reference-group parameterization (Fitzmaurice et al., 2004, p. 113). In this example with six measurement occasions, then, the estimate of the population mean at the final or sixth measurement occasion is β_0; however, the estimate of the population mean at the k^{th} measurement occasion for k = 1 to 5 is ($\beta_0 + \beta_k$) (see diagram below).

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 \\ \beta_0 + \beta_2 \\ \beta_0 + \beta_3 \\ \beta_0 + \beta_4 \\ \beta_0 + \beta_5 \\ \beta_0 \end{bmatrix}$$

Notice that this parameterization corresponds directly to $\mathbf{X_i}$, the subset of the design matrix **X** for the i^{th} subject.

$$\mathbf{X}_i = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The first five rows of this matrix determine the first five parameter estimates by adding a value determined by indicators in the second through sixth columns to the β_0 value (the first column). The sixth row is simply set to the β_0 value without any further adjustment and therefore defines the sixth measurement occasion as the reference group.

Data Generation and Model Fitting Procedures Technical Notes: Phase I

For every dataset generated by the simulation, a unique seed value for the SAS random number generator was used. These seed values were generated by the "SEEDGEN" macro developed by Fan, Felsovalyi, Sivo, and Keenan (2001, p. 38).

SEEDGEN is a SAS macro that generates seed values insured to select unique or non-overlapping streams of quasi-random numbers. As a consequence, the seed values used in the current study guarantee the following: 1) the results of the current study can be reproduced at a later time with a subsequent execution of the simulation macro, and 2) no duplicate random numbers are used in data among replications of the simulation.

Technical details concerning the data generation and manipulation and general steps of the simulation are as follows. For each condition *i* and each replication *j*, a unique (N x t) data matrix of standard normal random variables ($\mathbf{D_{ij}}$)[37] was generated using SAS's PROC IML (where N = the sample size of the one homogeneous group which varies across experimental conditions taking on values of 10, 30, & 60, and t = the number of time points or measurement occasions also varying across experimental conditions and taking on values of 3 & 6). Thus, the number of rows in the data matrix ($\mathbf{D_{ij}}$) represent the number of subjects (N) and the number of columns of $\mathbf{D_{ij}}$ represent the number of measurement occasions (t):

$$\mathbf{D_{ij(Nxt)}} \sim \text{N}(\mathbf{0},\mathbf{I}). \tag{3.5}$$

At this point, there is no relationship or correlation among the data with respect to the t measurement occasions. In order to introduce the desired correlation in the data, a correlation matrix ($\mathbf{R}_i$) consistent with the parameters of each experimental condition (i) was specified. That is, the correlation matrix ($\mathbf{R}_i$) was systematically varied between conditions using SAS macro conditional statements. This matrix was dependent on three parameters: the general structure of the correct covariance matrix, the number of

[37] $\mathbf{D_{ij}}$ is the data matrix of standard normal variates for the i^{th} condition and the j^{th} replication.

measurement occasions, and the maximum magnitude of the serial correlation for a given experimental condition (see Methods section Generation of Correlation Matrices for more details).

For example, one condition of the current study entailed generating data for a correct model with an IN covariance structure and three measurement occasions[38]. The $\mathbf{R_1}$ matrix is specified as follows for this condition.

$$\mathbf{R}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

However, for the condition where the correct model is specified to be compound symmetric with six measurement occasions and the magnitude of serial correlation set to .3, $\mathbf{R_2}$ takes the following form.

$$\mathbf{R}_2 = \begin{bmatrix} 1 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 1 \end{bmatrix}$$

Thus, $\mathbf{R_i}$ is systematically varied in order to meet the differing conditions under which the data were generated. Because $\mathbf{R_i}$ entirely specifies the covariance structure of

[38] It should be noted that the independence model is a special case where no serial correlation is present by definition of the structure of the covariance model.

the data to be generated in the i^{th} condition, the subsequent steps for generating the data in the current study can be expressed in generality as a function of $\mathbf{R_i}$.

In order to induce the desired amount and pattern of serial correlation in the raw data matrix, a variation on Ripley's (1987, p. 98) method was used. Ripley's method entails pre-multiplying a matrix of standard normal variates by a lower triangular decomposition (the Cholesky decomposition) of the desired correlation matrix ($\mathbf{R_i}$) for the i^{th} condition. However, the SAS PROC IML function HALF(), which performs the Cholesky factorization, returns the upper triangular decomposition of the $\mathbf{R_i}$ matrix. Consequently, the upper triangular decomposition of $\mathbf{R_i}$ was post-multiplied by the data matrix $\mathbf{D_{ij}}$ (see 3.4):

$$\mathbf{K_{ij(Nxt)}} = \mathbf{D_{ij(Nxt)}} * \text{HALF}(\mathbf{R_{i(txt)}}). \tag{3.6}$$

These operations yield identical results: the resulting matrix ($\mathbf{K_{ij(Nxt)}}$) is a matrix of standard normal variates with the desired correlation structure (Ripley, 1987, pp. 98):

$$\mathbf{K_{ij(Nxt)}} \sim \text{N}(\mathbf{0},\mathbf{\Sigma_i}). \tag{3.7}$$

The correlated data matrix $\mathbf{K_{ij}}$ was then transformed to the familiar *t*-scaling with a mean of 50 and a variance of 100. This was accomplished by post-multiplying $\mathbf{K_{ij}}$ by a diagonal matrix with the standard deviations on the main diagonal, and then adding a comparably-sized matrix of population mean values of 50. Thus, the data generated under the condition of constant variances were post-multiplied by a $\mathbf{V_{i(txt)}}$ matrix =

diag(10, 10, …, 10). However, the data generated under the condition of non-constant variances were post-multiplied by a $\mathbf{V}_{i(txt)}$ matrix = diag(10, 10√2, …, 10√6) so that the variances were linearly increasing to where the variance at the final measurement occasion was proportional to six[39]:

$$\mathbf{L}_{ij(Nxt)} = \mathbf{K}_{ij(Nxt)} * \mathbf{V}_{i(txt)} + \boldsymbol{\mu}_{(Nxt)} \quad (3.8a)$$

where $\mathbf{V}_{i(txt)}$ = diag(10, $10\sqrt{f(\lambda_2)}$, …, $10\sqrt{\lambda_t}$),

$\boldsymbol{\mu}_{(Nxt)} = 50 * \mathbf{J}_{(Nxt)}$, and

$\mathbf{J}_{(Nxt)}$ = a (N x t) matrix of ones.

The resulting data matrix ($\mathbf{L}_{ij(Nxt)}$) represents t measurements on N subjects with a t-scaling and with the desired amount and pattern of serial correlation and variance pattern present. Subsequent steps in the SAS macro simply transposed this matrix to a column vector (i.e., from wide data format to long; necessary for model fitting using SAS's PROC MIXED procedure) and fit differing modern MMs corresponding to the differing covariance structures of interest using PROC MIXED.

An example of the SAS code is given below. The call to PROC MIXED was executed with a SAS macro named TEST. In this example, PROC MIXED is being called to fit the saturated mean model with a different mean estimate at each time point (hence, TIME is specified as a class variable). In addition, the KR approximation is being requested (hence, the DDFM=KR option in the REPEATED statement). Finally, &STRUC in the REPEATED statement is a SAS macro variable that takes on different values corresponding to the different covariance models to be fit.

[39] See Methods section "Explication of Experimental Factors" and Table 3.3, p. 98 for more detail.

The SAS macro processor reads the code between the %MACRO and %MEND statements and then generates SAS code for each %TEST statement read while substituting in the value of the &STRUC macro variable in the code (Delwiche & Slaughter, 2003, pp. 200-201). For example, the first %TEST statement directs the SAS macro processor to write code for PROC MIXED while substituting in "SIMPLE" for &STRUC in the TYPE option in the REPEATED statement and thereby fitting the model with an IN covariance structure.

```
%MACRO TEST(STRUC=);

PROC MIXED DATA=DA4 IC;
     CLASS TIME ID;
     MODEL RESPONSE = TIME / DDFM=KR;
     REPEATED TIME / TYPE=&STRUC SUBJECT=ID;
     ODS OUTPUT MIXED.MODELINFO=INFO;
     ODS OUTPUT Mixed.InfoCrit=IC;
     ODS OUTPUT MIXED.TESTS3=OUTDATA2;
     ODS OUTPUT Mixed.ConvergenceStatus=CON1;
RUN;

%MEND TEST;

%TEST(STRUC=SIMPLE);
%TEST(STRUC=CS);
.
.
.
%TEST(STRUC=UN);
```

Other modern MMs were fit in like manner. For example, the models fit without the KR approximation were done using the similar code, but without the DDFM=KR option. Furthermore, the models with the sandwich estimator were also fit using similar code, however, without the DDFM=KR option and with the EMPIRICAL option in the PROC MIXED statement (Fitzmaurice et al., 2004, p. 184).

As a result, a total of nine modern MMs (with varying covariance models: one correct and eight candidate models) by four test statistic options were fit to each generated dataset. Four classical models were fit to each dataset as well using PROC GLM. The SAS Output Delivery System (ODS) was then used within each call to PROC MIXED (or PROC GLM) in order to save relevant information concerning each model fit to a permanent SAS dataset for subsequent analysis. Information considered relevant here included a variable identifying the covariance structure used in the model fit; degrees of freedom, F-statistic, and associated p-value for the test of Time; the negative two log-likelihood statistic; the number of covariance parameters used to fit the model; values for the five information criteria of interest; and indicator variables and a description of the convergence status of the model. This information was then merged with all information needed to uniquely identify the conditions under which the data were generated. See Appendix, Table A1 (pp. 186 - 189) for an example of this dataset.

Data Generation Overview: Phase II

The second phase involved data generation for the investigation of empirical power estimates for all nine potential candidate covariance models and all four classical models. To this end, data were generated under the condition that the alternative hypothesis for the test of Time was known to be true in order to obtain empirical power estimates. As before, a SAS macro was used to generate normally-distributed random variates under each experimental condition. During this phase, however, each generated dataset had a Time mean effect imposed upon it corresponding to small, medium, and large ω^2 values (Cohen, 1977, pp. 284-288). The mean effects in Time for all conditions have been derived based on data meeting the assumptions of the CLM: independent observations and constant variance across time. These mean effects are displayed in Table 3.4 (p. 99). The SAS macro then proceeded by fitting each of the eight candidate models, the correct model to the data, and the four classical models.

Data Generation Technical Notes: Phase II

As mentioned previously, only those conditions where the test for Time was found to be robust were evaluated during the power phase of the study. Up to the transformation involving the t-scaling (Equation 3.8a), data generation for phase II proceeded exactly as described for phase I. At this stage, however, the phase II procedure deviated from the aforementioned by imposing a mean effect in Time onto the generated data. This was done by including the matrix $\boldsymbol{\alpha}_{\mathbf{i(Nxt)}}$ in the transformation of $\mathbf{K}_{\mathbf{ij(Nxt)}}$:

$$\mathbf{L}_{\mathbf{ij(Nxt)}} = \mathbf{K}_{\mathbf{ij(Nxt)}} * \mathbf{V}_{\mathbf{i(txt)}} + \boldsymbol{\mu}_{\mathbf{i}}{}^{*}{}_{\mathbf{(Nxt)}} \qquad (3.8b)$$

where $\mathbf{V}_{\mathbf{i(txt)}}$ matrix = diag(10, $10\sqrt{f(\lambda_2)}$, …, $10\sqrt{\lambda_t}$),

$\boldsymbol{\mu}_{\mathbf{i}}{}^{*}{}_{\mathbf{(Nxt)}} = 50 * \mathbf{J}_{\mathbf{(Nxt)}} + \boldsymbol{\alpha}_{\mathbf{i(Nxt)}}$,

$\mathbf{J}_{\mathbf{(Nxt)}}$ = a (N x t) matrix of ones, and

$\boldsymbol{\alpha}_{\mathbf{i(Nxt)}}$ = a (N x t) matrix of mean effects for the i^{th} condition (see Table 3.4, p. 99, for actual values).

The matrix $\boldsymbol{\alpha}_{\mathbf{i(Nxt)}}$ introduced the desired amount of difference in the means with respect to Time depending on the specified effect size for a given experimental condition.

See Table 3.4 (p. 99) for the different values of $\boldsymbol{\alpha}_{i(Nxt)}$ corresponding to ω^2 values. These mean effects were chosen to closely approximate the Cohen (1977, pp. 284-288) effect size interpretations of small, medium, and large. As demonstrated in Table 3.4, the actual ω^2 values obtained by these mean configurations do closely approximate the target values of .01, .06, and .14.

An example of the composition of $\boldsymbol{\alpha}_{i(Nxt)}$ for a particular condition of the power analysis appears below. In this example, the condition is characterized by a sample size of ten, three measurement occasions, and a small effect size. The $\boldsymbol{\alpha}_{i(Nxt)}$ matrix takes on the following form.

$$\boldsymbol{\alpha}_{1(10x3)} = \begin{bmatrix} 0.00 & 1.75 & 3.50 \\ 0.00 & 1.75 & 3.50 \\ . & . & . \\ . & . & . \\ . & . & . \\ 0.00 & 1.75 & 3.50 \end{bmatrix}$$

Notice that the matrix has ten rows and three columns corresponding to the size of the $\mathbf{K}_{ij}$ matrix for this condition with a sample size of ten and three measurement occasions. The numerical values for this condition and all other variations are displayed in Table 3.4 (p. 99).

The SAS macro for phase II then executed the same steps as the phase I macro, most notably including those to call PROC MIXED and PROC GLM and fit the appropriate models, extract the pertinent information, and save that information to a permanent SAS dataset for further analysis.

Generation of Correlation and Covariance Matrices

This section is designed to provide further explication concerning the procedure and matrices involved in inducing the desired amount of serial correlation in the data. Five out of the seven covariance structures serving as population covariance patterns from which data were generated are fully specified by the general structure of the covariance matrix (IN, VC, CS, etc.), the number of measurement occasions, and the magnitude of the serial correlation for a given experimental condition. These five structures are the IN, VC, CS, CSH, and ARH models.

The remaining two structures require more information to be fully specified. The following is provided in order to explain how these structures were obtained. First, the TOEPH structure allows for differing rates of decreasing (or increasing) correlation by estimating a unique covariance parameter for each time lag. Consequently, in order to fully specify these structures in the current study, an arbitrary rate of decreasing population correlation was specified by the multipliers {1, 0.65, 0.50, 0.35, 0.20} for the five distinct time lags. For example, at time lag 1, the magnitude of serial correlation for a given condition (i.e., .50) is multiplied by 1 in order to obtain the serial correlation at time lag 1 ($r = .50$). In like manner, in order to obtain the serial correlation present at time lag 2, the magnitude of serial correlation for a given condition (.50) was multiplied by 0.65 to obtain a serial correlation at time lag 2 of $r = .325$. This rule was applied uniformly to all TOEPH matrices with only the first three elements used when the number of measurement occasions was specified as three {1, 0.65, 0.50}. Because both the AR and TOEPH structures model the correlation as a function of time lag, these specific multipliers were chosen to create distinct matrices from the AR structure where the serial correlation is modeled as decreasing exponentially with increasing time lags. These population correlation matrices are displayed in Table 3.5 (p. 100). Furthermore, all population covariance matrices are displayed in tables A2 – A7 (pp. 190 – 195) in the Appendix.

The UN covariance model also requires further information in order to be fully specified. Because this structure more or less implies a random pattern of covariances among the measurement occasions, the SAS uniform random number generator was used to generate random numbers. These numbers were then transformed so that they had a mean of the magnitude of serial correlation specified for a given experimental conditions (either r = .30 or .50) and fell within the interval (r +/- 0.20, where r = .30 or .50). For example, if r = .30, then uniform random numbers were generated with a mean of .30 and a range of (.10, .50). These numbers were then randomly assigned elements with the correlation matrix to be used to induce serial correlation. As a result, there were two unique correlation matrices ($\mathbf{R_i}$) used for the UN matrix: one for each of the two levels of serial correlation. Like Gomez et al. (2005), matrices for three time points were subsets of those created for six time points. For example, the **R** matrix for the UN pattern with a serial correlation of .30 and number of time points equal to three was a sub-matrix of the corresponding **R** matrix for six time points. These population correlation matrices are displayed in Table 3.6 (p. 101). The specific population covariance matrices for all structures used in the generation of the data are provided in tables A2 – A7 (pp. 190 – 195) in the Appendix.

Analysis & Results

As mentioned earlier, a computer macro program was developed using a desktop personal computer with SAS for Windows. However, the final executions of this program were submitted to the AMD Opteron SAS server at the University of Georgia Research Computing Center (RCC) for greater efficiency. The SAS server is composed of two Opteron CPUs operating with 4 Gigabytes of RAM and the Linux operating system. Therefore, the macro was developed with SAS for Windows; however, the final executions of the program were performed by SAS for UNIX.

The SAS datasets generated by the macro were exported to a comma-delimited data file from which they were imported into R for further analysis. R was chosen as a superior environment for subsequent analyses because it facilitates comparisons between cases (or rows) in the data matrix where this is more difficult to accomplish in SAS.

The Evaluation of Empirical Type I Error Rates

As mentioned previously, the 95% confidence interval method for evaluating α_e's was chosen over the Bradley liberal criterion at the onset of the current study. Additionally, it also was decided at the onset to report α_e's at the thousandths level of precision. Thus, for both phases I and II, where the number of replications was 10,000, the 95% confidence interval for evaluating α_e's for a single condition was .050 +/- .004 or (.046, .054). For phase III, where the number of replications was 5,000, the 95% confidence interval for a single condition was .050 +/- .006 or (.044, .056).

However, the majority of results reported in the next chapter are aggregated across conditions in order to arrive at α_e's for seven marginal conditions: low or high correlation, $t = 3$ or 6 measurement occasions, and $N = 10$, 30, or 60 sample sizes. As a result, estimates of α_e's for these marginal conditions are based on varying numbers of replications from 240,000 to 420,000. In addition, estimates for an eighth overall marginal condition were based on 720,000 replications. Because the number of replications differed among these marginal conditions, varying confidence interval widths were necessary in order to evaluate the results at the 95% confidence level. However, because the α_e's were reported to the nearest thousandth, these differing confidence intervals reduced to the same interval after rounding. Therefore, the 95% confidence interval for evaluating α_e's in the majority of tables provided in the next chapter is .050 +/- .001 or (.049, .051)[40].

To further aid in the description and evaluation of α_e's, the current author proposed a criterion of "close proximity." This criterion was used in the next chapter to distinguish between those α_e's that are extreme (i.e., substantially conservative or liberal) and those that, while they do not fall within the limits of the appropriate 95% confidence interval, they still fall within close proximity of the nominal α level. Therefore, α_e's that fall within the intervals (.046, .048) and (.052, .054) were considered only slightly conservative or liberal and therefore within close proximity of the nominal α value.

The Evaluation of Empirical Power Estimates

As mentioned previously, a review of the research literature produced no previous rules or heuristics in place for evaluating the comparability of statistical power. Therefore, the current author proposed a criterion for comparing the empirical power of

[40] Exceptions to this rule are noted when applicable.

two different statistical methods based on the same dataset. Specifically, one method was considered comparable to an arbitrary base method if its empirical power estimate was within .10 of the other methods power estimate.

In the context of identifying surrogate covariance models, this general rule takes on the following form. First, a lower bound for the true model was constructed by subtracting .10 from the true model's empirical power estimate. Then, a given candidate model's power estimate was compared to this lower bound. The candidate model was declared a surrogate if its power estimate was greater than that of the true model's lower bound. Otherwise, the candidate model was not considered to produce sufficiently comparable power in order to be considered a surrogate.

This section has been a brief discussion of the criteria used in the current study to evaluate α_e's and empirical power estimates. When analysis of both of these measures was appropriate[41], evaluation of α_e's was conducted first. Then, empirical power estimates were evaluated for only those methods and only those conditions where acceptable Type I error control was observed.

Limitations

Longitudinal data analysis often constitutes a potentially complex statistical modeling situation. Therefore, it is not reasonable to expect any one simulation study to address all potential variables operative in this context. As a result, the current study was designed to only address a finite number of the relevant issues in order to remain within a manageable scope. Some of the issues that were not addressed include the analysis of

[41] Evaluation for both α_e's and statistical power were not proposed for all investigations in the current study (e.g., phase III).

longitudinal data with unequal measurement occasions, missing values, and categorical responses. Also, the current study focuses on data with one repeated measures factor (Time) and, to a lesser extent, data with one repeated measures factor and one between-subjects factor (Group). However, many permutations of these designs are possible including multiple within- and between-subjects factors. Finally, the current study evaluates continuous response data that are normally-distributed. Investigations of non-normally distributed data may be of interest in the future.

In addition, some methodologists in the field promote a theoretical model for the variance in longitudinal data (Diggle, Heagerty, Liang, & Zeger, 2002; Fitzmaurice et al., 2004, pp. 36-43). This model partitions the variance of the responses over time into three types: serial correlation, between-subject random effects, and measurement error. However, primary interest of the current study was to investigate how well candidate models approximate correct models of serial correlation. Therefore, the data generated in the current study do not exhibit between-subjects random effects. Further research may concern itself with introducing the existence and magnitude of between-subjects random effects and how these factors influence the modeling of serial correlation.

As previously demonstrated, modern MM methods are a flexible class of statistical models. Consequently, there are many different options for modeling the mean response over time such as profile analysis and growth curve analysis approaches. Therefore, another limitation of the current study is that only one approach to modeling the mean response was used: the profile analysis approach. Other options are possible.

Furthermore, care has been taken here to simulate conditions that often occur in the analysis of applied longitudinal data in the social sciences (including small sample sizes, low and moderate serial correlation, etc.); however, every possible variable that may be operative in applied analysis cannot be included. For example, in phase III of the current study, data were generated with variance heterogeneity with respect to Time, however, not with respect to Group. Previous research has shown that group variance heterogeneity, unequal group sample sizes, and positive or negative pairings of these two factors have a substantial effect on α_e's (Gomez et al., 2005; Keselman, Algina et al., 1998; Vallejo & Livacic-Rojas, 2005). This may reduce the external validity of the current study for researchers working with data under these conditions or with highly unusual datasets.

