REPEATED MEASURES MULTIPLE COMPARISON PROCEDURES

WITH A MIXED MODEL ANALYSIS

Rhonda K. Kowalchuk

A Thesis
Submitted to the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

University of Manitoba
Winnipeg, MB

0-612-51642-3

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION PAGE

Repeated Measures Multiple Comparison Procedures With a Mixed Model Analysis

BY

Rhonda K. Kowalchuk

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Table of Contents

Acknowledgments

To my advisor and mentor, Dr. H. J. Keselman, for his hard work, time, and effort throughout the completion of this thesis and other projects undertaken during my doctoral program.

To my committee members, Dr. J. Keselman, Dr. M. Aftanas, and Dr. K. Mount, for their participation and many helpful comments and suggestions.

To my external examiner, Dr. S. Maxwell, for his insightful evaluation.

To my dear friends, for their encouragement to keep my work and life in perspective.

Finally, to my family, especially my parents, for their enduring support.

Abstract

One approach to the analysis of repeated measures designs allows researchers to model the variance-covariance structure of their data rather than presume a certain structure as is the case with conventional univariate and multivariate test statistics (Littell, Milliken, Stroup, & Wolfinger, 1996). This mixed-model approach was evaluated for testing all possible pairwise differences among repeated measures marginal means in a between- by within-subjects design. Specifically, Type I error control and power were examined for simultaneous and stepwise multiple comparison procedures using SAS' (1996) PROC MIXED in an unbalanced repeated measures design when normality and variance-covariance homogeneity assumptions did not hold. The potential advantage of the MIXED procedure with its ability to specify various variance-covariance structures was compared to known robust multiple comparison procedures based on a between-subjects heterogeneous unstructured form of the variance-covariance matrix with Satterthwaite (1941, 1946) degrees of freedom (Keselman, 1994; Keselman, Keselman, & Shaffer, 1991; Keselman & Lix, 1995). Specifically, the testing strategies of always fitting an unstructured variance-covariance matrix, fitting the true population structure, or allowing two model selection criteria available through PROC MIXED to select the best structure were investigated. Rates of Type I error control were similar across the testing strategies for each of the multiple comparison procedures. The recommendation of always fitting an unstructured variance-covariance matrix to the data was based on the fact that a researcher does not need prior knowledge about the true population structure and does not need to rely on a model selection criterion to provide good Type I error control.

Furthermore, results showed two stepwise multiple comparison procedures as particularly notable. Shaffer's (1986) sequentially rejective Bonferroni and Hochberg's (1988) sequentially acceptive Bonferroni procedures had superior performance with regards to Type I error control and power to detect true pairwise differences across the investigated conditions.

Repeated Measures Multiple Comparison Procedures with a Mixed Model Analysis

A common experimental design in psychological and educational research is the repeated measures (RM) design in which the same measurement is taken on a unit of analysis (e.g., subject) on more than one occasion. Other names for this type of design include within-subjects and correlated groups designs. A reason for the popularity of this design is that in many situations repeated measurements on the same subject occur naturally. For example, a developmental study that examines a child's motor development across age levels or a learning study where the same individual is measured across various treatment conditions that may represent different drug dosage levels. Because the same subjects are measured repeatedly the measurements are correlated. Therefore, the pattern of variances and covariances (or correlations) among the levels of the repeated factor require special consideration when analysing data from such designs. There are two main advantages of RM designs that also contribute to their popularity in the literature (Maxwell & Delaney, 1990). First, the units of analysis act as their own control, thereby eliminating individual differences between subjects from the error variability and thus creating a more sensitive design for testing effects. Second, a RM design requires fewer subjects to obtain the same level of power as in a between-subjects design (also known as an independent groups design). This is appealing for a researcher who may be under constraints of time and money when conducting a study.