Finally, criteria currently used in order to define acceptable and comparable levels of α_e's and power estimates are not orthodox. Many alternative definitions are possible. Even with these limitations in mind, the current investigation yields important and substantial information regarding the analysis of longitudinal data.

Summary

This chapter outlined the methods employed in the current study to answer the research questions of interest through the evaluation of α_e's and statistical power estimates. The chapter began by identifying the current study as a Monte Carlo simulation that was executed and analyzed using both SAS and R. Next, a brief review of the CLM was provided. Following that, the MM was introduced and issues regarding model parameter estimation, model-based inference, and model fitting procedures were discussed.

Next, the design of the current study was outlined with respect to the three distinct phases under which data were generated. The first phase of the current study was designed to evaluate α_e's. Therefore, data were generated so that the null hypothesis was known to be true. Furthermore, five design features were identified: 1) population covariance structure, 2) number of measurement occasions, 3) magnitude of serial correlation, 4) presence of non-constant variances, and 5) sample size. Finally, the number of replications was set to 10,000 and the 95% confidence interval method was chosen for evaluating α_e's.

The second phase of the current study was designed to evaluate empirical statistical power. Therefore, data were generated in this phase so that the alternative hypothesis was known to be true. For these purposes, the five experimental factors for phase I were retained and an additional factor was added: the magnitude of the mean effect. The Mean Effect factor was designed with three levels corresponding to values of ω^2 delineated as small, medium, and large effects by Cohen (1977, pp. 284-288). Finally, the number of replications was set to 10,000 and a criterion was proposed for evaluating empirical power estimates.

The third and final phase of the study was designed to evaluate the α_e's of the interaction test for data with one repeated measures factor and one between-subjects factor. Therefore, like phase I, data were generated here so that the null hypothesis was known to be true. The design of this phase closely approximates the design of phase I; however, the sample size factor changed slightly in order to reflect the two group case. Therefore, sample sizes in phase III were n_j, = 5, 15, or 30. Finally, the number of replications was set to 5,000 and, like phase I, the 95% confidence interval method was chosen for evaluating α_e's.

This has been a summary of the methods that were employed in the current study. The next chapter reports the results with respect to each of the research questions of interest. General discussion and implications of these results follow in Chapter V.

Table 3.1

Number of Experimental Conditions

Structures	Number of Structures	Number of Measurement Occasions	Magnitude of Serial Correlation	Presence of Non-constant Variance	Sample Size	Number of Conditions
Independence (IN)	1	2	.	1	3	6
Variance Components (VC)	1	2	.	1	3	6
Homogeneous Variance Structures (CS)	1	2	2	1	3	12
Heterogeneous Variance Structures (CSH, ARH, TOEPH, UN)*	4	2	2	1	3	48
Total						72

*CSH=Heterogeneous Compound Symmetry, ARH=Heterogeneous Autoregressive, TOEPH=Heterogeneous Toeplitz, UN=Unstructured

Table 3.2

Structures of Parameterized Covariance Matrices

Independence (IN)	σ^2	0	0	0	0	0
		σ^2	0	0	0	0
			σ^2	0	0	0
				σ^2	0	0
					σ^2	0
						σ^2
Variance Components (VC)	σ^2_1	0	0	0	0	0
		σ^2_2	0	0	0	0
			σ^2_3	0	0	0
				σ^2_4	0	0
					σ^2_5	0
						σ^2_6
Compound Symmetry (CS)	$\sigma^2+\sigma_1$	σ_1	σ_1	σ_1	σ_1	σ_1
		$\sigma^2+\sigma_1$	σ_1	σ_1	σ_1	σ_1
			$\sigma^2+\sigma_1$	σ_1	σ_1	σ_1
				$\sigma^2+\sigma_1$	σ_1	σ_1
					$\sigma^2+\sigma_1$	σ_1
						$\sigma^2+\sigma_1$

Table 3.2 (continued)

Heterogeneous Compound Symmetry (CSH)	σ^2_1	$\sigma_1\sigma_2\rho$	$\sigma_1\sigma_3\rho$	$\sigma_1\sigma_4\rho$	$\sigma_1\sigma_5\rho$	$\sigma_1\sigma_6\rho$
		σ^2_2	$\sigma_2\sigma_3\rho$	$\sigma_2\sigma_4\rho$	$\sigma_2\sigma_5\rho$	$\sigma_2\sigma_6\rho$
			σ^2_3	$\sigma_3\sigma_4\rho$	$\sigma_3\sigma_5\rho$	$\sigma_3\sigma_6\rho$
				σ^2_4	$\sigma_4\sigma_5\rho$	$\sigma_4\sigma_6\rho$
					σ^2_5	$\sigma_5\sigma_6\rho$
						σ^2_6
Autoregressive (AR)	σ^2	ρ	ρ^2	ρ^3	ρ^4	ρ^5
		σ^2	ρ	ρ^2	ρ^3	ρ^4
			σ^2	ρ	ρ^2	ρ^3
				σ^2	ρ	ρ^2
					σ^2	ρ
						σ^2
Heterogeneous Autoregressive (ARH)	σ^2_1	$\sigma_1\sigma_2\rho$	$\sigma_1\sigma_3\rho^2$	$\sigma_1\sigma_4\rho^3$	$\sigma_1\sigma_5\rho^4$	$\sigma_1\sigma_6\rho^5$
		σ^2_2	$\sigma_2\sigma_3\rho$	$\sigma_2\sigma_4\rho^2$	$\sigma_2\sigma_5\rho^3$	$\sigma_2\sigma_6\rho^4$
			σ^2_3	$\sigma_3\sigma_4\rho$	$\sigma_3\sigma_5\rho^2$	$\sigma_3\sigma_6\rho^3$
				σ^2_4	$\sigma_4\sigma_5\rho$	$\sigma_4\sigma_6\rho^2$
					σ^2_5	$\sigma_5\sigma_6\rho$
						σ^2_6

Table 3.2 (continued)

Toeplitz (TOEP)	$\begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 & \sigma_5 \\ & \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 \\ & & \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ & & & \sigma^2 & \sigma_1 & \sigma_2 \\ & & & & \sigma^2 & \sigma_1 \\ & & & & & \sigma^2 \end{bmatrix}$
Heterogeneous Toeplitz (TOEPH)	$\begin{bmatrix} \sigma^2_1 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_2 & \sigma_1\sigma_4\rho_3 & \sigma_1\sigma_5\rho_4 & \sigma_1\sigma_6\rho_5 \\ & \sigma^2_2 & \sigma_2\sigma_3\rho_1 & \sigma_2\sigma_4\rho_2 & \sigma_2\sigma_5\rho_3 & \sigma_2\sigma_6\rho_4 \\ & & \sigma^2_3 & \sigma_3\sigma_4\rho_1 & \sigma_3\sigma_5\rho_2 & \sigma_3\sigma_6\rho_3 \\ & & & \sigma^2_4 & \sigma_4\sigma_5\rho_1 & \sigma_4\sigma_6\rho_2 \\ & & & & \sigma^2_5 & \sigma_5\sigma_6\rho_1 \\ & & & & & \sigma^2_6 \end{bmatrix}$
Unstructured (UN)	$\begin{bmatrix} \sigma^2_1 & \sigma_{21} & \sigma_{31} & \sigma_{41} & \sigma_{51} & \sigma_{61} \\ & \sigma^2_2 & \sigma_{32} & \sigma_{42} & \sigma_{52} & \sigma_{62} \\ & & \sigma^2_3 & \sigma_{43} & \sigma_{53} & \sigma_{63} \\ & & & \sigma^2_4 & \sigma_{54} & \sigma_{64} \\ & & & & \sigma^2_5 & \sigma_{65} \\ & & & & & \sigma^2_6 \end{bmatrix}$

Table 3.3

Variance multipliers for Data Transformation

	Time = 3		Time = 6	
	Variance multiplier at the final measurement occasion			
Time	1	6	1	6
1	10	10	10	10
2	10	$10\sqrt{3}$	10	$10\sqrt{2}$
3	10	$10\sqrt{6}$	10	$10\sqrt{3}$
4			10	$10\sqrt{4}$
5			10	$10\sqrt{5}$
6			10	$10\sqrt{6}$

* Each column represents a vector of multipliers which constitutes the main diagonal of the $\mathbf{V}_{\mathbf{i(txt)}}$ matrix that is to be post-multiplied to $\mathbf{K}_{\mathbf{ij(Nxt)}}$ in (3.8a & 3.8b) in order to obtain the desired amount of variance at each measurement occasion.

Table 3.4

Mean Configuration Values for Power Analysis

	Time = 3			Time = 6		
	Cohen's (1977) interpretation of effect size					
Time	Small	Medium	Large	Small	Medium	Large
1	0.00	0.00	0.00	0.00	0.00	0.00
2	1.75	2.95	4.95	0.60	1.50	2.35
3	3.50	5.90	9.90	1.20	3.00	4.70
4				1.80	4.50	7.05
5				2.40	6.00	9.40
6				3.00	7.50	11.75
Target ω^2 value	.0100	.0600	.1400	.0100	.0600	.1400
Actual ω^2 value	.0200	.0548	.1404	.0104	.0616	.1387

* Each column represents the variations of the row composition of $\boldsymbol{\alpha}_{\mathbf{i(Nxt)}}$ for conditions differing in number of measurement occasions and effect size.

Table 3.5

Population Correlation Matrices for the Toeplitz Pattern

Correlation	3 measurement occasions			6 measurement occasions					
	1	.300	.195	1	.300	.195	.150	.105	.060
		1	.300		1	.300	.195	.150	.105
.3			1			1	.300	.195	.150
							1	.300	.195
								1	.300
									1
	1	.500	.325	1	.500	.325	.250	.175	.100
		1	.500		1	.500	.325	.250	.175
.5			1			1	.500	.325	.250
							1	.500	.325
								1	.500
									1

Table 3.6

Population Correlation Matrices for the Unstructured Pattern

Correlation	3 measurement occasions			6 measurement occasions					
.3	1	.272	.434	1	.272	.434	.427	.285	.220
		1	.147		1	.147	.177	.179	.205
			1			1	.394	.455	.395
							1	.285	.129
								1	.286
									1
.5	1	.609	.572	1	.609	.572	.657	.473	.449
		1	.370		1	.370	.581	.395	.428
			1			1	.510	.403	.519
							1	.455	.382
								1	.495
									1

CHAPTER IV: RESULTS

This chapter presents the following results that correspond directly to the research questions of interest in the current study: i) empirical Type I error rates (α_e's) for four modern mixed model (MM) test statistic options; 1) α_e's and power estimates of all true model by potential surrogate model combinations as well as identification of surrogate models; 2) selection rates of the a) correct, b) surrogate, and c) appropriate models as well as the rates of a) underfitting, and b) overfitting the true model for five information criteria; 3) α_e's and power estimates for models selected by each of the five information criteria; 4) α_e's and power estimates for both modern MMs and classical methods; 5) α_e's for the interaction test of modern MMs when a between-subjects factor is added to the single-group repeated measures design.

As described in Chapter III, raw data were generated, the appropriate statistical models fit, and model fit information was coerced and saved in permanent SAS datasets. Once these datasets were obtained, they were imported into R and the analysis phase of the study began. R programs that analyzed the data were run on both PC and IBM pcluster machines.

This chapter uses a set of tables to summarize the results of the current study. These tables are located both at the end of the current chapter as well as in the Appendix. An attempt was made to construct these tables so that they are as uniform as possible for ease of reading and interpreting the results. However, at least two comments are worthy of note.

First, most tables reporting α_e's and power estimates summarize these results with respect to eight marginal conditions: low correlation, high correlation, 3 measurement occasions, 6 measurement occasions, small sample size, moderate sample size, and large sample size, as well as the overall marginal summary[42]. Tables of this variety for power estimates report results aggregated across the Mean Effect factor. In addition, tables that summarize results with respect to the twelve combinations of correlation, measurement occasions, and sample size are provided when appropriate (usually when an interaction was observed and further explanation is offered).

Second, the high correlation condition for both the independence (IN) and the variance components (VC) true structures is non-applicable because, by definition, no correlation is present in these structures. For ease of reading the tables, however, the IN and VC data are summarized under the low correlation condition. In actuality, there is no correlation present in the IN and VC population covariance matrices (see Table 3.1, p. 94). This affects the format of the tables because, as a result, the number of replications may vary among columns. For example, in Table 4.i.1 (p. 133), column 1 estimates are based on $N \approx$ 420K replications, while column 2 estimates are based on $N \approx$ 300K replications.

Therefore, in tables reporting α_e's, the number of replications used in the calculation of confidence intervals varied among columns. However, α_e point estimates are reported to the nearest thousandth[43]. As a consequence of this level of precision, the limits of respective 95% confidence intervals were also rounded to the nearest thousandth. As a result, the 95% confidence interval of (.049, .051) was used in all instances, except where otherwise noted. Point estimates of α_e's that fell within the limits of this interval are presented in bold print in all tables. The criterion of close proximity proposed in Chapter III was also used in order to evaluate results

[42] The overall marginal summary aggregates data across all twelve conditions.
[43] In contrast, empirical power estimates are reported to the nearest hundredths.

summarized in these tables: (.046, .048) & (.052, .054). However, no special formatting was used in the summary tables in order to identify values that met this criterion.

Before addressing the research questions outlined in the first several chapters of this document, an analysis of nonconvergence rates of the modern MM methods is provided.

Nonconvergence Rates

Mixed models fit in SAS's PROC MIXED are estimated using restricted maximum likelihood (REML) methods. REML is an iterative estimation procedure that requires a convergence criterion to be met in order to produce a final set of model parameter estimates. As a result of this estimation method, it is possible that some models will not meet the convergence criteria, and therefore will not arrive at final model parameter estimates. Model fitting attempts that do not meet the convergence criterion may provide estimates that are unstable. Results obtained under these conditions are generally considered unreliable. Nonconvergence usually occurs when the data do not conform to the model to be fit.

Convergence-related issues are summarized in Table A8 (see Appendix, p. 196). In order to facilitate ease in reading this table, nonconvergence rates are reported as opposed to convergence rates. Table A8 reports frequencies of models that did not converge for each of the true covariance structures under which the data were generated (aggregated across the nine fitted models). Empty cells indicate that all models obtained convergence for a given condition.

Inspection of Table A8 (p. 196) demonstrates that nonconvergence was not prevalent enough to be problematic in the current study. For all but one true structure, counts of nonconvergence ranged from 11 to 14 out of either 540,000 or 1,080,000. The exception to these figures is the VC true structure where 161 models did not converge. Further investigation demonstrated the existence of a 3-way interaction: true model by fitted model by sample size. That is, all 161 incidents of nonconvergence occurred when the heterogeneous compound symmetry (CSH) model was fit to the VC data (true model) and the sample size was small. In this instance, the rate of nonconvergence was 0.81% (161 models out of 20,000). Furthermore, approximately only three tenths of a percent of the CSH models fit to the VC data did not converge (regardless of sample size). While these rates are elevated with respect to nonconvergence rates in all of the other conditions, they are still a very small percent of the total number of models attempted to be fit under these conditions. Therefore, this problem is unlikely for the typical applied researcher who may be using these methods under these conditions.

In conclusion, Table A8 (p. 196) demonstrates that the overwhelming majority of MMs fit to the data generated in the current study successfully obtained convergence status as evaluated by the PROC MIXED default settings. Therefore, non-convergence related issues were not considered to be a threat to validity of the results of the current study. Nonetheless, all subsequent results were calculated conditionally on model convergence. That is, only those models that met the convergence criteria were used in the calculations[44] to answer the research questions of the current study. In light of these findings, the remainder of this chapter reports and discusses the results that specifically address each of the research questions initially presented in Chapters I and II.

[44] That is, calculations of α_e's, power estimates, selection rates, etc.

Preliminary Research Question i

i) *How do test statistics for the fixed effects of the mixed model compare with respect to* α_e's *when the SAS PROC MIXED default (the Between/Within option), the Satterthwaite approximation, the Kenward-Roger approximation, and the sandwich estimator options are used?*

The main objective of this question was to investigate the relative robustness of the modern MM test statistics when differing methods of obtaining those statistics are used. This was accomplished by investigating α_e's under the differing conditions implemented in the current study.

Tables 4.i.1 (p. 133) and 4.i.2 (p. 134) report α_e's with respect to the four test statistic options under each of the eight marginal conditions. In Table 4.i.1, only the correct model fit was used from the seven true structures from which the data were generated. These seven α_e's were then averaged across true structure in order to obtain an aggregated α_e point estimate for each test statistic option for a given marginal condition. In contrast, Table 4.i.2 aggregates the α_e's across the eight incorrect candidate models for each of the seven true models and then across all seven true models in order to obtain estimates for each of the four test statistic options.

Inspection of Table 4.i.1 (p. 133) shows that the Kenward-Roger (KR) approximation obtained α_e's within the limits of the 95% confidence interval for six out of the eight marginal conditions. The two marginal conditions where point estimates did not fall within these limits were the high correlation and the small sample size marginal conditions. However, the KR test statistics obtained values of .048 in both of these cases. Therefore, these estimates were not liberal, only slightly conservative, and still within close proximity of the nominal α value of .05. In fact, when comparing across test statistic options, the KR value of .048 for the small sample marginal average is more desirable than the estimates from the other test statistic options, which were liberal in all cases, and were excessively liberal for both the Between/Within and sandwich estimator options. Finally, Table 4.i.1 shows that the KR approximation was the only option that obtained an overall marginal mean that fell within the 95% confidence interval limits.

Further investigation demonstrated that α_e's for the KR approximation were within the 95% confidence interval limits in six out of the twelve conditions studied (see Appendix, Table A9, p. 197). Of the other test statistic options, only two point estimates fell within the confidence interval limits for the Satterthwaite approximation, one for the Between/Within option, and none for the sandwich estimator. Finally, all twelve of the KR point estimates fell within close proximity of the nominal value[45].

Further analysis under all twelve conditions also revealed that the Between/Within option did not perform well, with the majority of α_e's in the neighborhood of .063 with a range of .053 to .115. Furthermore, this option proved especially problematic in instances where both a large number of measurement occasions (6) were used and the sample size was small. Under these conditions, the Between/Within option obtained an α_e estimate of .115. The Satterthwaite approximation obtained elevated α_e's under the same conditions. However, the Satterthwaite approximation provided more control over these errors (.088 was the largest estimate).

As expected from the available literature (Fitzmaurice et al., 2002, p. 303-305), the sandwich estimator did not perform well under small sample conditions and obtained values ranging from .112 to .399. Similar to the Between/Within and Satterthwaite results, there does appear to be a two-way interaction operative between the number of measurement occasions and

[45] That is, all estimates met the close proximity criterion used in this study (0.046, 0.048) & (0.052, 0.054).

sample size. This interaction manifests itself through excessive Type I errors for the sandwich estimator under small sample conditions with a large number of measurement occasions. Specifically, the difference between α_e's at measurement occasions $t = 3$ & 6 at the $N = 10$ factor level of sample size (.290) is 7.18 times *greater than* the average difference between $t = 3$ & 6 for $N = 30$ & 60 levels (.040). However, because $N = 10$ is an extremely low sample size condition and the existing literature has already documented the poor performance of the sandwich estimator under these conditions (Fitzmaurice et al., 2002, p. 303-305), it may be of more interest to note the effect of this interaction on moderate sample sizes. Results show the sandwich estimator obtained inflated α_e's of the magnitude of .124-.125 under conditions with as many as 30 subjects (considered a moderate sample size in many social science statistical applications) when a large number of measurement occasions are analyzed.

Because the sandwich estimator is resistant to covariance model misspecification, it may be inappropriate to compare it with the other options when only the correct covariance model was fit to the data as is the case in Table 4.i.1 (p.130). Therefore, a separate table (4.i.2, p. 134) was constructed that aggregates the α_e's across the eight incorrect candidate models for each of the seven true models and then across all seven true models in order to obtain estimates for each of the four test statistic options. These estimates provide information to the approximate α_e's across a number of modeling circumstances including when the covariance model was both underfit and overfit. Table 4.i.2 demonstrates that while none of the KR α_e's fell within the 95% confidence interval limits, none of them exceed the upper limit and the majority met the criterion of close proximity. In contrast, the sandwich estimator continued to obtain excessively inflated estimates for small sample cases and substantially inflated estimates for $N = 30$ moderate sample cases when the number of measurement occasions was large ($t = 6$). Furthermore, both the

Between/Within and Satterthwaite methods do not appear to be in contention as they continue to obtain inflated α_e's for small sample/large measurement occasion conditions. Therefore, the KR approximation still appears to provide superior error rate control, even when the covariance model is misspecified.

Overall, these results demonstrate that the KR approximation provides superior Type I error control when compared to the other options investigated in the current study. Even when the KR α_e's did not fall within the 95% confidence interval limits, they never exceeded the upper limit of that interval and therefore never obtained results suggesting a liberal test. Furthermore, the majority of α_e's fell within close proximity to the nominal α value of .05, providing further evidence of superior Type I error rate control. Because no other option came close to the KR level of control, a comparison of empirical power estimates among these test statistic options was considered unnecessary. As a result of these findings, the KR approximation is used in all modern MMs in subsequent analyses.

Primary Research Question 1

1) *Do surrogate covariance structures exist? If so, which structures serve as acceptable approximations for a given population or correct structure?*

The objective of this research question was to assess the comparability of different covariance structures in the modeling of longitudinal data using modern MM methods. This was considered important for three main reasons. *First*, the existence of surrogate covariance structures was hypothesized has having an impact on the accuracy rates of information criteria within the context of modeling covariance patterns. *Second*, an investigation into surrogate structures was expected to yield information regarding the severity of covariance model misspecification in modern MM use. *Third*, results from this investigation were expected to yield helpful heuristics for applied researchers employing modern MM methods in similar contexts.

As mentioned previously, surrogate-hood was based on two criteria: 1) Type I error control as measured by the 95% confidence interval method with $\alpha_n = .05$, and 2) a potential candidate model's comparability to the correct model in terms of empirical power estimates. Once again, a candidate model was considered comparable in statistical power if its empirical power estimate was greater than the lower bound (LB) formed by the correct model's power estimate:

$$LB = (1 - \beta)_{\text{correct model}} - .10.$$

Table 4.1.1 (p. 135) reports overall marginal α_e's and power estimates[46] for each of the 63 true model by candidate model combinations. Table 4.1.2 (p. 136) reports surrogate status of the candidate models with respect to each of the true models. Table 4.1.3 (p. 137) displays values of the multivariate index of distance, *u*. Tables 4.1.4 (p. 138) and 4.1.5 (p. 139) report α_e's and power estimates, respectively, that have been aggregated across the seven true structures in order to provide further insight into how well each candidate model performs when applied to longitudinal data on average. If further inspection of α_e's and power estimates is desired, fourteen tables are provided in the Appendix that report this information in disaggregated form. (see Appendix, tables A10 – A23, pp. 198 - 211).

[46] These α_e's and power estimates are averaged over all twelve conditions delineated by crossing the correlation, measurement occasion, and sample size factors.

The format of tables 4.1.1 (p. 135) and 4.1.2 (p. 136) are very similar. Both tables list the seven true covariance structures by row and the nine candidate models by column. The main diagonal[47] represents the situation where the correct model was fit to the data. Empirical Type I error rates (α_e's) are presented in each cell with power estimates immediately below the α_e's and in parentheses. For both tables, cells below the intersection (the lower triangle) represent situations were the data were underfit. Cells in the upper triangle represent situations where the data were overfit. In Table 4.1.1, both α_e's and power estimates that met their respective criteria are printed in bold. When both of these criteria were met, the candidate model was identified as a surrogate for that particular true model. These cells are shaded in Table 4.1.1 for improved interpretability.

Table 4.1.1 (p. 135) demonstrates that only in five of the seven instances when the *correct* model was fit to the data was the α_e criterion for surrogate-hood met. This phenomenon may be a reflection on the level of strictness of the α_e criterion or the level of precision (i.e., the number of replications) used in the current study. Or, this may be an indication of the modern MM methods' inability to properly model data under these conditions. This issue is discussed further in the next chapter.

Further inspection of Table 4.1.1 (p. 135) shows that 14 surrogates were identified in all. *First*, 13 of the 14 surrogates identified overfit the data; that is, these were situations where the modeled covariance structure overfit the true covariance structure. The exception was where the CSH model was found to be a surrogate for the UN true model. *Second*, models that overfit the variance structure typically demonstrated acceptable levels of Type I error control and comparable power (no substantial loss of power). In contrast, candidate models that underfit the

[47] Table 4.1.1 is not a square matrix and, therefore, technically does not have a main diagonal. However, the cells appearing on the diagonal running from the top left to the bottom right do have a special interpretation.

variance structure were typically rejected as surrogates due to inflated α_e's. *Third*, the number of surrogates for each true structured differed. Three surrogates for the IN model were identified: CS, AR, and TOEP. For the VC model, four surrogates were identified: CSH, ARH, TOEPH, and UN. Next, the TOEP and UN structures were identified as surrogates for the CS model. The TOEPH and UN structure were identified as surrogates for the ARH true model. In contrast, only the UN structure was identified as a surrogate for the TOEPH true model. *Finally*, as previously mentioned, the CSH model was identified as a surrogate for the UN structure. Implications of these results are addressed in the next chapter.

Table 4.1.2 (p. 136) identifies surrogate relationships in a more readable format. Candidate models that met the requirements for surrogate-hood for a given true model are identified as such by a "1" in the appropriate cell of the table. All other cells are left blank.

To further investigate the comparability of candidate models with true population structures, a measure of the multivariate distance between the mixed model-estimated and sample-based covariance matrices was obtained using the index u (Rencher, 2002, p. 248-249):

$$u = v[\ln|\mathbf{S}| - \ln|\mathbf{M}| + \mathrm{tr}(\mathbf{M}\mathbf{S}^{-1}) - p] \qquad (4.1)$$

where v = degrees of freedom of $\mathbf{M}$,
$\mathbf{S}$ = sample covariance matrix,
$\mathbf{M}$ = model estimated covariance matrix, and
p = the number of variables = the number of time points.

Therefore, u was considered a measure of how well the MM estimated or reproduced the sample covariance matrix. Specifically, smaller values of u indicate that the estimated covariance matrix more closely approximated the sample matrix while larger values indicate more "distance" between the two matrices, and therefore greater lack of fit.

Mean values of *u* are presented in Table 4.1.3 (p. 137) for each of the 63 true model-by-candidate model combinations[48]. Combinations where a candidate model was found to be a surrogate for a given true model are identified with corresponding shaded cells. Inspection of this table demonstrates that the behavior of *u* met a priori expectations. For example, the table shows that when the IN model was fit to simpler data (IN & VC), mean values of *u* were relatively low. However, when this highly restrictive covariance structure was fit to more complex structures, it was unable to adequately approximate the sample covariance matrix as indicated by higher mean values of *u*. Specifically, mean values of 4.27 and 6.89 were obtained for *u* when the IN model was fit to the IN and VC data, respectively. In contrast, values of 9.73 and 11.62 were obtained when the IN model was fit to the TOEPH and UN data.

It is worth noting that the UN structure obtained the lowest values of *u*, even when compared to situations when the correct covariance model was fit to the data. Under these circumstances (i.e., when the correct model was fit to the data), values of *u* ranged from 2.92 to 4.27. However, when the UN structure was fit to these same structures, lower values of *u* were obtained that ranged from 1.91 to 1.92. These results demonstrate the UN structure's ability to approximate other covariance structures and further support its role as a universal surrogate for the true structures evaluated in the current study.

Furthermore, lower values of *u* were found to correspond with surrogate-hood. This is the general trend, but not the case in every instance. For example, when the TOEPH candidate model was fit to data generated from each of the seven true models, mean values of *u* were tightly clustered and obtained a range from 2.90 to 3.33. In fact, TOEPH obtained the same value of 2.91 when fit to both the VC and CS data. However, the TOEPH model was only found to be a surrogate for the VC true structure. In general, however, values of *u* support the conclusions

[48] That is, 63 sample-by-estimated matrix combinations.

concerning surrogate-hood drawn from the comparison of α_e's and power estimates. That is, lower values of *u* were observed for combinations where the candidate model was found to be a surrogate for a given true model.

Tables 4.1.4 (p. 138) and 4.1.5 (p. 139) provide some insight as to the general applicability of the nine candidate models averaged across true model with respect to the eight marginal conditions. Table 4.1.4 demonstrates that, on average, α_e's were controlled when the CSH model was used in moderate and large sample size situations, regardless of the true model. Additionally, power estimates for this model under these conditions demonstrate a reasonable degree of statistical power ranging from .40 to .94 (see Table 4.1.5). Tables 4.1.4 and 4.1.5 also show that the AR model controlled α_e's in five of the eight marginal conditions; however, power estimates were considerably lower than the CSH model fits with differences as great as 12%. Finally, the UN structure provided superior Type I error control, obtaining α_e's within the 95% confidence limits in all eight marginal conditions. Furthermore, the UN structure demonstrated comparable power to the CSH model with estimates only 1 to 3% lower than the CSH model.

These results identified 14 surrogate covariance models. Moreover, results suggest that both the CSH and UN structures may serve as acceptable approximates to any one of the seven true models with both of these structures providing reasonable Type I error control and comparable power estimates. These results are discussed further in the next chapter.

Primary Research Question 2

> *2) What are the selection rates of a particular information criterion with respect to selecting a) the correct model, b) a surrogate model, c) and an appropriate*

model? What are the selection rates with respect to a) underfitting or b) overfitting the data?

The objective of this question was to provide information concerning the accuracy of information criteria in selecting *appropriate* covariance models for modern MM methods in the analysis of longitudinal data. For purposes in the current study, appropriate covariance models were defined as both the correct model and any surrogate models that had been identified. This objective was achieved by first identifying surrogate covariance structures (see previous section) and then estimating the accuracy of these criteria by accounting for the frequency with which they selected the correct model or a surrogate model.

Table 4.2.1 (p. 140) displays the selection rates of the correct model, all surrogate models, and appropriate models (a total of both correct and surrogates models) for each of the seven true covariance structures for the five information criteria under investigation in the current study. Marginal rates of selecting the correct model averaged across the seven true structures ranged from .51 to .54 for all information criteria. However, these selection rates were higher for simpler models (IN, VC & CS), with a mean of .70 and a range of .62 to .91; less so for moderately complex models (CSH & ARH), with a mean of .64 and a range of .64 to .75; and, as expected, even lower for more complex models (TOEPH & UN), with a mean of .13 and a range of .01 to .34.

Further inspection of Table 4.2.1 (p. 140) demonstrates that across the five ICs and the seven true models, selection rates for surrogates ranged from .11 (CAIC) to .19 (HQIC) with a mean of approximately .16. Trends for surrogate selection were similar to those observed for correct models: rates were higher for the simpler models (IN & VC), ranging from .06 to .31[49]; and less so for more complex models (CS, CSH, ARH, & TOEPH), ranging from .01 to .11. The exception to this rule was the UN structure: a complex structure that obtained relatively high rates of surrogate selection across the five information criteria: mean of .52 and a range from .39 to .62. This is no doubt explained by the fact that the one surrogate for UN (CSH) underfit the data and all five information criteria were observed to select simpler models than more complex ones, on average.

Finally, Table 4.2.1 (p. 140) also displays accuracy rates for selecting *appropriate* covariance models. Over all conditions studied, these results suggestion that one can expect to select an appropriate covariance model .68 of the time across the five information criteria, approximately .16 higher than the percent of selecting the correct model alone. The rates of selecting an appropriate model ranged from .64 to .71 across the five information criteria. As expected, the introduction of surrogate model selection into these accuracy rates exhibited the most impact on accuracy rates with respect to simpler models (especially IN & VC) where a larger number of surrogate models were identified. The impact was also high for the UN structure where a surrogate with a less complex structure was identified.

Similar to Table 4.2.1 (p. 140), Table 4.2.2 (p. 141) displays the selection rates with respect to underfitting and overfitting. A total incorrect selection rate (the sum of the underfitting and overfitting rates) is also included in this table. Marginal rates of underfitting the covariance structure ranged from .24 (HQIC) to .37 (CAIC) across the seven true models. The mean rate of underfitting across the five criteria was .29. In contrast, marginal rates of overfitting ranged from .05 to .11 with a mean of .08.

---

[49] Inspection of Table 4.1.1 shows that the simpler models obtained more surrogates and; therefore, information criteria had a greater chance of selecting surrogates for these true models.

In summary, selection rates for the correct, surrogate, appropriate models as well as rates of underfitting and overfitting the covariance structure do not provide evidence that one IC is superior to others in this context. Specifically, no substantial differences among ICs were observed in regard to these rates. Further discussion of these results and their impact on the use of information criteria is provided in the next chapter.

Primary Research Question 3

> *3) Will the analysis be statistically valid if one uses a particular information criterion to select a covariance model? That is, under what conditions are the* α_e's *controlled for models selected by a given information criterion?*

The objective of this question was to assess how reliable the five information criteria are in selecting modern MMs that produce robust test statistics under certain conditions. This question was considered especially important for two reasons. *First*, in applied research settings, researchers never know the true covariance structure of the data. Therefore, it was considered important to provide information concerning the robustness of modern MM test statistics when information regarding the true model was not available. *Second*, the articles reviewed in Chapter II that addressed this issue did so only with degrees of freedom methods other than the KR approximation and only with respect to the interaction test, not the test for the main effect of Time (Keselman et al., 1999; Robertson, 1996).

In order to answer this question, α_e's were obtained for the average selection of each of the five information criteria. These values were computed by transposing the column vector of selection rates of the nine candidate models for a given true model and multiplying by the vector of α_e's for that true model for each information criteria and each particular condition. This was repeated seven times for each of the true models and then the mean of these values was computed for each of the eight marginal conditions. These results are reported in Table 4.3.1 (p. 142). This table also presents α_e's for each of the four classical method tests to facilitate comparison among information criteria selected modern MMs and the classical methods.

Table 4.3.1 (p. 142) displays these results by each of the five information criteria and classical methods (rows) and each of the eight marginal conditions (columns). Table 4.3.1 reveals that only three of the information criteria obtained α_e's that fell within the 95% confidence interval limits: AIC, AICC, and HQIC. Moreover, these instances only occurred in large sample cases. However, closer inspection shows that none of the models chosen by the five information criteria obtained inflated α_e's and the most conservative estimate was only slightly so with a value of .045. That is, on average, all five information criteria selected models that produced α_e's that fell within close proximity of the nominal α value the large majority of the time. Under none of the conditions studied were the α_e's greater than $\alpha_n = .05$. Results for the classical methods demonstrate that the conventional *F*-test was liberal and the G-G corrected test was conservative under these circumstances. However, both the H-F corrected test and the MANOVA test provided exceptional Type I error control under these conditions with all α_e's falling within their respective 95% confidence interval limits.

Further, Table 4.3.2 (p. 143) displays empirical power estimates by information criteria. This table was constructed in a similar manner as Table 4.3.1 (p. 142); however, power estimates were averaged across the three levels of the Mean Effect factor. Inspection of Table 4.3.2 demonstrates that very little variation among the statistical power of models selected by the five information criteria was observed. Differences under any given condition were not greater than .01. Therefore, the statistical power of selected models does not appear to be an issue when choosing which information criterion to use. Results for the G-G showed that differences in power between it and the modern MMs as well the H-F and MANOVA approaches were negligible even though it was found to be conservative with respect to α_e's. Results further demonstrated that both the H-F and MANOVA tests provide power comparable to the modern MM approach under these circumstances.

In order to further investigate α_e's, two additional tables were constructed for the two most common information criteria: AIC and BIC. Tables 4.3.3 (p. 144) and 4.3.4 (p. 145) report α_e's for models selected by AIC and BIC, respectively. These tables show a disaggregation of the information summarized in Table 4.3.1 (p. 142) by reporting the observed α_e's for each true structure. This format is important for two reasons. *First*, it demonstrates that these estimates are stable across true covariance structures of differing degrees of complexity. *Second*, due to this stability, it appears reasonable to aggregate these data across true structure, as was done in Table 4.3.1.

Indeed, both tables 4.3.3 (p. 144) and 4.3.4 (p. 145) demonstrate that α_e's are remarkably stable across true covariance structures of differing complexity with the majority of estimates falling within close proximity of the nominal α value of .05. Furthermore, these tables provide additional evidence to conclude that the information criteria tend not to select models that produce inflated α_e's. For example, Table 4.3.3 shows that the most excessive α_e value for models selected by AIC was .054. This value was only observed once (the VC true model and three measurement occasions condition) and still fell within the criterion of close proximity used in the current study. Similar trends are evident in Table 4.3.4 for BIC.

In conclusion, results from the current study demonstrate that, on average, when information criteria are used to select modern MMs, the models chosen will produce α_e's that closely approximate the nominal α value. Further, there does not appear to be any substantial difference in the statistical power of the models selected by these information criteria. These results will be discussed further in the next chapter.

Secondary Research Question 4

> *4) How does the modern MM approach compare to the classical methods of repeated measures analysis in the context of covariance model misspecification? More specifically, how does the modern MM F-statistic compare to the classical methods: RM ANOVA conventional F-statistic, the Greenhouse-Geisser (G-G) or Huynh-Feldt (H-F) corrections, or the MANOVA Wilks' Λ test statistic with respect to α_e's and empirical power estimates?*

The objective of this question was to compare the performance of the modern MM methods with the classical methods in the analysis of longitudinal data. The KR test statistic for the modern MM methods was compared to the RM ANOVA conventional *F*-test, the G-G and H-F corrected univariate statistics, and the MANOVA Wilks' Λ statistic in terms of α_e's and power estimates.

As discussed in earlier chapters, longitudinal data often exist with covariance structures that are not spherical in nature. However, the RM ANOVA conventional *F*-test assumes sphericity. Furthermore, both the RM ANOVA G-G and RM ANOVA H-F adjustments are attempts to correct for the degree of non-sphericity in a given set of data. Therefore, it may be problematic to use these classical methods depending on the degree of non-sphericity of the data and/or the performance of the G-G and H-F adjustments. The modern MM methods have been suggested as a practical modeling alternative to these classical methods because they do not require sphericity and therefore may be more powerful. For these reasons, the extent of non-sphericity in the data is a critical factor in the comparison of these methods. Consequently, it was considered important to assess the degree of non-sphericity in the data generated in the current study.

Assessing Non-sphericity

Initially, two measures of non-sphericity were collected during the data generation and modeling phases of the study: the G-G and H-F ε statistics. For both of these, values in the proximity of 1.0 indicate that the data are spherical in nature. Lower values indicate greater departure from sphericity.

According to Maxwell and Delaney (2000, pp. 476-477), the H-F ε statistic tends to underestimate the degree of non-sphericity, while the G-G statistic slightly over-estimates it. Because the objective here was to demonstrate the degree of non-sphericity in the data, the H-F ε statistic was chosen for further analysis. These values provide an upper limit for the degree to which sphericity was violated in the data generated in the current study.

The H-F ε statistic was aggregated and inspected across the seven true covariance structures and the eight marginal conditions. The H-F ε overall marginal means for the two true covariance models where the assumption of sphericity was met (those being the IN and CS models) were both 1.013. As expected, these values indicate that the sample data are nearly spherical under these conditions.

However, H-F ε statistic means for more complex structures failed to demonstrate an extreme departure from sphericity. While mean estimates did fall to values as low as 0.795, this value does not suggest a great degree of non-sphericity. Furthermore, the range of values for the non-spherical data was not much different from those of the spherical data. For example, the UN true covariance model data have the greatest potential to be non-spherical. However, the overall marginal range of those H-F ε values was (0.32, 1.69). This interval is not substantially different from the range for the IN model where sphericity was met (0.43, 1.89). This trend proved to be the rule based on further inspection.

Furthermore, histograms of the H-F ε values were obtained and inspected. These graphics showed substantial differences in the distributional shapes of the H-F ε values among the levels of sample size[50]. However, distributions were relatively homogeneous across the non-spherical true covariance structures (i.e., VC, CSH, ARH, TOEPH, UN). Evidently, sampling error seems to influence these values to a greater degree than expected and the true covariance model to a lesser degree.

To further investigate this issue, ε, the parameter that the H-F ε statistic estimates, was obtained for each of the population covariance matrices used in the current study (Huynh & Feldt, 1976). Values of ε appear in Table 4.4.1 (p. 146) as well as tables A2 – A7 (see Appendix, pp. 190-195). The ε parameter was considered an indicator of non-sphericity of the population covariance matrices that is independent of the effect of sampling error. Table 4.4.1 shows that ε values ranged from .693 to 1.000. The population covariance matrix demonstrating the most non-

[50] Distributions were relatively homogeneous across two other design features: magnitude of correlation and, surprisingly, number of measurement occasions.

sphericity was the ARH structure with six measurement occasions and a high correlation (r = .50) with ε = .693. As expected, the IN and CS structures obtained ε values of 1.00 regardless of the number of measurement occasions and the magnitude of correlation present. That is, these population covariance matrices were perfectly spherical.

Based on these findings, another phase of the current study was added in order to allow for the comparison of the modern MM and classical methods under extremely non-spherical conditions. Results of analyses based on these data appear in the section entitled: Comparing Methods: Extremely Non-Spherical Data.

In conclusion, the ε values and H-F ε statistics suggest that extremely non-spherical data were not generated in the original phases of the current study. Furthermore, the moderately non-spherical data that were generated appear to be more of a function of sampling error rather than non-spherical population covariance structures. Results of analyses based on these data appear in the next subsection.

Comparing Methods: Moderately Non-Spherical Data

While remaining mindful of the caveats provided in the last section concerning the degree of non-sphericity of the data, Table 4.4.2 (p. 147) presents α_e's and power estimates for the modern MM KR test statistic when only the correct covariance model was fit to the data. Table 4.4.2 also presents this information for the classical methods (i.e., RM ANOVA conventional F-test, the RM ANOVA G-G and H-F corrected F-tests, and the MANOVA Wilks' Λ test statistic) when these models were fit to the correct data. That is, the RM ANOVA conventional F-test, G-G and H-F corrected tests were fit to the CS data and the MANOVA models was fit to the UN data.

As mentioned in the first chapter, the modern MM for longitudinal data is a generalization of the split-plot model, which, in turn, is a generalization of the CLM ((Fitzmaurice et al., 2004, pp. 187-197; Rencher, 2000, pp. 426-429; Vallejo & Livacic-Rojas, 2005). As a result, the mixed model with the compound symmetry covariance structure [MM(CS)] *F*-test reduces to the RM ANOVA conventional *F*-test under certain conditions[51] (Schaalje et al., 2002; Wright & Wolfinger, 1997). Similarly, the *F*-test associated with the unstructured [MM(UN)] covariance model is related to the MANOVA Lawley-Hotelling statistic (Wright & Wolfinger). Therefore, these models produce highly similar results in Table 4.4.2 (p. 147).

Inspection of Table 4.4.2 (p. 147) shows that the MM(CS)/RM ANOVA conventional *F*-test, MM(UN)/MANOVA models obtained α_e's with the 95% confidence interval limits[52] in all cases. This, of course, coincides with expectations because the assumptions of these models have been met under these circumstances. Further inspection shows that the majority of the other modern MMs (i.e., with IN, VC, CSH, ARH, & TOEPH) controlled Type I errors well with only three instances of inflated rates. However, α_e's were found to be slightly conservative in small sample situations in general. The G-G correction was found to be slightly to moderately conservative when the covariance structure was CS. However, the loss in statistical power[53] was only substantial (5% or greater) for small sample situations. In contrast, the H-F correction was found to approximate the nominal α level more closely and loss in statistical power was negligible even in small sample situations.

---

[51] Those conditions being design balance and *F*-test statistic degrees of freedom based on the independence model (i.e., unadjusted degrees of freedom).
[52] 95% confidence interval for columns 1-7 of Table 4.4.2: (.048, 0.52); interval for column 8: (.049, .051).
[53] That is, when the G-G power results are compared to the RM ANOVA conventional *F*-test.

Table 4.4.3 (p. 148) presents α_e's for the modern MM KR test statistic[54] and the classical method test statistics when results were aggregated across all seven true models. Inspection of Table 4.4.3 shows that both the H-F corrected univariate test statistic and the MM(UN)/MANOVA statistic were found to be robust under the conditions simulated in the current study with all α_e values falling within the 95% confidence interval limits.