A design that contains a single repeated factor is called a simple RM design. The inclusion of an additional repeated factor (or factors) in which the units of analysis are measured under each combination of the repeated factors is called a single group factorial

RM design or higher order within-subjects design. When a grouping factor (or factors) is added to a RM design it is called a between- by within-subjects design or mixed design. In a mixed design, the units of analysis are classified into independent groups and measured under all levels (or combinations) of a repeated factor (or factors). A design that contains one between-subjects and one within-subjects factor is the simplest example of a mixed design, sometimes referred to as a split-plot design.

The discussion of RM designs in the psychological literature dates back to the 1940s (see Lovie, 1981). The univariate analysis of variance (ANOVA) approach to data from these designs was the focus of early research. Not surprisingly the univariate approach remains the most common analysis method for RM designs in the psychological and educational literature (Keselman et al., 1998; Kowalchuk, Lix, & Keselman, 1996). However, data encountered by behavioral researchers is unlikely to satisfy the strict assumptions required for valid univariate F-tests. Recommendations in the literature about the "best" analysis strategy to adopt for RM designs can be found in numerous articles and book chapters (e.g., Barcikowski & Robey, 1984; Everitt, 1995; Keselman & Algina, 1996; Keselman & Keselman, 1993; Lewis, 1993; Looney & Stanley, 1989; Maxwell & Delaney, 1990; McCall & Appelbaum, 1973; O'Brien & Kaiser, 1985). These recommendations are typically based on simulation or Monte Carlo studies examining the "robustness" of various analysis methods. The term robustness refers to a statistical test's insensitivity with regard to Type I error control under violations of its assumptions (Box, 1954). The predominant method of evaluating robustness is the use of simulation techniques, in contrast to theoretical explications.

The majority of research on RM designs has focused on the analysis of omnibus tests of RM factors (i.e., within-subjects main and within-subjects interaction effects). Although omnibus tests of these factors are informative, a researcher usually has hypotheses of interest that require the use of more specific contrast (or comparison) tests (e.g., marginal mean comparisons). Assuming the effect of interest is a RM main effect (with more than two levels) a researcher has the option of testing specific a priori contrast(s) or choosing a multiple comparison procedure (MCP) to test all possible pairwise contrasts. The latter is the focus of the present study.

To develop the background for the present study, an overview of omnibus analysis approaches to RM designs will be presented. Typically the data from these designs are analysed by conventional univariate or multivariate methods. In addition, there are degrees of freedom (df) adjusted procedures that may be used when the assumptions of the conventional univariate approach is not tenable. Another approach to the analysis of RM designs recommended in the literature, the mixed model approach, allows one to model the variance-covariance structure of the data rather than presume a certain structure as is the case with conventional univariate and multivariate test statistics (see Littell, Milliken, Stroup, & Wolfinger, 1996; SAS Institute, 1999; Wolfinger, 1993, 1996). The purpose of the present study was to evaluate this mixed model approach for testing all possible pairwise differences among RM marginal means in a between- by within-subjects design. Accordingly, a Monte Carlo study was done to examine the robustness of several MCPs using SAS' (1999) PROC MIXED.

The decision to investigate a mixed RM design is based on the popularity of this design in the literature. A methodological content analysis of 13 educational and psychological journals published in 1994 found 84% of the articles that used a RM design analysed a mixed design (i.e., a design that contains both between-subjects and within-subjects factors; Keselman et al., 1998; Kowalchuk et al., 1996). Furthermore, the authors found that unbalanced designs (i.e., unequal numbers of subjects in each group/cell) were more common than balanced designs. The journals reviewed are considered representative of the education and psychology disciplines and are considered prominent by researchers in the respective areas and thus provide a good indication of designs likely to be encountered by applied researchers.

Mixed Repeated Measures Design

Omnibus Tests

The simplest of the higher-order mixed RM design contains one between-subjects and one within-subjects factor in which subjects ($i = 1, \ldots, n_j$, $\Sigma_j n_j = N$) are randomly selected for each level of the between-subjects factor ($j = 1, \ldots, J$) and observed and measured under all levels of the within-subjects factor ($k = 1, \ldots, K$).