Results for the modern MMs were previously discussed in section 4.1. Briefly, the CSH and AR models appear to perform well when α_e's are aggregated across true model with four and five estimates falling within the acceptable 95% confidence interval criterion, respectively. In addition, other α_e's for these models fell within the close proximity criterion. On average, however, the modern MM *F*-statistic obtained α_e's ranging from .042 to .047 across the eight marginal conditions.

The MM(CS)/conventional *F*-test results show that this test statistic is liberal under the conditions studied here with α_e's ranging from .057 to .061. Table 4.4.3 (p. 148) demonstrates that the G-G corrected test statistic is conservative under the eight marginal conditions with values ranging from .037 to .048. Further investigation demonstrated that the G-G corrected test statistic provided better Type I error control than initially suggested by Table 4.4.3. In fact, the G-G test statistic was substantially conservative only in conditions where the sample size was small and the number of measurement occasions was large. Under these conditions the G-G corrected statistic obtained α_e's as conservative as .028. Otherwise, the α_e's for the G-G test statistic are only moderately to slightly conservative ranging from .041 to .049.

Therefore, α_e's suggest that four models are in contention as superior methods for analyzing data under these conditions: MM(CSH), MM(AR), H-F corrected *F*-test, and MM(UN)/MANOVA. An investigation into empirical power estimates for these models follows.

---

[54] The mixed model information in Table 4.4.3 is a duplicate of information initially presented in Table 4.1.2.

Power estimates among the models in contention are presented in tables 4.4.4 (p. 149) & 4.4.5 (p. 150). Table 4.4.4 reports empirical power estimates for the modern MMs and classical methods across conditions relating to magnitude of correlation and number of measurement occasions. As expected, Table 4.4.4 shows that these conditions did not have a substantial impact on statistical power as estimates are relatively homogeneous across these conditions (columns). Furthermore, the MM(CSH) obtained higher power estimates under all conditions, however, by a negligible amount in some cases. Nonetheless, the MM(CSH) model obtained the highest marginal power estimate of .401 and was .078 greater than the MM(AR) power estimate of .323. Strictly speaking, the criterion for comparing power estimates proposed earlier is not directly applicable here because these estimates have been averaged across all true models. Therefore, the power of the correct model is irrelevant. However, if one were to apply this criterion in this instance, it may be concluded that all of these models provide comparable statistical power. However, there is evidence to support the conclusion that the MM(CSH) provides substantially greater statistical power than MM(AR).

To further investigate power issues, Table 4.4.5 (p. 150) reports power estimates by sample size and the magnitude of mean effect. Similar trends are evident, with MM(CSH) obtaining higher power estimates in the majority of conditions as well as the highest value for the overall marginal condition. However, differences in power were found not to be greater than .09. Additionally, MM(AR) consistently obtained the lower power, suggesting that the other models may be preferred.

In summary, the MM(CSH), H-F corrected test, and the MM(UN)/MANOVA models provide superior Type I error control compared to the other models considered here, although in many cases the difference in α_e's was quite small. Furthermore, these models provide comparable statistical power under these conditions. Implications of these findings are discussed in the next chapter.

Extremely Non-Spherical Data

As mentioned earlier, another phase of the current study was added in order to allow for the comparison of the modern MM and classical methods under extremely non-spherical conditions.

In order to induce a greater degree of non-sphericity, the structures that demonstrated the most non-sphericity[55] in the original data generation were chosen for inclusion: the ARH and UN patterns. As expected, greater non-sphericity was observed with more measurement occasions as well as higher degrees of correlation. Therefore, data for this phase were generated with only six measurement occasions and $r = .50$. Additionally, the heterogeneity of the variance with respect to Time was increased from 6 to 12 times greater at the final measurement occasion than the initial occasion. Thus, previous experimental variables of the number of measurement occasions, the degree of serial correlation, and the presence of non-constant variances were held fixed for this phase. Data were generated under the three differing sample size conditions ($N = 10, 30, 60$). Due to time constraints, 5,000 replications were generated.

These two new population covariance matrices obtained ε values of .670 and .492 for the ARH and UN structures, respectively. Sample-based H-F ε statistics for these structures were as follows: the ARH data obtained H-F ε values with a mean of 0.709 and a range of 0.306 – 1.700. The UN data obtained H-F ε values with a mean of 0.524 and a range of 0.237 – 1.330. Table A23 (p. 211) displays these two additional covariance matrices.

[55] As measured by ε.

Comparing Methods: Extremely Non-Spherical Data

Table 4.4.6 (p. 151) reports α_e's for both the extremely non-spherical ARH and UN data across the three sample size conditions and a marginal condition. Values of α_e's for the modern MM have been aggregated across all nine fitted models. Estimates in columns 1-3 of Table 4.4.6 are based on 5,000 replications. Consequently, the corresponding 95% confidence interval is (.044, .056). In contrast, column 4 estimates are based on 15,000 replications, and, as a result, the corresponding 95% confidence interval is (.047, .053).

For the less non-spherical ARH true model, the modern MMs, the H-F univariate statistics, and the MANOVA approach all produced results demonstrating acceptable Type I error control across all conditions. As expected, the conventional *F*-test obtained consistently liberal results and the G-G statistic was consistently conservative (especially so in the small sample size condition).

Table 4.4.6 (p. 151) also reports results for the extremely non-spherical UN data. Under these conditions, the modern MM α_es are inflated ranging from .057 to .062. Furthermore, the H-F statistic also obtained inflated rates except in the large sample situation. The G-G statistic, while found conservative under previous conditions, demonstrated acceptable levels of Type I error control, however, these results were not as stable as the MANOVA results. As expected, the MANOVA test provides consistent error control, even under these extreme conditions.

In order to further investigate the performance of the modern MMs in these extreme non-spherical conditions (for the UN data only), Table 4.4.7 (p. 152) reports a dissaggregation of the modern MMs so that the performance of specific fitted covariance matrices can be assessed. As expected, both models that assume independence of observations (IN & VC), obtained excessively conservative α_e's ranging from .014 to .032. In addition, the MM(CS)/RM ANOVA conventional *F*-test demonstrated excessively inflated α_e's under these extreme non-spherical conditions with rates ranging from .080 to .099. Further inspection of Table 4.4.7 demonstrates that only the G-G corrected test statistic and the MM(UN)/MANOVA approaches provided consistent Type I error control under these extreme conditions.

Table 4.4.8 (p. 153) reports power estimates for these non-spherical conditions. While results are ambiguous for the moderately non-spherical ARH data, the MM(UN)/MANOVA models are clearly more powerful under the extreme conditions. Specifically, the MM(UN)/MANOVA model was found to be 2.5 times more powerful than the G-G univariate test with a difference of .28 (the G-G test being the only other that obtained acceptable α_e's under these conditions)[56].

In summary, the MM(CSH), H-F corrected test, and the MM(UN)/MANOVA model provide superior Type I error control and comparable statistical power under moderately non-spherical conditions. In contrast, when the data are extremely non-spherical (ε = .492), both the G-G corrected test and the MM(UN)/MANOVA approaches control Type I errors, however, the MM(UN)/MANOVA was found to be clearly more powerful. Implications of these findings will be discussed in the next chapter.

Secondary Research Question 5

5) What are the modern MM α_e's *for the test of the interaction in repeated measures data with a between-subjects factor?*

[56] One should be mindful that under this extreme non-spherical condition, the MM(UN)/MANOVA is the correct model for the population covariance structure used to generate the data.

The objective of this question was to ascertain the effect of the conditions considered in the current study on the α_e's of the interaction test when a between-subjects factor is present in the design. It is well known that the main effect test for Group in a design including both within and between-subjects factors is not dependent on the covariance structure of the data with respect to Time (Keppel, 1991, p. 378). In contrast, it has been demonstrated that the covariance structure of the data with respect to Time has an impact on the Group x Time interaction test in these designs.

Table 4.5.1 (p. 154) reports α_e's for the interaction test for data with two groups and a variable number of measurement occasions (t = 3 or 6) and group sample sizes (n_j = 5, 15, & 30). These results have been aggregated across the seven true covariance structures for all models appearing in the table. Results show that the MM(UN)/MANOVA model provided acceptable Type I error control across all conditions studied. Most other covariance structures for the modern MM did not perform well, except for the AR pattern. The AR model obtained acceptable α_e's under all but one of the eight marginal conditions[57]. Furthermore, the G-G and H-F univariate tests, which were found to be in contention under conditions discussed earlier, were found not to perform as well here. Specifically, the G-G statistic proved to be substantially conservative under many of the conditions. Furthermore, while the H-F statistic closely approximated the nominal α level under all conditions, it only met the confidence interval criterion for robustness used in the current study under the large sample condition.

Thus, results from Table 4.5.1 (p. 154) demonstrate that the MM(AR) and MM(UN)/MANOVA models provide acceptable Type I error control for the interaction

[57] It is very possible that the inflated α_e value for the AR model in the moderate sample size condition (α_e = 0.052) is the result of sampling error because the same model obtained an acceptable α_e value in small sample conditions.

test under these conditions. Alternatively, results suggest that the H-F corrected test would be statistically valid in large sample situations. These results are discussed in the next chapter.

Summary

This chapter presented the results of the current simulation study with respect to the one preliminary, three primary, and two secondary research questions of interest. Briefly, the preliminary investigation demonstrated that the KR option for computing test statistics was found to provide superior Type I error control over the Between/Within method, Satterthwaite approximation, and the sandwich estimator.

Results from primary investigations identified 14 surrogate covariance structures; at least one for each of the seven true models. Further, results suggested that the either the CSH or UN structures may be the best models to fit to data in applied settings when the true covariance structure is unknown. Next, rates of information criteria selecting an appropriate covariance model were estimated and reported. Overall, these rates were found to be higher than "naïve" accuracy rates[58] reported in many previous studies that entertained a comparable number of candidate models (Gomez et al., 2005; Guerin & Stroup, 2000; Keselman, Algina, et al., 1998; Keselman et al., 1999; Vallejo & Livacic-Rojas, 2005). Finally, α_e's were estimated by information criteria. These values were found to be slightly conservative under small and moderate sample conditions, however, within close proximity of the nominal α level. Perhaps more importantly, though, none of these α_e's were found to be inflated beyond the upper limit of the 95% confidence

[58] "Naive" accuracy rates being those that do not consider the impact of surrogate covariance models.

intervals used to evaluate robustness in the current study. Furthermore, empirical power estimates were found to be comparable among models selected by the five information criteria.

Results from secondary investigations demonstrated that the MM(CSH), H-F corrected test, and the MM(UN)/MANOVA model provide superior Type I error control and comparable statistical power under moderate non-sphericity conditions. However, with extremely non-spherical data, only the G-G corrected test and the MM(UN)/MANOVA approaches provided acceptable Type I error control, however, the MM(UN)/MANOVA approach was found to be clearly more powerful. Finally, results showed that the MM(AR) and MM(UN)/MANOVA models provided acceptable levels of Type I error control for the interaction test of the Group x Time design. These results and their implications for statistical modeling of longitudinal data in applied settings are discussed in the next chapter.

Table 4.i.1

Empirical Type I Error Rates by Test Statistic Option and Marginal Conditions

Test statistic Option	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Overall Marginal
N ≈	420K	300K	360K	360K	240K	240K	240K	720K
Between/Within	.068	.072	.057	.079	.089	.060	.054	.068
Satterthwaite	.060	.062	.053	.066	.072	.055	.052	.060
Kenward/Roger	**.050**	.048	**.050**	**.049**	.048	**.050**	**.050**	**.049**
Sandwich Estimator	.141	.139	.080	.202	.256	.097	.071	.141

(a) Error rates when only the correct model was fit to the data and then aggregated across all true models

Table 4.i.2

Empirical Type I Error Rates (b) by Test Statistic Option and Each Individual Condition

Test statistic Option	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Overall Marginal
N ≈	420K	300K	360K	360K	240K	240K	240K	720K
Between/Within	.062	.056	.052	.070	.080	.054	**.049**	.061
Satterthwaite	.056	**.050**	**.050**	.060	.066	**.050**	.048	.055
Kenward/Roger	.047	.041	.046	.045	.045	.046	.045	.045
Sandwich Estimator	.143	.143	.081	.204	.259	.098	.071	.143

(b) Empirical Type I error rates aggregated across all eight incorrect candidate models and then all seven true models

Table 4.1.1

Empirical Type I Error Rates and Power Estimates for All True by Potential Surrogate Combinations

True Covariance Structure	Potential Surrogate Structures								
	IN	VC	CS	CSH	AR	ARH	TOEP	TOEPH	UN
IN	**.049** (**.54**)	.048 (**.52**)	**.049** (**.53**)	.048 (**.52**)	**.049** (**.53**)	.048 (**.52**)	**.050** (**.52**)	.047 (**.50**)	.048 (**.49**)
VC	.061 (**.28**)	**.051** (**.31**)	.060 (**.28**)	**.051** (**.30**)	.061 (**.27**)	**.050** (**.30**)	.058 (**.27**)	**.049** (**.30**)	**.051** (**.29**)
CS	.008 (.49)	.008 (.48)	**.050** (**.65**)	.047 (**.63**)	.039 (.52)	.041 (.52)	**.051** (**.63**)	.048 (**.61**)	**.051** (**.59**)
CSH	.016 (.22)	.012 (**.25**)	.059 (**.36**)	.046 (**.35**)	.048 (**.25**)	.043 (**.28**)	.056 (**.32**)	.047 (**.34**)	**.049** (**.33**)
ARH	.032 (**.24**)	.027 (**.27**)	.065 (**.32**)	.052 (**.32**)	.057 (**.26**)	**.049** (**.28**)	.056 (**.25**)	**.049** (**.28**)	**.050** (**.27**)
TOEPH	.024 (**.23**)	.020 (**.26**)	.062 (**.33**)	.048 (**.33**)	.053 (**.26**)	.046 (**.28**)	.056 (**.27**)	.048 (**.29**)	**.050** (**.28**)
UN	.020 (.22)	.013 (.25)	.065 (**.35**)	**.049** (**.35**)	.052 (.25)	.043 (**.27**)	.061 (**.33**)	.046 (**.34**)	**.051** (**.37**)

* IN=Independence, VC=Variance Components, CS=Compound Symmetry, CSH=Heterogeneous Compound Symmetry, AR=Autoregressive, ARH=Heterogeneous Autoregressive, TOEP=Toeplitz, TOEPH=Heterogeneous Toeplitz, UN=Unstructured

Table 4.1.2

Surrogate Covariance Structures

True Covariance Structure	Potential Surrogate Structures								
	IN	VC	CS	CSH	AR	ARH	TOEP	TOEPH	UN
IN	NA		1		1		1		
VC		NA		1		1		1	1
CS			NA				1		1
CSH				NA					1
ARH						NA		1	1
TOEPH								NA	1
UN				1					NA

* IN=Independence, VC=Variance Components, CS=Compound Symmetry, CSH=Heterogeneous Compound Symmetry, AR=Autoregressive, ARH=Heterogeneous Autoregressive, TOEP=Toeplitz, TOEPH=Heterogeneous Toeplitz, UN=Unstructured

Table 4.1.3

Mean Multivariate Distance (u) between the Model-Estimated and Sample-Based Covariance Matrices

True Covariance Structure	Potential Surrogate Structures								
	IN	VC	CS	CSH	AR	ARH	TOEP	TOEPH	UN
IN	4.27	3.62	4.12	3.49	4.12	3.47	3.57	2.93	1.92
VC	6.89	3.60	6.72	3.47	6.69	3.45	6.18	2.91	1.91
CS	8.10	7.46	4.10	3.78	5.32	4.59	3.56	2.91	1.91
CSH	11.06	7.47	7.42	3.79	8.24	4.59	6.78	2.92	1.91
ARH	9.58	5.98	8.15	4.74	6.82	3.46	6.31	2.90	1.91
TOEPH	9.73	6.13	7.71	4.28	6.96	3.55	6.36	2.92	1.91
UN	11.62	8.01	8.32	4.39	9.11	5.28	7.66	3.33	1.91

* IN=Independence, VC=Variance Components, CS=Compound Symmetry, CSH=Heterogeneous Compound Symmetry, AR=Autoregressive, ARH=Heterogeneous Autoregressive, TOEP=Toeplitz, TOEPH=Heterogeneous Toeplitz, UN=Unstructured

Table 4.1.4

Empirical Type I Error Rates for Mixed Models Aggregated Across the Seven True Models

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.035	.013	.031	.028	.032	.029	.028	.030
Variance Components	.029	.011	.027	.024	.027	.025	.025	.026
Compound Symmetry	.058	.061	.057	.060	.059	.059	.058	.058
Heterogeneous Compound Symmetry	**.049**	.048	**.049**	.048	.046	**.050**	**.050**	.049
Autoregressive	.052	**.049**	.052	**.051**	**.051**	**.051**	**.051**	.051
Heterogeneous Autoregressive	.046	.045	.048	.044	.045	.047	.046	.046
Toeplitz	.056	.055	.055	.056	.056	.056	.056	.056
Heterogeneous Toeplitz	.048	.047	.048	.047	.044	**.049**	**.049**	.048
Unstructured	**.050**	**.050**	**.051**	**.049**	**.050**	**.050**	**.050**	**.050**
Mean	.047	.042	.046	.045	.046	.046	.046	.046

* 95% confidence interval: (.049, .051), except for the overall marginal condition: (.050, .050)

Table 4.1.5

Empirical Power Rates for Mixed Models Aggregated Across the Seven True Models

Fitted Model	Small Mean Effect *N* = 10	*N* = 30	*N* = 60	Moderate Mean Effect *N* = 10	*N* = 30	*N* = 60	Large Mean Effect *N* = 10	*N* = 30	*N* = 60	Marginal
Independence	.04	.10	.24	.07	.29	.59	.13	.50	.86	.31
Variance Components	.04	.09	.23	.07	.30	.66	.14	.56	.92	.33
Compound Symmetry	.08	.17	.34	.14	.40	.71	.23	.62	.92	.40
Heterogeneous Compound Symmetry	.07	.14	.31	.13	.40	.73	.23	.65	.94	.40
Autoregressive	.06	.12	.25	.11	.31	.61	.17	.52	.87	.34
Heterogeneous Autoregressive	.06	.11	.24	.10	.32	.66	.17	.57	.91	.35
Toeplitz	.07	.14	.28	.13	.36	.65	.22	.57	.89	.37
Heterogeneous Toeplitz	.06	.12	.26	.13	.38	.70	.22	.62	.93	.38
Unstructured	.07	.11	.23	.13	.37	.69	.22	.62	.93	.38
Mean	.06	.12	.26	.11	.35	.67	.19	.58	.91	.36

Table 4.2.1

Information Criteria Selection Rates for Correct and Surrogate Models

True Model	Selection	AIC	AICC	HQIC	BIC	CAIC
Independence	Correct	.63	.67	.67	.82	.91
	Surrogate	.23	.23	.20	.13	.08
	Appropriate	.86	.90	.87	.95	.99
Variance Components	Correct	.62	.65	.67	.74	.72
	Surrogate	.31	.23	.28	.15	.06
	Appropriate	.93	.88	.95	.89	.78
Compound Symmetry	Correct	.63	.67	.66	.73	.75
	Surrogate	.11	.08	.11	.04	.01
	Appropriate	.74	.74	.77	.78	.76
Heterogeneous Compound Symmetry	Correct	.64	.66	.67	.70	.66
	Surrogate	.07	.03	.09	.03	.01
	Appropriate	.72	.69	.75	.73	.66
Heterogeneous Autoregressive	Correct	.61	.62	.64	.65	.60
	Surrogate	.07	.03	.08	.03	.01
	Appropriate	.68	.65	.72	.68	.60
Heterogeneous Toeplitz	Correct	.11	.08	.07	.03	.01
	Surrogate	.07	.03	.08	.03	.01
	Appropriate	.18	.11	.15	.06	.02
Unstructured	Correct	.34	.23	.26	.12	.06
	Surrogate	.39	.46	.50	.62	.62
	Appropriate	.73	.69	.76	.74	.68
Marginal	Correct	.51	.51	.52	.54	.53
	Surrogate	.18	.15	.19	.15	.11
	Appropriate	.69	.67	.71	.69	.64

Table 4.2.2

Information Criteria Selection Rates for Incorrect Models

True Model	Selection	AIC	AICC	HQIC	BIC	CAIC
Independence	Underfit	NA	NA	NA	NA	NA
	Overfit	.14	.10	.13	.05	.01
	Total Incorrect	.14	.10	.13	.05	.01
Variance Components	Underfit	.06	.09	.03	.09	.19
	Overfit	.02	.03	.02	.02	.02
	Total Incorrect	.07	.12	.05	.11	.22
Compound Symmetry	Underfit	.07	.08	.06	.09	.13
	Overfit	.20	.18	.17	.14	.11
	Total Incorrect	.26	.26	.23	.22	.24
Heterogeneous Compound Symmetry	Underfit	.09	.13	.07	.12	.21
	Overfit	.20	.18	.18	.15	.13
	Total Incorrect	.28	.31	.25	.27	.34
Heterogeneous Autoregressive	Underfit	.22	.28	.20	.28	.38
	Overfit	.10	.07	.08	.04	.02
	Total Incorrect	.32	.35	.28	.32	.40
Heterogeneous Toeplitz	Underfit	.82	.89	.85	.94	.98
	Overfit	.00	.00	.00	.00	.00
	Total Incorrect	.82	.89	.85	.94	.98
Unstructured	Underfit	.27	.31	.24	.26	.32
	Overfit	NA	NA	NA	NA	NA
	Total Incorrect	.27	.31	.24	.26	.32
Marginal	Underfit	.25	.30	.24	.30	.37
	Overfit	.11	.09	.10	.07	.05
	Total Incorrect	.31	.33	.29	.31	.36

Table 4.3.1

Empirical Type I Error Rates by Information Criteria

Information Criteria / Classical model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
AIC	.047	.047	.047	.048	.045	**.049**	**.049**	.047
AICC	.047	.047	.046	.048	.045	**.049**	**.049**	.047
HQIC	.048	.047	.047	.048	.045	**.049**	**.049**	.047
BIC	.047	.047	.046	.048	.045	.048	**.049**	.047
CAIC	.046	.046	.045	.048	.045	.048	**.049**	.047
RM ANOVA Conventional F-test	.058	.061	.057	.060	.059	.059	.058	.058
RM ANOVA G-G	.043	.043	.048	.039	.037	.045	.047	.043
RM ANOVA H-F	**.051**	**.051**	**.051**	**.050**	**.051**	**.051**	**.050**	**.050**
MANOVA - Wilks' Λ	**.050**	**.050**	**.051**	**.049**	**.050**	**.050**	**.050**	**.050**