The general linear model for RM data is (see Timm, 1975)

$$\mathbf{Y} = \mathbf{XB} + \xi \, , \qquad (1)$$

where $\mathbf{Y}$ is an N x K matrix of scores on K repeated measurements, N is the total sample size, $\mathbf{X}$ is an N x J design matrix that codes for between-subjects effects (rank($\mathbf{X}$) = J), $\mathbf{B}$ is a J x K matrix of nonrandom parameters (i.e., population means), and ξ is an N x K matrix of random error components. The rows of ξ are assumed to independently and identically distributed as $N_k(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a K x K variance-covariance matrix.

The unknown elements σ_{ij} of the matrix $\mathbf{\Sigma}$ are estimated by

$$\mathbf{S} = \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\mathbf{b}\,, \tag{2}$$

where ^{T} refers to the transpose operator. The df due to error is N - J = v_e , and $(1/v_e)\mathbf{S}$ is an unbiased estimator of $\mathbf{\Sigma}$.

The general linear null hypothesis can be written as

$$H_0\text{: } \mathbf{DBU} = \mathbf{0}\,. \tag{3}$$

where $\mathbf{D}$ is a df_D x J contrast matrix on the between-subjects effect, with rank($\mathbf{D}$) = $df_D \le$ J and $\mathbf{U}$ is a K x df_U contrast matrix on the within-subjects effect, with rank($\mathbf{U}$) = $df_U \le$ K. The sum of squares and cross products matrix due to the hypothesis is computed as

$$\mathbf{S}_h = (\mathbf{DbU} - \mathbf{0})^T\ (\mathbf{D}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{D}^T)^{-1}\ (\mathbf{DbU} - \mathbf{0})\,. \tag{4}$$

and the sum of squares and cross products matrix due to error is computed as

$$\mathbf{S}_e = \mathbf{U}^T\mathbf{Y}^T\left[\mathbf{I}_N - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right]\mathbf{YU}\,. \tag{5}$$

where $\mathbf{I}_N$ is an identity matrix of dimension N. A test of the null hypothesis is made by comparing the matrices $\mathbf{S}_e$ and $\mathbf{S}_h$.

Multivariate Approach. Several multivariate (MANOVA) test statistics can be used to test an omnibus null hypothesis in a RM design. The most common include Hotelling's (1931) T^2 statistic, Pillai (1955)- Bartlett (1939) trace statistic, Wilks' (1932) likelihood ratio, Hotelling (1951)- Lawley (1938) trace criterion, and Roy's (1953) largest root criterion. When the minimum of (df_D, df_U) is equal to one, all criteria are equivalent to Hotelling's (1931) T^2 statistic.

Hotelling's (1931) T^2 statistic is defined by $T^2 = v_e\,[\mathrm{tr}\,(\mathbf{S}_h\,\mathbf{S}_e^{-1})]$, where tr refers to the trace operator, Pillai (1955)- Bartlett (1939) trace (PB) statistic is given by $PB = \mathrm{tr}(\mathbf{S}_h\,\mathbf{T}^{-1})$ where $\mathbf{T} = \mathbf{S}_h + \mathbf{S}_e$, Wilks' (1932) likelihood ratio (W) is defined by $W = \det\,(\mathbf{S}_e\,\mathbf{T}^{-1})$, where det refers to the determinant of a matrix, and the Hotelling (1951)- Lawley (1938) trace (HLT) criterion is defined by $HLT = \mathrm{tr}\,(\mathbf{S}_h\,\mathbf{S}_e^{-1})$. Each of these statistics can be expressed as an F-variate. For example, a test of a RM main effect based on Hotelling's (1931) T^2 can be expressed as

$$F = \frac{N - J - K + 2}{(N - J)\,(K - 1)}\,T^2 \approx F\left[\alpha;\,(K - 1),\,(N - J - K + 2)\right]. \qquad (6)$$

The F-approximations for PB, W, and HLT can be found in Muller, LaVange, Ramey, and Ramey (1992).

The multivariate test statistics are based on an Unstructured (UN) form of the variance-covariance matrix, where (K(K+1))/2 parameters must be estimated [i.e., K variances and (K(K-1))/2 covariances]. The UN variance-covariance matrix has the following form (assuming K = 4)

$$UN = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ & & \sigma_3^2 & \sigma_{34} \\ & & & \sigma_4^2 \end{bmatrix}. \qquad (7)$$

Valid use of multivariate test statistics do not place any restrictions on the form of the variance-covariance matrix but do require the structure of the matrix to be constant across the levels of the grouping variable, in addition to the assumptions of normality and independence of observations. According to Olson (1974), the Pillai-Bartlett trace criterion is the most robust of the MANOVA test statistics.

Univariate Approach. The univariate approach to the analysis of RM designs is a special case of the more general multivariate analysis. However, the matrix **U** must be orthonormal such that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{(K-1)}$, where **I** is an identity matrix of dimension K - 1. A test of the null hypothesis is given by

$$F = \left[SSH / df_h\right] / \left[SSE / df_e\right] \approx F\left[\alpha;\ df_h,\ df_e\right]. \qquad (8)$$

where $df_h = df_D \operatorname{rank}(\mathbf{U})$, $df_e = \nu_e \operatorname{rank}(\mathbf{U})$, SSH = trace($\mathbf{S}_h$), and SSE = trace($\mathbf{S}_e$).

Unlike the multivariate tests, the validity of the univariate *F*-tests (main and interaction effects) is dependent on a particular form of the variance-covariance matrix. The assumption of equal population variances and equal population covariances defines a specific type of variance-covariance structure known as Compound Symmetry (CS) or uniformity and requires two parameters to be estimated, a homogeneous variance (σ^2) and a constant correlation (ρ). The CS variance-covariance matrix, when K = 4, has the following form

$$CS = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ & 1 & \rho & \rho \\ & & 1 & \rho \\ & & & 1 \end{bmatrix}. \qquad (9)$$

Compound symmetry is a sufficient condition for valid univariate F-tests but it is not a necessary condition. A less restrictive condition for valid F-tests is that the variances of all paired differences among the levels of the RM variable are equal (Huynh & Feldt, 1970). This can be expressed as

$$\sigma^2_{k-k'} = \sigma^2_k + \sigma^2_{k'} - 2\sigma_k\sigma_{k'} \qquad (\text{for all } k \neq k')\,. \tag{10}$$

A variance-covariance matrix that satisfies this less restrictive assumption is said to possess a spherical pattern and is known as a Huynh-Feldt (HF) structure which has the following form (when K = 4)

$$\mathrm{HF} = \begin{bmatrix} \sigma_1^2 & \frac{\sigma_1^2+\sigma_2^2}{2}-\lambda & \frac{\sigma_1^2+\sigma_3^2}{2}-\lambda & \frac{\sigma_1^2+\sigma_4^2}{2}-\lambda \\ & \sigma_2^2 & \frac{\sigma_2^2+\sigma_3^2}{2}-\lambda & \frac{\sigma_2^2+\sigma_4^2}{2}-\lambda \\ & & \sigma_3^2 & \frac{\sigma_3^2+\sigma_4^2}{2}-\lambda \\ & & & \sigma_4^2 \end{bmatrix}, \tag{11}$$

where λ is a scalar value greater than zero. Thus, the necessary and sufficient condition for valid univariate F-tests is called the sphericity assumption. When a RM variable contains two levels, only one paired difference exists and the sphericity assumption is said to be trivially satisfied.