* 95% confidence interval: (.049, .051), except for the overall marginal condition: (.050, .050)

Table 4.3.2

Empirical Power Estimates by Information Criteria

Information Criteria / Classical model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
AIC	0.38	0.39	0.37	0.39	0.16	0.41	0.60	0.39
AICC	0.37	0.39	0.37	0.39	0.16	0.41	0.60	0.39
HQIC	0.38	0.39	0.37	0.39	0.16	0.41	0.59	0.39
BIC	0.37	0.39	0.36	0.39	0.16	0.40	0.59	0.39
CAIC	0.37	0.39	0.36	0.39	0.16	0.40	0.59	0.39
RM ANOVA Conventional F-test	0.38	0.43	0.38	0.43	0.20	0.42	0.60	0.41
RM ANOVA G-G	0.35	0.39	0.36	0.38	0.15	0.39	0.57	0.37
RM ANOVA H-F	0.37	0.40	0.36	0.41	0.18	0.40	0.58	0.39
MANOVA - Wilks' Λ	0.36	0.38	0.38	0.37	0.14	0.40	0.59	0.37

Table 4.3.3

Empirical Type I Error Rates for Fitted Models Selected by AIC by True Model

True Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.049	NA	**.050**	.048	.053	.047	.047	.049
Variance Components	.052	NA	.054	**.050**	**.051**	.053	.052	.052
Compound Symmetry	.045	.047	.044	.049	.041	.048	**.050**	.046
Heterogeneous Compound Symmetry	.044	.046	.044	.046	.040	.048	.048	.045
Heterogeneous Autoregressive	**.050**	.048	.047	**.049**	.045	**.051**	**.051**	.048
Heterogeneous Toeplitz	.045	.047	.045	.046	.043	.048	.047	.046
Unstructured	.047	.048	.046	.048	.044	.049	**.050**	.047
Mean	.047	.047	.047	.048	.045	.049	**.049**	.047

Table 4.3.4

Empirical Type I Error Rates for Fitted Models Selected by BIC by True Model

True Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.049	NA	**.050**	.048	.053	.047	.047	.049
Variance Components	.052	NA	.054	**.050**	.052	.054	.052	.052
Compound Symmetry	.044	.046	.042	.049	.039	.047	**.049**	.045
Heterogeneous Compound Symmetry	.043	.045	.042	.046	.039	.047	.047	.044
Heterogeneous Autoregressive	.049	.048	.046	**.050**	.045	**.050**	**.051**	.048
Heterogeneous Toeplitz	.044	.046	.044	.046	.042	.047	.046	.045
Unstructured	.046	.047	.045	.047	.043	.047	**.049**	.046
Mean	.047	.046	.046	.048	.045	.048	.049	.047

Table 4.4.1

ε Values for Population Covariance Matrices

Covariance Structure	Amount of Correlation	Number of Measurement Occasions 3	6
IN	NA	1.000	1.000
VC	NA	.840	.840
CS	.3	1.000	1.000
	.5	1.000	1.000
CSH	.3	.850	.851
	.5	.854	.854
ARH	.3	.835	.787
	.5	.807	.693
TOEPH	.3	.845	.823
	.5	.824	.742
UN	.3	.785	.791
	.5	.750	.773

Table 4.4.2

Empirical Type I Error Rates and Statistical Power: Correct Model Fit Only

True/Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	**.049**	NA	**.050**	**.048**	.053	.046	.047	**.049**
	(.54)	NA	(.51)	(.57)	(.28)	(.59)	(.74)	(.54)
Variance Components	**.051**	NA	.053	**.049**	**.048**	.054	**.052**	**.051**
	(.31)	NA	(.27)	(.34)	(.11)	(.30)	(.50)	(.31)
Compound Symmetry	**.051**	**.049**	**.049**	**.050**	**.049**	**.050**	**.051**	**.050**
	(.61)	(.68)	(.63)	(.67)	(.40)	(.70)	(.83)	(.65)
Heterogeneous Compound Symmetry	.047	.046	.047	.046	.042	**.049**	**.048**	.046
	(.34)	(.37)	(.33)	(.37)	(.13)	(.36)	(.56)	(.35)
Heterogeneous Autoregressive	**.051**	**.048**	**.050**	**.049**	.046	**.052**	**.051**	**.049**
	(.28)	(.28)	(.28)	(.28)	(.10)	(.27)	(.47)	(.28)
Heterogeneous Toeplitz	**.048**	**.048**	**.049**	.047	.044	**.051**	**.049**	.048
	(.29)	(.29)	(.29)	(.28)	(.10)	(.28)	(.49)	(.29)
Unstructured	**.051**	**.051**	**.051**	**.051**	**.052**	**.050**	**.051**	**.051**
	(.35)	(.39)	(.38)	(.36)	(.12)	(.39)	(.60)	(.37)
RM ANOVA Conventional F-test	**.051**	**.049**	**.049**	**.050**	**.049**	**.050**	**.051**	**.050**
	(.61)	(.68)	(.63)	(.67)	(.40)	(.70)	(.83)	(.65)
RM ANOVA G-G	.041	.039	.044	.036	.031	.042	.047	.040
	(.58)	(.66)	(.62)	(.63)	(.35)	(.69)	(.83)	(.62)
RM ANOVA H-F	**.049**	.047	.047	**.048**	.045	**.049**	**.050**	.048
	(.60)	(.68)	(.62)	(.66)	(.39)	(.70)	(.83)	(.64)
MANOVA - Wilks' Λ	**.051**	**.051**	**.051**	**.051**	**.052**	**.050**	**.051**	**.051**
	(.35)	(.39)	(.38)	(.36)	(.12)	(.39)	(.60)	(.37)

Table 4.4.3

Empirical Type I Error Rates for Comparing Mixed Models with Classical Methods under Marginal Conditions

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.035	.013	.031	.028	.032	.029	.028	.030
Variance Components	.029	.011	.027	.024	.027	.025	.025	.026
Compound Symmetry	.058	.061	.057	.060	.059	.059	.058	.058
Heterogeneous Compound Symmetry	**.049**	.048	**.049**	.048	.046	**.050**	**.050**	.049
Autoregressive	.052	**.049**	.052	**.051**	**.051**	**.051**	**.051**	.051
Heterogeneous Autoregressive	.046	.045	.048	.044	.045	.047	.046	.046
Toeplitz	.056	.055	.055	.056	.056	.056	.056	.056
Heterogeneous Toeplitz	.048	.047	.048	.047	.044	**.049**	**.049**	.048
Unstructured	**.050**	**.050**	**.051**	**.049**	**.050**	**.050**	**.050**	**.050**
Mean	.047	.042	.046	.045	.046	.046	.046	.046
RM ANOVA Conventional F-test	.058	.061	.057	.060	.059	.059	.058	.058
RM ANOVA G-G	.043	.043	.048	.039	.037	.045	.047	.043
RM ANOVA H-F	**.051**	**.051**	**.051**	**.050**	**.051**	**.051**	**.050**	**.050**
MANOVA - Wilks' Λ	**.050**	**.050**	**.051**	**.049**	**.050**	**.050**	**.050**	**.050**

* Values appearing in bold print fall within the 95% confidence interval of (.049, .051)

Table 4.4.4

Empirical Power Estimates for Comparing Mixed Models with Classical Methods under Marginal Conditions

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Marginal
Independence	.32	.27	.30	.33	.29
Variance Components	.34	.29	.31	.35	.32
Compound Symmetry	.38	.43	.38	.41	.41
Heterogeneous Compound Symmetry	.39	.41	.38	.41	.40
Autoregressive	.33	.31	.34	.33	.32
Heterogeneous Autoregressive	.35	.33	.35	.35	.34
Toeplitz	.36	.38	.36	.37	.37
Heterogeneous Toeplitz	.37	.38	.37	.38	.38
Unstructured	.37	.39	.38	.37	.38
Mean	.36	.35	.35	.36	.36
RM ANOVA Conventional F-test	.38	.43	.38	.41	.41
RM ANOVA G-G	.35	.39	.35	.37	.37
RM ANOVA H-F	.37	.40	.36	.39	.39
MANOVA - Wilks' Λ	.37	.39	.38	.37	.38

Table 4.4.5

Empirical Power Estimates for Comparing Mixed Models with Classical Methods under Marginal Conditions

Fitted Model	Small Mean Effect *N* = 10	Small Mean Effect *N*= 30	Small Mean Effect *N* = 60	Moderate Mean Effect *N* = 10	Moderate Mean Effect *N*= 30	Moderate Mean Effect *N* = 60	Large Mean Effect *N* = 10	Large Mean Effect *N*= 30	Large Mean Effect *N* = 60	Marginal
Independence	.04	.10	.24	.07	.29	.59	.13	.50	.86	.31
Variance Components	.04	.09	.23	.07	.30	.66	.14	.56	.92	.33
Compound Symmetry	.08	.17	.34	.14	.40	.71	.23	.62	.92	.40
Heterogeneous Compound Symmetry	.07	.14	.31	.13	.40	.73	.23	.65	.94	.40
Autoregressive	.06	.12	.25	.11	.31	.61	.17	.52	.87	.34
Heterogeneous Autoregressive	.06	.11	.24	.10	.32	.66	.17	.57	.91	.35
Toeplitz	.07	.14	.28	.13	.36	.65	.22	.57	.89	.37
Heterogeneous Toeplitz	.06	.12	.26	.13	.38	.70	.22	.62	.93	.38
Unstructured	.07	.11	.23	.13	.37	.69	.22	.62	.93	.38
Mean	.062	.122	.264	.112	.349	.667	.193	.581	.909	.362
RM ANOVA Conventional F-test	.08	.17	.34	.14	.40	.71	.23	.62	.92	.40
RM ANOVA G-G	.05	.12	.27	.12	.37	.67	.21	.59	.91	.37
RM ANOVA H-F	.07	.15	.32	.13	.38	.68	.22	.59	.91	.38
MANOVA - Wilks' Λ	.07	.11	.23	.13	.37	.69	.22	.62	.93	.38

Table 4.4.6

Empirical Type I Error Rates for Extremely Non-Spherical Data

True Structure	Information Criteria	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
ARH (ε = .670)	Mixed Model (KR Approximation)	**.048**	**.051**	**.049**	**.049**
	RM ANOVA Conventional F-test	.073	.073	.070	.072
	RM ANOVA G-G	.035	.042	**.044**	.041
	RM ANOVA H-F	**.054**	**.050**	**.048**	**.051**
	MANOVA - Wilks' Λ	**.049**	**.053**	**.048**	**.050**
UN (ε = .492)	Mixed Model (KR Approximation)	.060	.062	.057	.060
	RM ANOVA Conventional F-test	.096	.099	.080	.092
	RM ANOVA G-G	**.047**	**.054**	**.044**	**.048**
	RM ANOVA H-F	.063	.058	**.046**	.056
	MANOVA - Wilks' Λ	**.051**	**.048**	**.053**	**.051**

*ARH = Heterogeneous Autoregressive; UN = Unstructured
**RM ANOVA = Repeated Measures Analysis of Variance; G-G = Greenhouse-Geisser; H-F = Huynh-Feldt; MANOVA = Multivariate Analysis of Variance
*** $N \approx$ 5,000 for columns 1-3; 95% confidence interval = (.044, .056)
**** $N \approx$ 15,000 for marginal condition; 95% confidence interval = (.046, .054)

Table 4.4.7

Empirical Type I Error Rates for Extremely Non-Spherical Data: Expanded View of UN

Fitted Model	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.032	.024	.020	.025
Variance Components	.017	.014	.014	.015
Compound Symmetry	.097	.097	.078	.091
Heterogeneous Compound Symmetry	.068	.082	.072	.074
Autoregressive	.079	.076	.069	.075
Heterogeneous Autoregressive	.059	.068	.062	.063
Toeplitz	.086	.084	.082	.084
Heterogeneous Toeplitz	**.051**	.068	.064	.061
Unstructured	**.051**	**.048**	**.054**	**.051**
Mean	.060	.062	.057	.060
RM ANOVA Conventional F-test	.096	.099	.080	.092
RM ANOVA G-G	**.047**	**.054**	**.044**	**.048**
RM ANOVA H-F	.063	.058	**.046**	.056
MANOVA - Wilks' Λ	**.051**	**.048**	**.053**	**.051**

* Columns 1-3: N = 5,000; 95% confidence interval = (.044, .056)
**Columns 4: N = 15,000; 95% confidence interval = (.047, .053)

Table 4.4.8

Empirical Power Estimates for Models Fit to Extremely Non-Spherical Data

True Structure	Fitted Model	Small Mean Effect *N* = 10	*N*= 30	*N* = 60	Moderate Mean Effect *N* = 10	*N*= 30	*N* = 60	Large Mean Effect *N* = 10	*N*= 30	*N* = 60	Marginal
ARH (ε = .670)	Mixed Model (KR)	.05	.06	.08	.07	.14	.27	.10	.31	.59	.19
	RM ANOVA Conventional F-test	.07	.10	.11	.11	.20	.34	.16	.40	.67	.24
	RM ANOVA G-G	.04	.06	.07	.06	.14	.27	.09	.31	.59	.18
	RM ANOVA H-F	.05	.07	.08	.09	.16	.28	.13	.33	.61	.20
	MANOVA - Wilks' Λ	.05	.06	.08	.06	.13	.26	.07	.28	.59	.18
UN (ε = .492)	Mixed Model (KR)	.06	.08	.10	.08	.20	.35	.13	.40	.70	.23
	RM ANOVA Conventional F-test	.10	.11	.13	.13	.23	.39	.20	.46	.81	.28
	RM ANOVA G-G	.04	.06	.07	.06	.14	.27	.10	.30	.65	.19
	RM ANOVA H-F	.06	.07	.08	.08	.15	.27	.13	.32	.67	.20
	MANOVA - Wilks' Λ	.06	.12	.22	.12	.59	.93	.25	.95	1.00	.47

Table 4.5.1

Empirical Type I Error Rates for the Test of Interaction (Aggregated across all seven true models)

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.035	.013	.031	.029	.033	.030	.028	.030
Variance Components	.029	.011	.026	.024	.027	.025	.024	.025
Compound Symmetry	.058	.061	.057	.060	.059	.059	.058	.059
Heterogeneous Compound Symmetry	**.049**	.048	.048	.048	.046	**.050**	**.049**	.048
Autoregressive	**.051**	**.050**	**.051**	**.051**	**.050**	.052	**.051**	**.051**
Heterogeneous Autoregressive	.045	.045	.047	.044	.044	.046	.046	.046
Toeplitz	.056	.056	.054	.058	.056	.056	.056	.056
Heterogeneous Toeplitz	.048	.047	.047	.048	.045	**.050**	**.049**	.048
Unstructured	**.051**	**.049**	**.050**	**.051**	**.051**	**.051**	**.049**	**.050**
Mean	.047	.042	.046	.046	.046	.047	.046	.046
RM ANOVA Conventional F-test	.058	.061	.057	.060	.059	.059	.058	.059
RM ANOVA G-G	.042	.043	.047	.038	.036	.045	.046	.042
RM ANOVA H-F	.052	.053	.052	.052	.054	.052	**.050**	.052
MANOVA - Wilks' Λ	**.051**	**.049**	**.050**	**.051**	**.051**	**.051**	**.049**	**.050**

* 95% confidence interval: (.049, .051)

CHAPTER V: DISCUSSION

This chapter provides a synopsis of the current study, a discussion of the results, acknowledgement of the limitations and suggestions for future research, and recommendations for applied researchers working with longitudinal data. The synopsis revisits the motivation, purpose, and the design and methods of the study as well as providing a summary of the results. The discussion of the results is organized by each of the research questions. The limitations and suggestions for future research summarize the shortcomings of the current study and offer suggestions for further scientific inquiry of this topic. Finally, recommendations for applied researchers are provided.

Synopsis

Collecting data from the same subjects through time is attractive to researchers for two main reasons. *First*, research designs of this nature allow investigators to evaluate individual changes over time. *Second*, data collection mechanisms of this type reduce the error variance as subjects serve as their own control. These sampling designs produce “repeated measures” or “longitudinal” data.

The analysis of longitudinal data represents a departure from many traditional statistical analysis approaches where observations are assumed to be independent. In contrast, longitudinal data are often related through time as a function of being collected from the same subjects repeatedly. Therefore, special statistical models have been developed in order to evaluate these data.

Classical univariate methods for analyzing longitudinal data assume independence of subjects, but allow observations within subjects to be dependent. However, a restrictive assumption is placed on the quality of those associations, known as sphericity. In longitudinal data analysis, this most often equates to assuming that data collected further apart in time have the same relationship as data collected closer in time. Unfortunately, this assumption is often found to be untenable in applied longitudinal data analysis. Further, when sphericity is violated, the conventional *F*-test for the Time effect in the repeated measures analysis of variance (RM ANOVA) has been shown to be excessively liberal (Everitt, 2001, pp. 139-141; Rogan et al., 1979). In order to address this problem, corrections have been developed to adjust the *F*-test and lessen the rates of falsely rejecting the null hypothesis (Huynh & Feldt, 1976). Also, a multivariate approach for analyzing longitudinal data based on the classical linear model (CLM) is available that does not require sphericity. However, this approach may require many parameters to be estimated and therefore may lack statistical power in these situations.

Another univariate conceptualization for analyzing longitudinal data is based on modern mixed model (MM) methods. These methods obviate the sphericity issue by modeling the covariance of the data with respect to time instead of imposing a specific form upon it. Therefore, these methods seem more appropriate for longitudinal data analysis where the sphericity assumption is typically not met. Furthermore, proponents of this approach claim that it is more powerful than the MANOVA approach because modern MM methods are able to more parsimoniously model the covariance structure of the data.

However, if one chooses to use the modern MM approach, one faces new challenges. Specifically, the researcher is required to select a model for the covariance of the data. While many tools are available to help guide this choice, most methodologists suggest the use of information criteria. Information criteria are a set of quantitative indicators that incorporate a form of the maximized model likelihood function and a penalty for model complexity[59]. Thus, these criteria measure the goodness of fit of the model to the data.

Unfortunately, the existing research literature suggests that information criteria are not very accurate in selecting the correct covariance model among a set of possible candidate models (Ferron et al., 2002; Gomez et al., 2005; Keselman, Algina, et al., 1998; Vallejo & Livacic-Rojas, 2005). Marginal accuracy rates have been reported as low as 47% and 35% correct for two of the most common information criteria (i.e., AIC and BIC) (Keselman, Algina, et al., 1998). Some authors have suggested that these low accuracy rates may be negatively influenced by the existence of surrogate models that approximate the true model well (Gomez et al., 2005; Keselman, Algina, et al.).

The current study was multifaceted with preliminary, primary, and secondary objectives. The preliminary investigation was concerned with comparing four test statistic options in modern MMs through examination of empirical Type I error rates (α_e's). The three primary purposes were to 1) investigate the possible existence of surrogate covariance models as measured by α_e's and statistical power of the tests of the fixed effects in MMs, 2) estimate the rates of five information criteria in selecting appropriate covariance models, and 3) investigate the α_e's and statistical power of models selected by the five information criteria. The two secondary purposes

[59] Among other things (e.g., sample size).

were to 1) compare modern MM methods with classical methods of analyzing longitudinal data by examining α_e's and statistical power estimates and 2) investigate the α_e's of the Group x Time interaction test when a between-subjects factor was added to the design.

The current study addressed these objectives by performing a Monte Carlo computer simulation that generated data, fit the appropriate models, and obtained outcome measures (e.g., α_e's, power estimates, selection rates, etc.). The computer simulation was performed in three phases. In all three phases, normally-distributed longitudinal data were generated. In phases I and II, data were generated with a single-group repeated measures design[60]. Specifically, the purpose of phase I was to obtain α_e's. Thus, data were generated so that the null hypothesis of the Time main effect was known to be true. In phase II, the purpose was to obtain statistical power estimates. Consequently, data were generated for phase II where the alternative hypothesis was known to be true. In contrast, phase III data were generated from a design with one within-subjects factor (Time) and one between-subjects factor (Group). The purpose here was to obtain α_e's for the Group x Time interaction test. As a result, phase III data were generated so that the null hypothesis of a Group x Time interaction was known to be true[61].

Results demonstrated that the Kenward-Roger (KR) approximation provided superior Type I error control when compared to the Between/Within and Satterthwaite methods as well as the use of the sandwich estimator. Next, 14 surrogate covariance models were identified for seven true models (each true model obtained at least one surrogate). Further, rates for selecting *appropriate* covariance models were estimated to be approximately 16% greater than rates of selecting only the correct model across both the seven true models and the five information criteria. Differences in performance among these five information criteria were negligible. When

[60] That is, one within-subjects factor and no between-subjects factors.
[61] The null hypotheses of the Group and Time main effects were also specified to be true.

α_e's were compared for models selected by the five information criteria, reasonable Type I error control was found with no α_e's indicating inflated rates of falsely rejecting the null hypothesis. Additionally, differences in statistical power of models selected by each of the criteria were slight. Next, the MM(CSH), Huynh-Feldt (H-F) corrected RM ANOVA test, and the MM(UN)/MANOVA models were found to perform comparably under moderately non-spherical circumstances. Further analysis demonstrated that only the MM(UN)/MANOVA models performed at acceptable levels when the data were extremely non-spherical. Finally, the MM(AR) and MM(UN)/MANOVA models were found the provide superior Type I error control for the Group x Time interaction test.

Discussion

This section provides a discussion of the results organized by research question.

Preliminary Research Question

i) How do test statistics for the fixed effects of the mixed model compare with respect to α_e's *when the SAS PROC MIXED default (the Between/Within method), the Satterthwaite or KR approximations, or the sandwich estimator options are used?*

The objective of this research question was to assess the effectiveness of four test statistic options in controlling α_e's in modern MMs. Each one of these options is readily available in

SAS, version 9.1; therefore, accessibility of these methods is not an issue for the typical researcher.