The sphericity condition can be expressed in matrix notation as

$$\mathbf{U}^{T}\Sigma\,\mathbf{U} = \lambda\,\mathbf{I}_{(K-1)} \tag{12}$$

(Rouanet & Lepine 1970). This implies that the variances of the set of K-1 contrasts on the repeated variable represented by the matrix **U** are constant. When the design contains a between-subjects variable, an additional assumption is required that can be expressed as

$$\mathbf{U}^T\Sigma_1\mathbf{U} = \mathbf{U}^T\Sigma_2\mathbf{U} = \cdots = \mathbf{U}^T\Sigma_J\mathbf{U} = \lambda\,\mathbf{I}_{(K-1)}\,, \tag{13}$$

which implies a constant variance across the levels of the grouping factor. These assumptions have been jointly referred to as multisample sphericity (Huynh, 1978). The univariate F-tests are valid if and only if the assumption of multisample sphericity is satisfied in addition to the normality and independence assumptions.

The literature uniformly agrees that the conventional univariate F-test should not be used due to its sensitivity to assumption violations. Because the data from educational and psychological research is unlikely to satisfy the strict assumptions required for valid univariate F-tests, this analysis approach is not recommended. Furthermore, the use of preliminary tests of the multisample sphericity assumption such as Box's modified criterion (see Huynh & Feldt, 1970) to test for heteroscedasticity of covariance matrices and Mauchly's (1940) sphericity criterion to test for departures from sphericity are not recommended (Keselman, Rogan, Mendoza, & Breen, 1980; Rogan, Keselman, & Mendoza, 1979). Both tests are extremely sensitive to nonnormality and even under conditions of normality the tests are sensitive to small departures from their respective hypotheses.

Box (1954) showed that the univariate F-tests are approximately distributed as F-variates when the df are adjusted by a factor representing the departure from sphericity. The degree of sphericity in a population variance-covariance matrix is measured by the parameter epsilon, notated as ε , where

$$\varepsilon = \frac{\left[\mathrm{tr}\,(\mathbf{U}^{T}\Sigma\mathbf{U})\right]^{2}}{(K-1)\,\mathrm{tr}\left[(\mathbf{U}^{T}\Sigma\mathbf{U})^{2}\right]}\,, \tag{14}$$

and

$$\Sigma = \frac{\sum_{j=1}^{J}(n_j - 1)\,\Sigma_j}{(N-J)}\,, \tag{15}$$

is the pooled population variance-covariance matrix. The parameter ε has an upper bound of one when sphericity is satisfied and a lower bound of $(K-1)^{-1}$ for a J x K design. Several df adjusted univariate analysis procedures have been proposed to correct for violation of the sphericity assumption in a RM analysis.

Adjusted Degrees of Freedom Univariate Approaches. Given the lower bound of epsilon, Geisser and Greenhouse (1958) presented a lower bound correction for univariate F-tests where ε is set equal to $(K-1)^{-1}$. When the df for effects involving a repeated factor in a J x K design are multiplied by this correction factor, a test of the within-subjects main effect is based on 1 and (N - J) df and a test of the within-subjects interaction effect is based on (J - 1), and (J - 1)(N - J) df. This method is conservative (i.e., smaller df correspond to a larger critical F-value; Maxwell & Delaney, 1990) and may lack sufficient power to reject the null hypothesis. Hence, this method is not recommended (Rogan et al., 1979).

Greenhouse and Geisser (1959) suggested a sample estimate $\hat{\varepsilon}$ for Box's (1954) ε. The conventional univariate df are multiplied by the adjustment factor $\hat{\varepsilon}$. For tests of within-subjects main and within-subjects interaction effects in a J x K design, the Greenhouse and Geisser (1959) $\hat{\varepsilon}$ adjusted F-tests are respectively,

$$F_K \approx F\left[\alpha;\ (K-1)\hat{\varepsilon},\ (N-J)(K-1)\hat{\varepsilon}\right], \tag{16}$$

and

$$F_{J\times K} \approx F\left[\alpha;\ (J-1)(K-1)\hat{\varepsilon},\ (N-J)(K-1)\hat{\varepsilon}\right], \tag{17}$$