Results demonstrated that the KR approximation provided superior Type I error control over the Between/Within, Satterthwaite, and sandwich estimator options. These results agreed with the conclusions of Guerin and Stroup (2000), who found the KR approximation to out perform the Containment and Satterthwaite methods. However, the Guerin and Stroup comparison was only made in small sample situations ($N = 12$, $n_j = 6$). The results of the current study extend the findings of Guerin and Stroup to both moderate and large sample conditions[62] ($N = 30$ & 60). Under these larger sample size conditions, the KR approximation was still found to provide superior Type I error control.

Furthermore, Guerin and Stroup (2000) found that the KR approximation did not provide acceptable levels of Type I error control when the covariance model of the modern MM was misspecified. In contrast to their findings, the current study found that, on average, the KR approximation provided slightly conservative Type I error control when the incorrect covariance model was fit to the data (i.e., $\alpha_e = .045$, on average; see Table 4.i.2, p. 134 and tables A2 – A7, pp. 190-195). Compared to the other methods evaluated in the current study, these results were deemed desirable and ultimately acceptable. Obvious exceptions to this general rule are: 1) fitting IN or VC models to data that exhibit autocorrelation yields consistently conservative test statistics with $\alpha_e = .018$, on average, and 2) fitting the CS model to non-spherical data (i.e., essentially fitting the RM ANOVA conventional F-test model) results in consistently liberal tests with an mean $\alpha_e = .060$. Empirical Type I error rates were not held to the nominal α level under these circumstances, even with the use of the KR approximation.

---

[62] Moderate and large sample sizes with respect to social science longitudinal research (Keselman, Huberty et al., 1998).

In conclusion, results of the current study support the recommendations of both Guerin and Stroup (2000) and Littell et al. (2006): the KR approximation should be used in modeling situations involving repeated measures data using modern MM methods. Furthermore, the SAS degrees of freedom default[63] (the Between/Within method) should be avoided as it produces test statistics with inflated α_e's of the magnitude of .064[64]. The Satterthwaite method was found to provide more control than that of the Between/Within method; however, still slightly liberal with $\alpha_e = .057$, on average.

Finally, the sandwich estimator was found to produce comparable α_e's regardless of whether the correct or incorrect covariance model was fit to the data. This supports the claims that the sandwich estimator is robust to covariance model misspecification (Fitzmaurice et al., 2004). On the surface, this makes the sandwich estimator seem like an attractive alternative to many of the covariance modeling issues discussed here. Unfortunately, results from the current study demonstrated that the sandwich estimator does not provide acceptable levels of Type I error control with sample sizes that are typical of social science longitudinal data analysis. This is not to say that it is not effective in larger sample situations.

Primary Research Questions

1) Do surrogate covariance structures exist? If so, which structures serve as acceptable approximations for a given population or correct structure and under what conditions?

[63] The Between/Within method is the default for models fit in PROC MIXED using only a REPEATED statement.
[64] Aggregated across situations where both the correct and incorrect models were fit to the data.

The objective of this question was to identify covariance structures that are comparable to others, if they exist. Surrogate models of this type could then serve as substitutes for the correct model. It was hypothesized that the existence of surrogate models would have a substantial impact on the rates of selecting an appropriate covariance model using information criteria. Furthermore, the current investigation was expected to yield valuable information regarding the severity of covariance misspecification on α_e's and statistical power.

Table 4.1.1 (p. 135) demonstrated that the CSH and TOEPH models did not obtain acceptable levels of α_e's when these were the correct models. That is, the CSH model was found to be conservative and the TOEPH model to be liberal, even when each was the correct model. This is consistent with the findings of both Gomez et al. (2005) and Robertson (1996) who found lack of Type I error control for complex models even when the correct model was fit to the data. This may indicate an inability of the modern MM approach to properly model data of these types. Conversely, it is possible that a sufficient level of precision of these Monte Carlo simulations was not obtained. Furthermore, the Robertson study used the Containment degrees of freedom method which has been shown to be problematic (Gomez et al.).

Overall, 14 surrogate covariance models were identified; at least one for each of the seven true covariance models used in the current study. As expected, simpler models (IN and VC) obtained more surrogates than more complex models. That is, most surrogates overfit the true model with only one occurrence of a surrogate underfitting the data (i.e., the surrogate CSH model for UN). Further, variance homogeneity/heterogeneity appears to have a substantial impact on which models are considered comparable. For example, the IN model (a homogeneous variance model) obtained four surrogates, all of which exhibited homogeneous variances. Similarly, the VC model (a heterogeneous variance model) obtained three surrogates, all of which exhibited heterogeneous variances. In fact, no surrogates were identified in the current study that did not match the variance structure of the true model. However, this is not surprising because it has been documented that homogeneous variance models that are fit to heterogeneous variance data produce test statistics with inflated α_e's (Guerin & Stroup, 2000). Results from the current study provide further evidence of this phenomenon. For example, when the IN model was fit to the VC data, $\alpha_e = .061$. Similarly, when the CS model was fit to the CSH data, $\alpha_e = .059$. More generally, all homogenous variance models[65] fit to the VC data obtained inflated α_e's ranging from .058 to .061.

In contrast, when heterogeneous variance models are fit to homogeneous variance data (i.e., overfitting the variance structure and therefore overfitting the covariance structure as a whole), we would generally expect acceptable Type I error control, but a loss of statistical power. Therefore, if a candidate model was not found to be a surrogate in this situation, one would expect the lack of comparability to be due to insufficient statistical power. However, this was not observed in the current study. Instead, results demonstrated in these situations that α_e's became slightly conservative and, on average, power decreased to a small extent, but this decrease was minor. Counter-intuitively, candidate models were rejected as surrogates in the majority of these cases because of conservative α_e's and not power issues. This, of course, may simply be an artifact of the criteria used in the current study to identify surrogates.

Results from this investigation also demonstrate that fitting models assuming independence of observations to data that exhibit autocorrelation result in substantially conservative tests. For example, fitting either the IN or VC models to data that exhibit autocorrelation results in α_e's of .018, on average. In these cases, substantial loss of statistical power may occur. For example, the most extreme case found in the current study took place

[65] That is, all homogeneous variance models that allowed for autocorrelation (i.e., excluding the IN model).

when the VC model was fit to the CS data. In this situation, the difference in power was .17. Thus, it is important to account for autocorrelation in the data in order to obtain tests that adhere to the nominal α level and exhibit reasonable statistical power.

Robertson (1996) concluded that the CS model was a good approximation for others. This conclusion was not supported by the results of the current study. This discrepancy is most likely attributable to two reasons. *First*, Robertson only generated data under three true models: CS, autoregressive plus a common covariance (AR+CC), and UN. While the AR+CC model was not included in the current investigation, results demonstrated that fitting the CS model to UN data[66] produced inflated α_e's of the magnitude of .065[67]. It is possible that the UN data in the Robertson study were less non-spherical than those generated in the current study. If this were the case, it may explain this discrepancy in results. *Second*, Robertson used the Containment degrees of freedom method exclusively. As mentioned earlier, problems with this method are suspected (Gomez et al., 2005). Results from the current study using the KR approximation demonstrate that the CS model does not in fact serve as a good approximation of other structures unless the true structure is IN. Otherwise, the CS model obtained inflated α_e's as great as .065. Therefore, the CS model should be avoided unless especially convincing evidence of the tenability of sphericity is available[68].

Keselman et al. (1999) found the ARH model to be a good approximation for others when the Satterthwaite method was used to obtain test statistics. While the ARH model did not result in inflated α_e's when fit to data with other covariance structures in the current study, α_e's were moderately conservative in some cases and loss in statistical power was as great as .13.

---

[66] Once again, this is analogous to assuming sphericity and using the RM ANOVA conventional *F*-test.
[67] However, power estimates were found to be comparable with the true model (UN): $(1\text{-}\beta)_{CS} = .35$; $(1\text{-}\beta)_{UN} = .37$.
[68] Because the MM(CS) model and the RM ANOVA conventional *F*-test are synonymous, this recommendation is the same offered by many other methodologist concerning the use of the conventional *F*-test when the assumption of sphericity is suspect.

Keselman et al. did not investigate statistical power and therefore may not have been aware of the loss of power when making this recommendation. Therefore, results from the current study demonstrate that the ARH model is only a surrogate for the VC model due largely to power issues.

Results from the current study suggest that the CSH, but especially the UN model may serve as an acceptable approximation to any one of the seven true structures studied. Specifically, the UN model produced either acceptable or slightly conservative α_e's when applied to data with any other true covariance structure. That is, the UN model never produced inflated α_e's. Furthermore, while the UN model uses the most degrees of freedom to estimate parameters, the average loss in statistical power observed in the current study was only .03. Of course, it is expected that the UN model would not perform as well in situations with a larger number of measurement occasions. That is, a greater loss in power is expected with designs incorporating a larger number of measurement occasions. However, these situations are not typical in applied social science research. Therefore, in modeling situations that resemble those simulated in the current study, the UN structure is considered the best overall covariance structure for modeling data with an unknown covariance structure.

2) What are the selection rates of a particular information criterion with respect to selecting a) the correct model, b) a surrogate model, c) and an appropriate model? What are the selection rates with respect to a) underfitting or b) overfitting the data?

The main objective of this research question was to re-assess the performance of five information criteria in selecting appropriate covariance models for modern MMs taking into account the influence of surrogate models. It was hypothesized that information criteria select appropriate covariance models at higher rates than their accuracy rates reported in previous research. If this was found to be the case, it was suspected that applied researchers may opt to use both information criteria and the modern MM approach more often[69].

The current study found that the five information criteria selected the correct model approximately 52% of the time, on average. Estimates ranged from 51% for AIC and AICC to 54% for BIC. Therefore, only negligible differences in accuracy among the five information criteria were observed. These results appear contradictory to those of Robertson (1996) and Ferron et al. (2002) who found accuracy rates to range from 75% to 95% and 66% to 79% for differing information criteria, respectively. However, in both of these investigations only a small number of candidate models were considered (3 and 2, respectively).

Rates found in the current study for AIC are comparable to those reported by Keselman, Algina et al. (1998): 51% in the current study and 47% in the Keselman, Algina et al. study. However, a large discrepancy was observed between the results for BIC. The current study found BIC to select accurately 54% of the time, while BIC was found to do so only 35% of the time in the Keselman, Algina et al. study. Because BIC is a function of the sample size of the data to be fit, it is possible that the differences in sample sizes between the two studies accounts for the differences in reported accuracy

[69] Reviews of longitudinal data analysis in the social sciences report that the majority of applied researchers use classical methods of analysis (Keselman, Huberty et al., 1998; Kowalchuk et al., 1996).

rates. However, samples sizes in the two studies were roughly comparable: N = 10, 30, & 60 in the current study and N = 30, 45, & 60 in the Keselman, Algina et al. study. Furthermore, these differences may be explained by the difference in the size of the set of candidate models: 9 in the current study and 11 in the Keselman, Algina, et al. study.

Similar to previous studies, the current investigation found that AIC tended to select more complex models, while BIC and CAIC tended to select more parsimonious models (Gomez et al., 2005; Guerin & Stroup, 2000; Robertson, 1996). Further, like Guerin and Stroup, information criteria were found to be more accurate for simpler true models, becoming less and less accurate as the structure of the true model became more complex.

Beyond "naïve" accuracy rates, the current study found that, on average, the five information criteria selected a surrogate model that controlled Type I errors and provided comparable statistical power to the true model .16 of the time. These rates ranged from .11 to .19 with CAIC selecting surrogates the least and HQIC and AIC selecting surrogates .19 and .18 of the time, respectively. As a result, the five information criteria were found to select appropriate covariance models .68 of the time, on average.

In contrast to selection rates of correct and appropriate covariance models, the current study also estimated the rates of these information criteria in underfitting and overfitting the data. The marginal rate of underfitting the data across the five information criteria was approximately .29, ranging from .24 to .37 for HQIC and CAIC, respectively. The marginal rate of overfitting the data was approximately .08, ranging from .05 to .11 for CAIC and AIC, respectively.

Guerin and Stroup (2000) found that underfitting the data generally results in inflated α_e's. Therefore, the fact that the five information criteria were found to underfit data approximately 3.5 times more often than overfitting the data is especially problematic. In contrast, however, the current study provided evidence to suggest that not necessarily underfitting the data, but rather not accounting for heterogeneous variances was a major source of inflated α_e's. Perhaps underfitting is not as problematic as previously expected.

In conclusion, the current study found that, on average, information criteria select an appropriate covariance model 68% of the time. While this is a substantial increase over many "naïve" accuracy estimates found in the literature, these rates are still low. Keselman, Algina, et al. (1998), Guerin and Stroup (2000), and Gomez et al. (2005) conclude that one can not expect to select the correct covariance model based on information criteria alone. Ultimately, results from the current study provide further support of this conclusion; however, in terms of applied research, a more interesting issue is discussed next.

3) *Will the analysis be statistically valid if one uses a particular information criterion to select a covariance model? That is, under what conditions are the α_e's controlled for models selected by a given information criterion?*

The objective of this research question was to ascertain the statistical properties of tests based on models selected by the five information criteria in terms of α_e's and statistical power estimates. This question was considered especially relevant because researchers working in applied settings never know the true/population covariance structure. In this context, it is of less interest that a given information criterion chooses the true model. Rather, it is more important that a given information criteria selects a model that provides acceptable Type I error control and reasonable statistical power.

The current study found that while information criteria can not be relied on solely to select the correct model or even an appropriate model for a given set of data, on average, the five information criteria evaluated will select a model that provides acceptable Type I error control. This was found to be true for all five information criteria with very little variation among them (see Table 4.3.1, p. 142). Furthermore, very little variation with respect to statistical power was found among the five information criteria.

In previous research, investigators have explored these issues with regard to the α_e's of the interaction test in a within-and between-subjects design (Gomez et al., 2005; Robertson, 1996). However, no reported studies were found that evaluated the α_e's and power of the main effect test for Time. Nonetheless, Robertson found α_e's for the interaction test of models selected by AIC, HQIC, BIC, and CAIC to exhibit inflated α_e's ranging from .069 to .082 (i.e., CAIC and AIC, respectively). Once again, the difference in these findings may be attributed to the difference in statistical tests being evaluated or to Robertson's use of the Containment method. Similarly, Gomez et al., who evaluated the performance of the KR approximation, reported that α_e's were always higher than the nominal α value when models were selected by AIC or BIC. Once again, however, these results were based on an evaluation of the Group x Time interaction test.

In conclusion, researchers can be assured that the α_e's will be adequately controlled, on average, for the main effect test for Time. Furthermore, there is evidence to suggest that all five information criteria select models that provide comparable Type I error control and statistical power under these circumstances.

Secondary Research Questions

> *4) How does the mixed model approach compare to the classical methods of repeated measures analysis in the context of covariance model misspecification? More specifically, how does the mixed model Wald-type F-statistic compare to the RM ANOVA conventional F-statistic, the G-G or H-F corrections, or the MANOVA Wilks' Λ test statistic with respect to* α_e's*?*

The objective of this question was to compare the modern MM methods with the classical methods in terms of α_e's and statistical power. As mentioned previously, there is some near overlap here with the MM(CS) model and conventional *F*-test as well as the MM(UN) model and the MANOVA approach fitting similar models. Nonetheless, this investigation was deemed important because only one study was found in the literature review that systematically compared the G-G and H-F corrected *F*-tests with the modern MM approach.

Results from the current study demonstrated that MM(CSH), H-F corrected test, and the MM(UN)/MANOVA approach all provided acceptable Type I error control and comparable statistical power when modeling moderately non-spherical data. Furthermore, only the MM(UN)/MANOVA approach demonstrated acceptable control Type I error control when modeling extremely non-spherical data.

These results suggest that the MM(UN)/MANOVA model is most appropriate for analyzing data when the degree of non-sphericity is unknown. Alternatively, one could initially fit the RM ANOVA model, obtain either the G-G or H-F ε statistic, and make a decision concerning the degree of non-sphericity in the data. With this information, one could then select the MM(CSH), H-F corrected test, or the MM(UN)/MANOVA models if the data are only moderately non-spherical; or, otherwise, select the MM(UN)/MANOVA model if the data are extremely non-spherical. Ultimately, however, results from the current study show that the loss in statistical power from choosing the MM(UN)/MANOVA at the onset, regardless of the degree of non-sphericity, is only 5% to 6%.

On the surface, results from the current study seem to suggest that the classical methods perform just as well as the modern MM methods, and, indeed, under the conditions studied in this simulation there is evidence to support this conclusion. However, the modern MM approach allows for the economical analysis of missing data, the use of time-varying covariates, and the use parameterized growth curve models for the mean response[70] that allow for the analysis of data with unequal measurement occasions. These options are not possible in classical methods for analyzing longitudinal data. Therefore, the current study has laid the foundation for the comparison of these models by demonstrating that certain methods within both frameworks are comparable with respect to α_e's and statistical power. However, further research is needed in order to compare the methods under more general situations that often arise in longitudinal data analysis involving missing data, time-varying-covariates, unequal measurement occasions, etc.

5) What are the mixed model α_e's for the test of the interaction in repeated measures data with a between-subjects factor?

[70] As well as profile analysis models for the mean response, like those used in the current study.

The objective of this question was to ascertain whether the mixed model test statistics provided acceptable Type I error control for the Group x Time interaction test. This was considered important because many of longitudinal studies in psychology and the social sciences in general involve at least one between-subjects factor and, in these cases, differential trends through time among groups is often of primary interest.

Results from the current study demonstrated that the MM(AR) and MM(UN)/MANOVA models provide acceptable Type I error control for the interaction test, regardless of the true model. Alternatively, results suggest that the H-F corrected test would be statistically valid in large sample situations.

While no reported studies were found that reported results in this manner, a few studies did investigate the use of AIC and the Satterthwaite or KR approximation in the modern MM Group x Time interaction test (Keselman et al., 1999; Vallejo & Livacic-Rojas, 2005). Keselman et al. investigated the α_e's of the interaction test for AIC selected MMs[71] under the following conditions: sample size inequality, variance heterogeneity with respect to Time, variance heterogeneity with respect to Group, positive and negative pairings of sample size inequality and Group variance heterogeneity, normal and lognormal data, etc. These researchers found α_e's to range from .054 to .099 when AIC was used and the data were normally-distributed. Further, these α_e's were found to be highly influenced by pairing: α_e's under positive pairing conditions were closer (yet still inflated) to the nominal α value while highly inflated α_e's were encountered for negative pairings.

[71] With the Satterthwaite approximation.

Vallejo and Livacic-Rojas (2005) also investigated the Group x Time interaction test[72] for MMs under similar circumstances to Keselman et al. (1999). These authors found that BIC was unable to select models that produced satisfactory Type I error control. Therefore, they preferred AIC even though AIC was found to select MMs that produced inflated α_e's in small sample situations ($N = 30$). In general, Type I error control was acceptable for AIC selected MMs when analyzing normal data. Results were similar for the lognormal data generated; however, the authors stated that conservative values of α_e's were encountered under conditions with extreme values of skewness and kurtosis (i.e., $\gamma_{skew} = 3.00$, $\gamma_{kurtosis} = 21.00$).

Limitations and Suggestions for Future Research

The limitations of the current study were initially discussed at the end of the third chapter. Some of the most salient limitations follow. *First*, the current study was primarily concerned with data generated from a single-group repeated measures design with only one within-subjects factor (Time). There are, of course, many variations of this design including multiple within and between-subjects factors. Further investigations into the properties of these test statistics may be warranted. *Second*, the current study generated normally-distributed data. While research investigating the effect of non-normally distributed data on the Group x Time interaction test from similar designs is available (Gomez et al., 2005; Guerin & Stroup, 2000; Keselman et al. 1999; Vallejo & Livacic-Rojas, 2005), the effect of non-normally distributed data on the main effect of Time may be of interest. *Third*, the current study did not focus on data

[72] With the KR approximation.

that exhibited between-subjects random effects[73]. A more in-depth investigation into this type of data is warranted.

Limitations of the current study are especially important for the last research question. It has been well documented that group size heterogeneity, variance heterogeneity, and the pairing of these two factors has a profound impact on α_e's (Gomez et al., 2005; Keselman et al., 1999; Vallejo & Livacic-Rojas, 2005). Specifically, positive pairings are known to produce slightly inflated α_e's, while negative pairings are known to produce excessively inflated α_e's. Therefore, this phase of the current study is only a preliminary investigation and further research including these operative experimental factors is needed.

More generally, the identification of surrogates in the current study is largely a result of which criteria for evaluating α_e's and statistical power were chosen. While the current author performed a review of the existing literature, and therefore made an informed decision, the results would have most certainly been different if other criteria were chosen for use in the current study. A more standardized definition of conservative and liberal test statistics and corresponding criteria are needed. Furthermore, development of statistical power criteria may be of interest as well.

Next, the current study found that many of the modern MM and classical methods are comparable in terms of α_e's and statistical power. However, this is only the first step in the comparison of these methods. As mentioned earlier, comparison of these methods with missing data, time-varying covariates, etc. is needed.

Finally, longitudinal data in the current study were analyzed solely in SAS's PROC MIXED routine with REML[74] estimation. However, many researchers in the social sciences use

[73] Data generated under the CS population covariance matrix exhibit subject-specific random effects.
[74] Restricted Maximum Likelihood.

the HLM software available from Scientific Software International, Inc. HLM 6 incorporates Bayesian estimation procedures. The current author is unaware of any research that compares these estimation procedures. Further research may investigate whether or not the HLM Bayes estimates are subject to the same bias problems of REML estimates that make adjustments like that of the KR approximation necessary.

Recommendations for Applied Researchers

This section offers some basic recommendations and suggestions for applied researchers working with longitudinal data. *First*, research study design is always critically important in any well-informed and systematic research endeavor. The same is true for longitudinal studies. Specifically, avoid studies with a small number of subjects and a large number of measurement occasions as much as possible. Results from the current study demonstrate that α_e's are often inflated in these situations. *Second*, results from the current study support the recommendations of Guerin and Stroup (2000) and Littell et al. (2006) to use the KR approximation when fitting modern MMs to repeated measures or longitudinal data. *Next*, during model fitting procedures, researchers need to pay special attention to properly fitting the variance structure of the data. It is especially important not to fit models that assume homogeneous variances to data that exhibit variance heterogeneity. Results from the current as well as previous studies demonstrate that α_e's are likely to be inflated in these situations (Guerin & Stroup). Various tools can be used in order to evaluate the existence of heterogeneous variances in a given dataset. Obvious examples include simply inspecting the sample covariance matrix as well as various graphical tools like residual plots with respect to Time, etc. *Fourth*, researchers need to account for autocorrelation in the data, if it exists. If models that assume independence of observations are fit to data that exhibit autocorrelation, test statistics are likely to be substantially conservative and statistical power drastically reduced. *Next*, results from the current study support the recommendation of Guerin and Stroup (2000) to err on the side of overfitting the data as opposed to underfitting. Furthermore, results demonstrate that AIC is more likely to overfit the data when compared to the other four information criteria. Therefore, the current author recommends the use AIC over other information criteria because it overfits data. *Sixth*, loss in statistical power is not crippling if one fits a completely saturated model for the covariance structure of data when the true covariance structure is unknown to the researchers. That is, if one fits either the MM(UN) or MANOVA models as a safe bet to guard against inflated α_e's, results from the current simulation suggest a loss in power in the neighborhood of 5 to 6%. *Finally*, fit either MM(UN) or MANOVA models when extreme non-sphericity is encountered. That is, the current study suggests if H-F ε values are encountered less than or equal to .65, fit one of these two models.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19, 716-723.

Bozdogan, H. (1987). Model selection and Akaike Information Criterion (AIC) – The general theory and its analytical extensions. *Psychometrika*, 52(3), 345-37.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.

Bryk, A. S. & Raudenbush, S. W. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. London: Sage Publications, Inc.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multiple Inference: A Practical Information-Theoretical Approach* (2nd ed.). New York: Springer-Verlag.

Cohen, J. (1977). *Statistical power analysis of the behavioral sciences* (rev. ed.). New York: Academic Press.

Christensen, R. (1987). *Plane Answers to Complex Questions: The Theory of Linear Models*. New York: Springer-Verlag.

Davis, C. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer.

Delwiche, L. D. & Slaughter, S. J. (2003). *The Little SAS Book: A primer* (3rd ed.). Cary, NC: SAS Institute.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics*, 44(4), 959-971.

Diggle, P. J., Heagerty, P., Liang, K., & Zeger, S. L. (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford University Press.

Everit, S. B. (2001). *Statistics for Psychologists: An Intermediate Course*. London: Lawerance Erlbaum Associates.

Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. C. (2001). *SAS for Monte Carlo Studies: A Guide for Quantitative Researchers*. Cary, NC: SAS Institute, Inc.

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*, 37(3), 379-403.

Fitzmaurice, G.M., Laird, N. M., & Ware, J. H. (2004). *Applied Longitudinal Analysis*. Boston: Wiley-Interscience.

Gomez, E. V., Schaalje, G. B., & Fellingham, G. W. (2005). Performance of the Kenward-Roger Method when the covariance structure is selected using AIC and BIC. *Communications in Statistics – Simulation and Computation*, 34, 377-392.

Guerin, L. & Stroup, W. (2000). A simulation study to evaluate PROC MIXED analysis of repeated measures data. In Proceedings of the Twelfth Annual Conference on Applied Statistics in Agriculture. Manhattan: Kansas State University.

Hannan, E. J. & Quinn, B. G. (1979). Determination of the order of an autoregression. *Journal of Royal Statistical Society*, 41(2), 190-195.

Hurvich & Tsai. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297-307.

Hutchinson, S.R. & Bandalos, D. L. (1997). A guide to Monte Carlo simulation research for applied researchers. *Journal of Vocational Education Research*. 22(4), 233-245.

Huynh, H, & 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.

Kackar, R. N. & Harvile, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79, 853-862.

Kauermann, G. & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96, 1387-1396.

Keppel, G. (1991). *Design and Analysis: A Researcher's Handbook* (3rd Ed.). Upper Saddle River, NJ: Prentice Hall.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics – Simulation and Computation*, 27(3), 591-604.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999). The analysis of repeated measurements: A comparison of mixed-model Satterthwaite *F* test and a nonpooled adjusted degrees of freedom multivariate test. *Communications in Statistics – Theory and Methodology*, 28: 2968-2999.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3), 350-386.

Kowalchuk, R. K., Lix, L. M., & Keselman, H. J. (1996). The analysis of repeated measures designs. Paper presented at the Annual Meetinig of the Psychometric Society. Banff, AB.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). Boston: McGraw-Hill/Irwin.

Liang, K.Y. & Zeger, S.L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Lindsey, J. K. (1999). *Models for Repeated Measurements* (2nd ed.). Oxford University Press.

Lindstrom, M. J. & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83, 1014-1022.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS Sytem for Mixed Models*. Cary, NC: SAS Institute.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for Mixed Models* (2nd ed.). Cary, NC: SAS Institute.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Tutorial in biostatistics: Modeling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, 19, 1793-1819.

Manor, O. & Zucker, D. M. (2004). Small sample inference for the fixed effects in the mixed linear model. *Computational Statistics & Data Analysis*, 46, 801-817.

Maxwell, S. E. & Delany, H. D. (2000). *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. New Jersey: Lawerence Erlbaum Associates.

Pinheiro, J. C. & Bate, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Rencher, A. C. (2000). *Linear Models in Statistics*. New York: Wiley-Interscience.

Rencher, A. C. (2002). *Methods of Multivariate Analysis* (2nd ed.) .New York: Wiley-Interscience.

Ripley, B. E. (1987). *Stochastic Simulation*. New York: Wiley.

Robertson, J. M. (1996). Covariance structure selection, alpha levels, distributions of test statistics, and model misspecification in mixed models as implemented in SAS PROC MIXED. M.S. thesis, Department of Statistics, Brigham Young University, Provo, UT.

Rogan, J. C., Keselman, H. J., and Mendoza, J. L. (1979). Analysis of repeated measurements. *The British Journal of Mathematical and Statistical Psychology*, 32, 269-286.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models using SAS® Proc MIXED. *Journal of Agricultural, Biological, and Environmental Statistics,* 7(4), 512-524.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 24, 323-355.

Tonidandel, S., Overall, J. E., & Smith, F. (2004). Use of resampling to select among alternative error structure specifications for GLMM analyses of repeated measurements. *International Journal of Methods in Psychiatric Research*, 13(1), 24-33.

Vallejo, G. & Livacic-Rojas, P. (2005). Comparison of two procedures for analyzing small sets of repeated measures data. *Multivariate Behavioral Research*, 40(2), 179-205.

Verbeke, G. & Molenbergs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics – Simulation and Computation*, 22(4), 1079-1106.

Wolfinger, R. (1996). Heterogeneous variance-covariance structures for repeated measurements. *Journal of Agricultural, Biological, and Environmental Statistics*, 1, 205-23.

Wright, S. P. & Wolfinger, R. D. (1997). Repeated measures analysis using mixed models: Some simulation results. In T. G. Gregoire (Ed.), *Modelling Longitudinal and Spatially Correlated Data* (pp. 147-157). New York: Springer-Verlag.

APPENDIX

Table A1

Example of Permanent SAS Dataset

PROGRAM	SET	CONDITION	REPLICATE	SEED_INDEX	TRUE_STRUC	MO	N	SIGMASQ
1	1	1	1	1	UN	3	10	6
1	1	1	1	1	UN	3	10	6
1	1	1	1	1	UN	3	10	6
1	1	1	1	1	UN	3	10	6
1	1	1	1	1	UN	3	10	6
1	1	1	1	1	UN	3	10	6
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
1	1	1	1	1	UN	3	10	6
1	1	1	1	1	UN	3	10	6
1	1	1	1	1	UN	3	10	6
1	1	1	1	1	UN	3	10	6
1	1	1	1	1	UN	3	10	6
1	1	1	1	1	UN	3	10	6
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
1	1130	5	130	1130	UN	6	30	6
1	1130	5	130	1130	UN	6	30	6

Table A1 (continued)

CORR	FITTED_MOD	STDERR	Neg2LogLik	Parms	AIC	AICC	HQIC	BIC
.3	Independence	Between/Within	247.1	1	249.1	249.2	248.7	249.4
.3	Independence	Satterthwaite	247.1	1	249.1	249.2	248.7	249.4
.3	Independence	Kenward/Roger	247.1	1	249.1	249.2	248.7	249.4
.3	Independence	Sandwhich	247.1	1	249.1	249.2	248.7	249.4
.3	Variance Components	Between/Within	233.4	3	239.4	24.4	238.4	24.3
.3	Variance Components	Satterthwaite	233.4	3	239.4	24.4	238.4	24.3
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.3	Unstructured	Kenward/Roger	228	6	240	244.2	238	241.8
.3	Unstructured	Sandwhich	228	6	240	244.2	238	241.8
.3	RM ANOVA CON-F	NA	NA	NA	NA	NA	NA	NA
.3	RM ANOVA G-G	NA	NA	NA	NA	NA	NA	NA
.3	RM ANOVA H-F	NA	NA	NA	NA	NA	NA	NA
.3	MANOVA - Wilks	NA	NA	NA	NA	NA	NA	NA
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.3	Unstructured	Satterthwaite	1474.6	21	1516.6	1522.7	1526	1546
.3	Unstructured	Kenward/Roger	1474.6	21	1516.6	1522.7	1526	1546

Table A1 (continued)

CAIC	NumDF	DenDF	FValue	ProbF	EPSILON_GG	EPSILON_HF	PDR	PDR_S	PDR_M	U_P_S
25.4	2	18	.52	.6030	NA	NA	1	1	1	1.7895
25.4	2	27	.52	.6002	NA	NA	1	1	1	1.7895
25.4	2	27	.52	.6002	NA	NA	1	1	1	1.7895
25.4	2	18	.37	.6945	NA	NA	1	1	1	1.7895
243.3	2	18	.31	.7385	NA	NA	1	1	1	1.7895
243.3	2	12.7	.31	.7400	NA	NA	1	1	1	1.7895
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
247.8	2	8	.30	.7505	NA	NA	1	1	1	1.7895
247.8	2	9	.37	.6995	NA	NA	1	1	1	1.7895
NA	2	18	.63	.5464	NA	NA	NA	NA	NA	NA
NA	1	1.3	.63	.4677	.5708	.599	NA	NA	NA	NA
NA	1	1.8	.63	.4744	.5708	.599	NA	NA	NA	NA
NA	2	8	.30	.7505	NA	NA	NA	NA	NA	NA
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
1567	5	29	.72	.6128	NA	NA	1	1	1	4.0155
1567	5	25	.62	.6845	NA	NA	1	1	1	4.0155

Table A1 (continued)

U_P_M	U_S_M	Reason	Status	pdG	pdH
6.4592	-3.2008	Convergence criteria met.	0	1	1
6.4592	-3.2008	Convergence criteria met.	0	1	1
6.4592	-3.2008	Convergence criteria met.	0	1	1
6.4592	-3.2008	Convergence criteria met.	0	1	1
1.6119	-.1604	Convergence criteria met.	0	1	1
1.6119	-.1604	Convergence criteria met.	0	1	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
1.7895	1.0456	Convergence criteria met.	0	1	1
1.7895	1.0456	Convergence criteria met.	0	1	1
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
4.0155	3.7355	Convergence criteria met.	0	1	1
4.0155	3.7355	Convergence criteria met.	0	1	1

Table A2

Population Covariance Matrices for both Independence & Variance Components Structures

Structure	3 measurement occasions			6 measurement occasions					
	100	0	0	100	0	0	0	0	0
		100	0		100	0	0	0	0
IN			100			100	0	0	0
							100	0	0
								100	0
									100
	ε = 1.000			ε = 1.000					
	100	0	0	100	0	0	0	0	0
		300	0		200	0	0	0	0
VC			600			300	0	0	0
							400	0	0
								500	0
									600
	ε = .840			ε = .840					

Table A3

Population Covariance Matrices for the Compound Symmetry Structure

Correlation	3 measurement occasions			6 measurement occasions					
	100	30	30	100	30	30	30	30	30
		100	30		100	30	30	30	30
.3			100			100	30	30	30
							100	30	30
								100	30
									100
		ε = 1.000				ε = 1.000			
	100	50	50	100	50	50	50	50	50
		100	50		100	50	50	50	50
.5			100			100	50	50	50
							100	50	50
								100	50
									100
		ε = 1.000				ε = 1.000			

Table A4

Population Covariance Structure for the Heterogeneous Compound Symmetry Structure

Correlation	3 measurement occasions			6 measurement occasions					
.3	100	51.96	73.48	100	42.43	51.96	60.00	67.08	73.48
		300	127.28		200	73.48	84.85	94.87	103.92
			600			300	103.92	116.19	127.28
							400	134.16	146.97
								500	164.32
									600
		ε = .850				ε = .851			
.5	100	86.60	122.47	100	7.71	86.60	100.00	111.80	122.47
		300	212.13		200	122.47	141.42	158.11	173.21
			600			300	173.21	193.65	212.13
							400	223.61	244.95
								500	273.86
									600
		ε = .854				ε = .854			

Table A5

Population Covariance Matrices for the Heterogeneous Autoregressive Structure

Correlation	3 Measurement Occasions			6 Measurement Occasions					
	100	51.96	22.05	100	42.43	15.59	5.40	1.81	0.60
		300	127.28		200	73.48	25.46	8.54	2.81
.3			600			300	103.92	34.86	11.46
							400	134.16	44.09
								500	164.32
									600
		ε = .835				ε = .787			
	100	86.60	61.24	100	70.71	43.30	25.00	13.98	7.65
		300	212.13		200	122.47	70.71	39.53	21.65
.5			600			300	173.21	96.82	53.03
							400	223.61	122.47
								500	273.86
									600
		ε = .807				ε = .693			

Table A6

Population Covariance Matrices for the Heterogeneous Toeplitz Structure

Correlation	3 Measurement Occasions			6 Measurement Occasions					
	100	51.96	47.77	100	42.43	33.77	30.00	23.48	14.70
		300	127.28		200	73.48	55.15	47.43	36.37
.3			600			300	103.92	75.52	63.64
							400	134.16	95.53
								500	164.32
									600
		ε = .845				ε = .823			
	100	86.60	79.61	100	70.71	56.29	50.00	39.13	24.49
		300	212.13		200	122.47	91.92	79.06	60.62
.5			600			300	173.21	125.87	106.07
							400	223.61	159.22
								500	273.86
									600
		ε = .824				ε = .742			

Table A7

Population Covariance Structures for the Unstructured Pattern

Correlation	3 Measurement Occasions			6 Measurement Occasions					
	100	47.11	106.31	100	38.47	75.17	85.40	63.73	53.89
		300	62.37		200	36.01	50.06	56.60	71.01
.3			600			300	136.49	176.22	167.58
							400	127.46	63.20
								500	156.65
									600
		ε = .785					ε = .791		
	100	105.48	140.11	100	86.13	99.07	131.40	105.77	109.98
		300	156.98		200	90.63	164.33	124.91	148.26
.5			600			300	176.67	156.08	22.19
							400	203.48	187.14
								500	271.12
									600
		ε = .750					ε = .773		

Table A8

Frequencies of Nonconvergence

True Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Overall Marginal
Independence	11 (540,000)	NA NA	3 (270,000)	8 (270,000)	11 (180,000)			11 (540,000)
Variance Components	161 540,000	NA NA	73 (270,000)	88 (270,000)	161 (180,000)			161 (540,000)
Compound Symmetry								
Heterogeneous Compound Symmetry	10 (540,000)	2 (540,000)	10 (540,000)	2 (540,000)	12 (360,000)			12 (1,080,000)
Heterogeneous Autoregressive	9 (540,000)	2 (540,000)	5 (540,000)	6 (540,000)	11 (360,000)			11 (1,080,000)
Heterogeneous Toeplitz	8 (540,000)	4 (540,000)	8 (540,000)	4 (540,000)	12 (360,000)			12 (1,080,000)
Unstructured	12 (540,000)	2 (540,000)	12 (540,000)	2 (540,000)	14 (360,000)			14 (1,080,000)
Marginal	211 (3,240,000)	10 (2,160,000)	111 (2,700,000)	110 (2,700,000)	221 (1,800,000)			221 (5,400,000)

* Total number of models fit for a given condition in parentheses

Table A9

Empirical Type I Error Rates (a) by Test statistic Option and Each Individual Condition

Correlation	.3	.3	.3	.3	.3	.3	.5	.5	.5	.5	.5	.5
Time Points	3	3	3	6	6	6	3	3	3	6	6	6
Sample Size	10	30	60	10	30	60	10	30	60	10	30	60
	Condition											
Test statistic Option	1	2	3	4	5	6	7	8	9	10	11	12
N ≈	70K	70K	70K	70K	70K	70K	50K	50K	50K	50K	50K	50K
Between/Within	.063	.055	.053	.114	.065	.057	.063	.055	**.050**	.115	.064	.057
Satterthwaite	.057	.054	**.052**	.087	.058	.054	.056	.053	**.049**	.088	.056	.054
Kenward/Roger	**.049**	**.052**	**.051**	.047	**.050**	**.050**	.048	**.051**	.048	.047	.048	**.050**
Sandwich Estimator	.114	.069	.059	.398	.125	.082	.112	.069	.058	.399	.124	.083

(a) Error rates when only the correct model was fit to the data and then aggregated across all true models

Table A10

Empirical Type I Error Rates for the Independence True Model

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	**.049**	NA	**.050**	.048	.053	.046	.047	**.049**
Variance Components	.048	NA	**.049**	.047	**.050**	.047	.047	.048
Compound Symmetry	**.049**	NA	**.050**	.048	.052	.047	.047	**.049**
Heterogeneous Compound Symmetry	.048	NA	**.049**	.047	**.049**	.047	.047	.048
Autoregressive	**.049**	NA	**.049**	**.049**	.053	.047	.048	**.049**
Heterogeneous Autoregressive	.048	NA	**.049**	.047	**.050**	.048	.047	.048
Toeplitz	**.050**	NA	**.050**	**.050**	.055	.047	.047	**.050**
Heterogeneous Toeplitz	.047	NA	.048	.046	.047	.047	.047	.047
Unstructured	.048	NA	**.049**	.048	**.050**	.048	.047	.048
Mean	.048	NA	**.049**	.048	**.051**	.047	.047	.048

Table A11

Empirical Type I Error Rates for the Variance Components True Model

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.061	NA	.062	.060	.063	.060	.059	.061
Variance Components	**.051**	NA	.053	**.049**	.048	.054	.052	**.051**
Compound Symmetry	.060	NA	.061	.059	.061	.060	.059	.060
Heterogeneous Compound Symmetry	**.051**	NA	.053	**.049**	.048	.053	.052	**.051**
Autoregressive	.061	NA	.061	.060	.061	.061	.060	.061
Heterogeneous Autoregressive	**.050**	NA	.053	.048	.048	.052	**.051**	**.050**
Toeplitz	.058	NA	.060	.057	.057	.059	.059	.058
Heterogeneous Toeplitz	**.049**	NA	.052	.046	.045	.052	**.051**	**.049**
Unstructured	**.051**	NA	.053	**.049**	**.050**	.052	**.050**	**.051**
Mean	.055	NA	.056	.053	.053	.056	.055	.055

Table A12

Empirical Type I Error Rates for the Compound Symmetry True Model

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.013	.003	.010	.005	.009	.007	.007	.008
Variance Components	.013	.003	.010	.006	.010	.007	.007	.008
Compound Symmetry	**.051**	**.049**	**.049**	**.050**	**.049**	**.050**	**.051**	**.050**
Heterogeneous Compound Symmetry	**.049**	.045	.047	.046	.042	**.049**	**.050**	.047
Autoregressive	.040	.039	.043	.036	.039	.040	.040	.039
Heterogeneous Autoregressive	.040	.041	.044	.038	.041	.041	.041	.041
Toeplitz	.052	**.050**	**.050**	.052	.052	**.050**	**.051**	**.051**
Heterogeneous Toeplitz	**.049**	.047	.047	.048	.045	.048	**.050**	.048
Unstructured	**.051**	**.050**	**.051**	**.051**	**.051**	**.050**	**.051**	**.051**
Mean	.040	.036	.039	.037	.037	.038	.039	.038

Table A13

Empirical Type I Error Rates for the Heterogeneous Compound Symmetry True Model

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.023	.009	.020	.011	.019	.015	.014	.016
Variance Components	.018	.007	.016	.009	.014	.012	.011	.012
Compound Symmetry	.059	.059	.058	.060	.058	.060	.058	.059
Heterogeneous Compound Symmetry	.047	.046	.047	.046	.042	**.049**	.048	.046
Autoregressive	.048	.048	**.050**	.047	.047	**.050**	**.049**	.048
Heterogeneous Autoregressive	.041	.044	.045	.040	.041	.044	.043	.043
Toeplitz	.056	.056	.055	.058	.054	.057	.057	.056
Heterogeneous Toeplitz	.047	.046	.047	.047	.043	**.049**	**.049**	.047
Unstructured	**.049**	**.049**	.048	**.050**	.047	**.051**	**.049**	**.049**
Mean	.043	.041	.043	.041	.041	.043	.042	.042

Table A14

Empirical Type I Error Rates for the Heterogeneous Autoregressive True Model

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.039	.024	.028	.035	.034	.030	.031	.032
Variance Components	.033	.021	.024	.031	.027	.027	.027	.027
Compound Symmetry	.063	.066	.061	.069	.066	.064	.064	.065
Heterogeneous Compound Symmetry	.052	.052	**.050**	.055	**.049**	.053	.054	.052
Autoregressive	.060	.055	.055	.060	.056	.058	.058	.057
Heterogeneous Autoregressive	**.051**	.048	**.050**	**.049**	.046	.052	**.051**	**.049**
Toeplitz	.058	.054	.055	.058	.056	.056	.057	.056
Heterogeneous Toeplitz	**.051**	.047	**.050**	.048	.046	**.050**	**.051**	**.049**
Unstructured	.052	**.049**	.052	**.049**	**.050**	**.050**	**.051**	**.050**
Mean	**.051**	.046	.047	**.050**	.048	**.049**	**.049**	**.049**