where

$$\hat{\varepsilon} = \frac{\left[\mathrm{tr}\,(\mathbf{U}^T\mathbf{S}\mathbf{U})\right]^2}{(K-1)\,\mathrm{tr}\left[(\mathbf{U}^T\mathbf{S}\mathbf{U})^2\right]}, \tag{18}$$

and

$$\mathbf{S} = \frac{\sum_{j=1}^{J}(n_j-1)\,\mathbf{S}_j}{(N-J)}, \tag{19}$$

where $\mathbf{S}_j$ is the sample variance-covariance matrix for the j^{th} group and $\mathbf{S}$ is the pooled sample variance-covariance matrix. The sample estimate of ε was found to be biased when ε was greater than or equal to .75 especially with small sample sizes (Collier, Baker, Mandeville, & Hayes, 1967).

Huynh and Feldt (1976) proposed an $\tilde{\varepsilon}$ adjustment approach to correct for the conservative nature of the $\hat{\varepsilon}$ adjustment method of Greenhouse and Geisser (1959). Tests of RM main and interaction effects in a J x K design are respectively,

$$F_K \approx F\left[\alpha;\ (K-1)\,\tilde{\varepsilon}\,,\ (N-J)(K-1)\,\tilde{\varepsilon}\right], \tag{20}$$

and

$$F_{J\times K} \approx F\left[\alpha;\ (J-1)(K-1)\,\tilde{\varepsilon},\ (N-J)(K-1)\,\tilde{\varepsilon}\right], \tag{21}$$

where

$$\tilde{\varepsilon} = \frac{N\,(K-1)\,\hat{\varepsilon}\,-\,2}{(K-1)\left[N-J-(K-1)\,\hat{\varepsilon}\right]}. \tag{22}$$

Although $\tilde{\varepsilon}$ can exceed a value of one its maximum value is restricted to one. Lecoutre (1991) offered a correction $\tilde{\varepsilon}_c$ to the $\tilde{\varepsilon}$ adjusted procedure when the number of groups is greater than or equal to two. Specifically, (N - J + 1) is substituted for N in the numerator of $\tilde{\varepsilon}$. Chen and Dunlap (1994) found $\tilde{\varepsilon}_c$ to be less biased than $\tilde{\varepsilon}$ when ε was greater than or equal to .75 (i.e., rates of Type I error were closer to the nominal level). This correction is important because a review of education and psychology publications found that ε rarely fell below .75 (Huynh & Feldt, 1976).

Additional ε adjusted procedures that essentially represent a combination of $\tilde{\varepsilon}_c$ and $\hat{\varepsilon}$ have been proposed. Quintana and Maxwell (1994) investigated seven ε adjusted approaches to the analysis of RM designs in terms of Type I error and power under violation of the sphericity assumption. The authors recommended two adjustment approaches: the $\tilde{\varepsilon}_c$ adjusted procedure and a method combining $\tilde{\varepsilon}_c$ and $\hat{\varepsilon}$ in which $\tilde{\varepsilon}_c$ is calculated and if it is greater than or equal to .75 then the $\tilde{\varepsilon}_c$ adjustment is used, otherwise the $\hat{\varepsilon}$ adjustment is used.

In general, the adjusted univariate methods are robust to heterogeneity of covariance matrices given equal sample sizes (Huynh, 1978) and are more robust than multivariate methods under nonnormality (Keselman, Keselman, & Lix, 1995; Rogan et al., 1979). However, when heterogeneous covariance matrices are combined with unequal group sizes, the univariate adjusted tests and the multivariate tests are generally not robust (Keselman & Keselman, 1990; Keselman et al., 1995; Olson, 1974). Specifically, when covariance matrices and group sizes are negatively paired (i.e., the covariance matrix with the largest element values is paired with the smallest group size), Type I error rates become liberal and when covariance matrices and group sizes are positively paired (i.e., the covariance matrix with the largest element values is paired with the largest group size), Type I error rates become conservative. For a review of the empirical literature, see Keselman, Lix, and Keselman (1996) who conducted a meta-analysis summarizing Type I error and power results of Monte Carlo studies on split-plot RM designs investigating univariate and multivariate approaches.