Table A15

Empirical Type I Error Rates for the Heterogeneous Toeplitz True Model

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.030	.018	.023	.024	.027	.023	.022	.024
Variance Components	.024	.015	.020	.020	.022	.019	.018	.020
Compound Symmetry	.060	.063	.059	.065	.062	.064	.060	.062
Heterogeneous Compound Symmetry	**.049**	.048	.048	**.049**	.046	**.051**	**.049**	.048
Autoregressive	.053	.052	.052	.053	**.051**	.055	.053	.053
Heterogeneous Autoregressive	.046	.046	.048	.044	.045	.047	.046	.046
Toeplitz	.057	.056	.055	.058	.054	.058	.056	.056
Heterogeneous Toeplitz	.048	.048	**.049**	.047	.044	**.051**	**.049**	.048
Unstructured	**.050**	**.050**	**.050**	**.050**	**.050**	.052	**.049**	**.050**
Mean	.046	.044	.045	.046	.044	.047	.045	.045

Table A16

Empirical Type I Error Rates for the Unstructured True Model

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.028	.012	.025	.015	.023	.020	.018	.020
Variance Components	.019	.006	.015	.010	.015	.012	.011	.013
Compound Symmetry	.064	.065	.064	.066	.066	.064	.065	.065
Heterogeneous Compound Symmetry	**.050**	.047	**.050**	.047	.046	**.050**	**.050**	**.049**
Autoregressive	**.051**	.052	.053	**.050**	.052	**.051**	.052	.052
Heterogeneous Autoregressive	.041	.045	.045	.041	.044	.044	.042	.043
Toeplitz	.060	.061	.060	.061	.059	.060	.062	.061
Heterogeneous Toeplitz	.048	.045	.045	.047	.042	.048	**.049**	.046
Unstructured	**.051**	**.051**	**.051**	**.051**	.052	**.050**	**.051**	**.051**
Marginal	.046	.043	.045	.043	.044	.044	.044	.044

Table A17

Empirical Power Estimates for the Independence True Model

Fitted Model	Small Mean Effect *N* = 10	*N*= 30	*N* = 60	Moderate Mean Effect *N* = 10	*N*= 30	*N* = 60	Large Mean Effect *N* = 10	*N*= 30	*N* = 60	Marginal
Independence	.09	.23	.53	.18	.63	.96	.32	.90	1.00	.54
Variance Components	.08	.20	.48	.17	.61	.96	.32	.90	1.00	.52
Compound Symmetry	.09	.22	.51	.17	.62	.96	.32	.90	1.00	.53
Heterogeneous Compound Symmetry	.08	.19	.46	.17	.61	.96	.32	.90	1.00	.52
Autoregressive	.08	.22	.50	.18	.62	.96	.32	.90	1.00	.53
Heterogeneous Autoregressive	.08	.19	.46	.17	.61	.96	.32	.90	1.00	.52
Toeplitz	.08	.20	.45	.17	.60	.95	.32	.89	1.00	.52
Heterogeneous Toeplitz	.07	.16	.39	.17	.59	.95	.31	.89	1.00	.50
Unstructured	.07	.14	.32	.16	.56	.94	.31	.89	1.00	.49
Marginal	.08	.19	.46	.17	.61	.96	.32	.90	1.00	.52

Table A18

Empirical Power Estimates for the Variance Components True Model

Fitted Model	Small Mean Effect *N* = 10	*N*= 30	*N* = 60	Moderate Mean Effect *N* = 10	*N*= 30	*N* = 60	Large Mean Effect *N* = 10	*N*= 30	*N* = 60	Marginal
Independence	.07	.09	.13	.10	.21	.39	.18	.49	.82	.28
Variance Components	.06	.09	.14	.10	.25	.47	.18	.58	.89	.31
Compound Symmetry	.07	.09	.12	.10	.21	.39	.18	.49	.82	.28
Heterogeneous Compound Symmetry	.06	.09	.14	.09	.24	.47	.17	.57	.89	.30
Autoregressive	.07	.09	.12	.10	.21	.39	.17	.49	.82	.27
Heterogeneous Autoregressive	.06	.09	.13	.09	.24	.47	.17	.57	.89	.30
Toeplitz	.06	.09	.12	.09	.20	.38	.16	.47	.81	.27
Heterogeneous Toeplitz	.06	.09	.13	.08	.24	.47	.15	.56	.88	.30
Unstructured	.06	.09	.13	.08	.22	.45	.14	.53	.88	.29
Marginal	.06	.09	.13	.09	.23	.43	.17	.53	.86	.29

Table A19

Empirical Power Estimates for the Compound Symmetry True Model

Fitted Model	Small Mean Effect			Moderate Mean Effect			Large Mean Effect			
	N = 10	*N*= 30	*N* = 60	*N* = 10	*N*= 30	*N* = 60	*N* = 10	*N*= 30	*N* = 60	Marginal
Independence	.03	.08	.21	.12	.59	.93	.47	.98	1.00	.49
Variance Components	.03	.08	.21	.11	.57	.92	.40	.98	1.00	.48
Compound Symmetry	.11	.28	.52	.35	.84	.98	.75	1.00	1.00	.65
Heterogeneous Compound Symmetry	.09	.27	.51	.30	.83	.98	.70	1.00	1.00	.63
Autoregressive	.07	.15	.30	.17	.62	.95	.49	.99	1.00	.52
Heterogeneous Autoregressive	.07	.15	.30	.16	.61	.95	.44	.98	1.00	.52
Toeplitz	.11	.27	.51	.31	.82	.98	.67	1.00	1.00	.63
Heterogeneous Toeplitz	.09	.26	.51	.25	.81	.98	.59	.99	1.00	.61
Unstructured	.09	.25	.50	.22	.80	.98	.51	.99	1.00	.59
Marginal	.08	.20	.40	.22	.72	.96	.56	.99	1.00	.57

Table A20

Empirical Power Estimates for the Heterogeneous Compound Symmetry True Model

	Small Mean Effect			Moderate Mean Effect			Large Mean Effect			
Fitted Model	*N* = 10	*N*= 30	*N* = 60	*N* = 10	*N*= 30	*N* = 60	*N* = 10	*N*= 30	*N* = 60	Marginal
Independence	.02	.03	.06	.05	.12	.31	.11	.43	.82	.22
Variance Components	.02	.03	.07	.04	.16	.40	.12	.52	.89	.25
Compound Symmetry	.07	.11	.17	.13	.31	.57	.25	.68	.93	.36
Heterogeneous Compound Symmetry	.06	.10	.16	.10	.30	.58	.22	.69	.95	.35
Autoregressive	.06	.08	.11	.08	.17	.35	.14	.46	.83	.25
Heterogeneous Autoregressive	.05	.08	.12	.07	.20	.41	.15	.52	.89	.28
Toeplitz	.07	.10	.15	.10	.26	.51	.19	.61	.91	.32
Heterogeneous Toeplitz	.06	.10	.16	.09	.29	.57	.19	.67	.95	.34
Unstructured	.06	.10	.16	.09	.28	.56	.17	.65	.94	.33
Marginal	.05	.08	.13	.08	.23	.47	.17	.58	.90	.30

Table A21

Empirical Power Estimates for the Heterogeneous Autoregressive True Model

Fitted Model	Small Mean Effect *N* = 10	*N*= 30	*N* = 60	Moderate Mean Effect *N* = 10	*N*= 30	*N* = 60	Large Mean Effect *N* = 10	*N*= 30	*N* = 60	Marginal
Independence	.04	.06	.09	.07	.16	.34	.15	.44	.78	.24
Variance Components	.04	.06	.10	.07	.20	.42	.15	.52	.86	.27
Compound Symmetry	.08	.11	.16	.12	.26	.47	.22	.58	.87	.32
Heterogeneous Compound Symmetry	.06	.10	.16	.10	.27	.51	.20	.61	.90	.32
Autoregressive	.06	.09	.13	.09	.19	.37	.16	.46	.80	.26
Heterogeneous Autoregressive	.06	.08	.13	.09	.21	.43	.16	.52	.86	.28
Toeplitz	.06	.09	.12	.09	.19	.36	.15	.44	.79	.25
Heterogeneous Toeplitz	.05	.08	.13	.08	.21	.42	.14	.51	.86	.28
Unstructured	.06	.08	.13	.08	.20	.41	.13	.49	.84	.27
Marginal	.06	.08	.13	.09	.21	.41	.16	.51	.84	.28

Table A22

Empirical Power Estimates for the Heterogeneous Toeplitz True Model

Fitted Model	Small Mean Effect *N* = 10	*N*= 30	*N* = 60	Moderate Mean Effect *N* = 10	*N*= 30	*N* = 60	Large Mean Effect *N* = 10	*N*= 30	*N* = 60	Marginal
Independence	.03	.05	.08	.06	.15	.33	.13	.43	.78	.23
Variance Components	.03	.05	.09	.06	.19	.41	.14	.51	.86	.26
Compound Symmetry	.07	.11	.17	.13	.28	.50	.23	.60	.89	.33
Heterogeneous Compound Symmetry	.06	.10	.16	.10	.28	.52	.21	.63	.91	.33
Autoregressive	.06	.08	.12	.09	.19	.36	.16	.46	.80	.26
Heterogeneous Autoregressive	.05	.08	.13	.08	.21	.42	.15	.52	.86	.28
Toeplitz	.06	.09	.13	.09	.20	.38	.15	.48	.82	.27
Heterogeneous Toeplitz	.05	.09	.14	.08	.22	.44	.15	.54	.87	.29
Unstructured	.06	.09	.14	.08	.21	.44	.14	.51	.87	.28
Marginal	.05	.08	.13	.09	.21	.42	.16	.52	.85	.28

Table A23

Empirical Power Estimates for the Unstructured True Model

Fitted Model	Small Mean Effect			Moderate Mean Effect			Large Mean Effect			
	N = 10	*N*= 30	*N* = 60	*N* = 10	*N*= 30	*N* = 60	*N* = 10	*N*= 30	*N* = 60	Marginal
Independence	.03	.04	.06	.05	.13	.30	.11	.42	.82	.22
Variance Components	.02	.03	.06	.05	.15	.38	.11	.52	.91	.25
Compound Symmetry	.08	.11	.16	.13	.30	.55	.25	.66	.94	.35
Heterogeneous Compound Symmetry	.06	.09	.15	.10	.29	.57	.21	.69	.96	.35
Autoregressive	.06	.07	.10	.08	.17	.33	.14	.44	.83	.25
Heterogeneous Autoregressive	.05	.07	.10	.08	.18	.40	.14	.52	.90	.27
Toeplitz	.07	.10	.15	.11	.26	.50	.20	.62	.93	.33
Heterogeneous Toeplitz	.05	.09	.16	.09	.28	.57	.18	.68	.96	.34
Unstructured	.06	.11	.19	.10	.33	.64	.20	.74	.98	.37
Marginal	.05	.08	.12	.09	.23	.47	.17	.59	.91	.30

Table A24

Population Covariance Matrices for both ARH & UN Structures for Follow-up Analysis

Correlation	6 Measurement Occasions					
ARH	100	89.44	58.09	34.46	19.57	1.83
		320	207.85	123.29	70.00	38.73
			540	32.31	181.87	10.62
				760	431.51	238.75
					980	542.22
						1200
	ε = .670					
UN	102	128.95	164.92	49.72	174.30	30.61
		399.7	196.01	41.56	218.74	423.53
			543.1	225.53	162.83	338.58
				785.8	215.81	184.35
					1016	971.62
						1253
	ε = .492					

Table A25

Empirical Type I Error Rates for the Independence True Model: Test of the Interaction

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	**.050**	NA	**.050**	**.050**	.044	.056	**.049**	**.050**
Variance Components	**.049**	NA	**.050**	**.048**	.043	.055	**.049**	**.049**
Compound Symmetry	**.050**	NA	**.051**	**.049**	.045	.056	**.050**	**.050**
Heterogeneous Compound Symmetry	**.048**	NA	**.049**	**.048**	.042	.054	**.049**	**.048**
Autoregressive	**.049**	NA	**.051**	**.048**	.044	.055	**.049**	**.049**
Heterogeneous Autoregressive	**.049**	NA	**.050**	.047	.043	.054	**.049**	**.049**
Toeplitz	**.052**	NA	**.052**	.053	**.052**	.057	**.049**	**.052**
Heterogeneous Toeplitz	**.049**	NA	**.049**	**.049**	.045	.054	**.048**	**.049**
Unstructured	**.052**	NA	**.051**	.053	.054	.056	**.047**	**.052**
Mean	**.050**	NA	**.050**	**.050**	.046	.055	**.049**	**.050**
RM ANOVA Conventional F-test	**.050**	NA	**.051**	**.049**	.045	.056	**.050**	**.050**
RM ANOVA G-G	.040	NA	.046	.035	.029	**.047**	.045	.040
RM ANOVA H-F	**.049**	NA	**.050**	**.048**	.043	.055	**.049**	**.049**
MANOVA - Wilks' Λ	**.052**	NA	**.051**	.053	.054	.056	**.047**	**.052**

* 95% confidence interval = (.048, .052), except for sample size conditions: (.047, .053)

Table A26

Empirical Type I Error Rates for the Variance Components True Model: Test of the Interaction

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.062	NA	.061	.062	.066	.058	.062	.062
Variance Components	**.048**	NA	**.051**	.046	**.049**	**.047**	**.049**	**.048**
Compound Symmetry	.061	NA	.060	.062	.063	.058	.062	.061
Heterogeneous Compound Symmetry	**.048**	NA	**.050**	.046	**.048**	**.047**	**.049**	**.048**
Autoregressive	.061	NA	.058	.063	.063	.058	.061	.061
Heterogeneous Autoregressive	**.049**	NA	**.051**	.046	**.048**	**.048**	**.049**	**.049**
Toeplitz	.058	NA	.057	.060	.059	.056	.060	.058
Heterogeneous Toeplitz	**.048**	NA	**.050**	.047	**.048**	**.048**	**.049**	**.048**
Unstructured	**.049**	NA	**.050**	**.048**	**.051**	**.048**	**.048**	**.049**
Mean	.054	NA	.054	.053	.055	**.052**	.054	.054
RM ANOVA Conventional F-test	.061	NA	.060	.062	.063	.058	.062	.061
RM ANOVA G-G	.044	NA	**.049**	.039	.039	.045	**.048**	.044
RM ANOVA H-F	.054	NA	.054	.054	.058	**.052**	**.052**	.054
MANOVA - Wilks' Λ	**.049**	NA	**.050**	**.048**	**.051**	**.048**	**.048**	**.049**

* 95% confidence interval = (.048, .052), except for sample size conditions: (.047, .053)

Table A27

Empirical Type I Error Rates for the Compound Symmetry True Model: Test of the Interaction

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.012	.003	.010	.005	.009	.008	.005	.007
Variance Components	.013	.003	.010	.006	.011	.008	.006	.008
Compound Symmetry	**.050**	**.050**	**.051**	**.049**	**.049**	**.052**	**.048**	**.050**
Heterogeneous Compound Symmetry	**.048**	.046	**.048**	.046	.042	**.050**	**.049**	.047
Autoregressive	.038	.040	.043	.035	.038	.040	.040	.039
Heterogeneous Autoregressive	.039	.041	.044	.036	.039	.041	.040	.040
Toeplitz	**.051**	**.052**	**.051**	**.052**	**.052**	**.052**	**.049**	**.051**
Heterogeneous Toeplitz	.047	**.048**	**.048**	.047	.043	**.051**	**.049**	.047
Unstructured	**.050**	**.049**	**.050**	**.050**	**.049**	**.051**	**.049**	**.050**
Mean	.039	.037	.039	.036	.037	.039	.037	.038
RM ANOVA Conventional F-test	**.050**	**.050**	**.051**	**.049**	**.049**	**.052**	**.048**	**.050**
RM ANOVA G-G	.040	.040	.045	.034	.030	.045	.044	.040
RM ANOVA H-F	**.049**	**.049**	**.050**	**.048**	**.048**	**.051**	**.048**	**.049**
MANOVA - Wilks' Λ	**.050**	**.049**	**.050**	**.050**	**.049**	**.051**	**.049**	**.050**

* 95% confidence interval = (.048, .052), except for sample size conditions: (.047, .053)

Table A28

Empirical Type I Error Rates for the Heterogeneous Compound Symmetry True Model: Test of the Interaction

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.023	.009	.021	.011	.020	.016	.013	.016
Variance Components	.017	.008	.016	.008	.016	.011	.010	.012
Compound Symmetry	.060	.057	.056	.060	.059	.058	.058	.058
Heterogeneous Compound Symmetry	**.048**	.045	.047	.046	.043	**.048**	**.047**	.046
Autoregressive	**.049**	.047	**.049**	.047	**.048**	**.049**	**.048**	**.048**
Heterogeneous Autoregressive	.043	.043	.045	.041	.043	.044	.042	.043
Toeplitz	.057	.055	.054	.058	.057	.056	.056	.056
Heterogeneous Toeplitz	.047	.045	.047	.045	.044	**.047**	.046	.046
Unstructured	**.050**	**.048**	**.049**	**.049**	**.050**	**.050**	**.047**	**.049**
Mean	.044	.040	.043	.041	.042	.042	.041	.042
RM ANOVA Conventional F-test	.060	.057	.056	.060	.059	.058	.058	.058
RM ANOVA G-G	.043	.041	.046	.037	.035	.044	**.047**	.042
RM ANOVA H-F	.054	**.051**	**.052**	.053	.054	**.052**	**.051**	**.052**
MANOVA - Wilks' Λ	**.050**	**.048**	**.049**	**.049**	**.050**	**.050**	**.047**	**.049**

* 95% confidence interval = (.048, .052), except for sample size conditions: (.047, .053)

Table A29

Empirical Type I Error Rates for the Heterogeneous Autoregressive True Model: Test of the Interaction

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.038	.025	.026	.037	.036	.029	.029	.031
Variance Components	.031	.022	.022	.031	.029	.026	.025	.027
Compound Symmetry	.063	.068	.060	.071	.068	.064	.064	.065
Heterogeneous Compound Symmetry	**.050**	.054	**.048**	.056	**.051**	**.052**	**.053**	**.052**
Autoregressive	.058	.057	.054	.061	.056	.057	.059	.057
Heterogeneous Autoregressive	**.048**	**.049**	.047	**.050**	**.047**	**.048**	**.050**	**.048**
Toeplitz	.058	.059	.054	.062	.059	.056	.058	.058
Heterogeneous Toeplitz	**.049**	**.049**	.047	**.051**	**.048**	**.050**	**.050**	**.049**
Unstructured	**.050**	**.051**	**.050**	**.052**	**.052**	**.048**	**.051**	**.051**
Mean	**.049**	**.048**	.045	**.052**	**.050**	**.048**	**.049**	**.049**
RM ANOVA Conventional F-test	.063	.068	.060	.071	.068	.064	.064	.065
RM ANOVA G-G	.044	.045	.047	.042	.040	**.047**	**.047**	.044
RM ANOVA H-F	.055	.055	.053	.057	.061	.054	**.050**	.055
MANOVA - Wilks' Λ	**.050**	**.051**	**.050**	**.052**	**.052**	**.048**	**.051**	**.051**

* 95% confidence interval = (.048, .052), except for sample size conditions: (.047, .053)

Table A30

Empirical Type I Error Rates for the Heterogeneous Toeplitz True Model: Test of the Interaction

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.030	.018	.024	.025	.029	.023	.021	.024
Variance Components	.026	.015	.020	.022	.023	.020	.019	.021
Compound Symmetry	.061	.063	.059	.065	.065	.061	.060	.062
Heterogeneous Compound Symmetry	**.050**	.047	**.048**	**.049**	**.047**	**.049**	**.050**	**.049**
Autoregressive	.054	**.052**	.053	.054	**.053**	**.053**	**.053**	.053
Heterogeneous Autoregressive	.046	.047	**.048**	.045	.045	**.047**	**.047**	.047
Toeplitz	.058	.054	.054	.058	.056	.056	.055	.056
Heterogeneous Toeplitz	**.050**	.047	**.049**	**.048**	.046	**.050**	**.050**	**.049**
Unstructured	.053	**.050**	**.051**	**.051**	**.051**	**.052**	**.051**	**.051**
Mean	**.048**	.044	.045	.046	.046	.046	.045	.046
RM ANOVA Conventional F-test	.061	.063	.059	.065	.065	.061	.060	.062
RM ANOVA G-G	.043	.043	.047	.039	.039	.045	.046	.043
RM ANOVA H-F	.053	.053	**.052**	.054	.059	**.051**	**.049**	.053
MANOVA - Wilks' Λ	.053	**.050**	**.051**	**.051**	**.051**	**.052**	**.051**	**.051**

* 95% confidence interval = (.048, .052), except for sample size conditions: (.047, .053)

Table A31

Empirical Type I Error Rates for the Unstructured True Model: Test of the Interaction

Fitted Model	Low Correlation Marginal	High Correlation Marginal	Time Points = 3 Marginal	Time Points = 6 Marginal	Small Sample Marginal	Moderate Sample Marginal	Large Sample Marginal	Marginal
Independence	.028	.013	.026	.015	.024	.019	.018	.020
Variance Components	.019	.007	.015	.010	.016	.012	.010	.013
Compound Symmetry	.060	.067	.062	.065	.063	.063	.064	.064
Heterogeneous Compound Symmetry	**.050**	.046	**.048**	**.048**	.046	**.050**	**.049**	**.048**
Autoregressive	**.048**	.053	**.050**	**.050**	**.050**	**.050**	**.051**	**.050**
Heterogeneous Autoregressive	.041	.046	.043	.044	.044	.043	.043	.043
Toeplitz	.058	.061	.058	.061	.057	.060	.062	.059
Heterogeneous Toeplitz	.046	.044	.043	**.048**	.042	**.047**	**.047**	.045
Unstructured	**.051**	**.049**	**.050**	**.050**	**.051**	**.049**	**.050**	**.050**
Mean	.045	.043	.044	.043	.044	.044	.044	.044
RM ANOVA Conventional F-test	.060	.067	.062	.065	.063	.063	.064	.064
RM ANOVA G-G	.042	.045	.047	.040	.037	.046	**.048**	.043
RM ANOVA H-F	**.051**	.056	.053	.054	.057	**.052**	**.052**	.054
MANOVA - Wilks' Λ	**.051**	**.049**	**.050**	**.050**	**.051**	**.049**	**.050**	**.050**

* 95% confidence interval = (.048, .052), except for sample size conditions: (.047, .053)