Some authors have recommended a test strategy that combines the adjusted univariate and multivariate tests (Barcikowski & Robey, 1984; Looney & Stanley, 1989).

For example, both univariate and multivariate tests are evaluated at $\alpha/2$ and if either test is significant then the hypothesis is rejected. Keselman et al. (1995) compared this testing strategy to uniformly adopting either an adjusted univariate or multivariate test. Their results do not favor the use of a combined testing strategy because the combined strategy is sensitive to the same conditions (i.e., unequal covariance matrices combined with unequal group sizes) that cause the adjusted univariate and multivariate tests to lack robustness.

Huynh (1978) proposed two approaches to deal with the case of unequal covariance matrices and arbitrary group sizes in a between- by within-subjects RM design; the General Approximation (GA) test and the Improved General Approximation (IGA) test. The RM main and interaction F-tests for the GA test are distributed respectively as,

$$F_K \approx b\, F\left[h',\, h\right], \tag{23}$$

and

$$F_{J \times K} \approx c\, F\left[h'',\, h\right], \tag{24}$$

where b, c, h, h' and h" are unknown constants. To correct for underestimation, "improved" estimates for h, h', and h" are given by $\tilde{h}$, $\tilde{h}'$, and $\tilde{h}''$. The formulae can be found in Huynh (1978). Algina (1994) presented a Lecoutre (1991) correction (CIGA) to the IGA test by replacing N in the numerators of $\tilde{h}'$ and $\tilde{h}''$ by (N - J + 1). Huynh (1978) found the $\tilde{\varepsilon}$ and IGA adjusted tests performed reasonably well in terms of Type I error control compared to the $\hat{\varepsilon}$ and GA adjusted tests. Because of the complexity of the IGA test Huynh recommended the simpler $\tilde{\varepsilon}$ approximate procedure. However, the $\tilde{\varepsilon}$ procedure is not robust to violation of the multisample sphericity assumption when group sizes are unequal.

Algina and Oshima (1994) found both the IGA and CIGA tests adequately controlled Type I error under violation of multisample sphericity but the combined effect of covariance heterogeneity and nonnormality resulted in conservative rates of error for the test of the within-subjects interaction. Algina and Oshima (1995) recommended the CIGA test for tests of the within-subjects main effect (unweighted hypothesis) when the design is unbalanced and covariance matrices are unequal. However, when the design is balanced the $\tilde{\varepsilon}_c$ adjusted procedure provides good control of Type I error. Algina (1997) extended the CIGA test for RM designs containing multiple between-subjects and multiple within-subjects factors and provided a computer program to compute the CIGA test. Thus, the complexity of the procedure is no longer an issue preventing researchers from adopting this method of analysis.

Approximate Degrees of Freedom Multivariate Approach. In addition to the univariate solution provided by Huynh (1978), an approximate df multivariate solution for unequal covariance matrices (i.e., allows $\mathbf{\Sigma}_j \neq \mathbf{\Sigma}_{j'}$, where $j \neq j'$) is based on a Welch (1947, 1951)-James (1951, 1954) (WJ) type statistic according to Johansen (1980) and Keselman, Carriere, and Lix (1993).

The general linear null hypothesis is expressed as

$$H_0: \mathbf{C}\mu = \mathbf{0}\,, \tag{25}$$

where $\mathbf{C} = \mathbf{D} \otimes \mathbf{U}^T$, where $\mathbf{D}$ and $\mathbf{U}$ have been defined previously and $\otimes$ is the Kronecker or direct product function, $\mathbf{C}$ is a contrast matrix with $df_D \times df_U$ rows and J x K columns, $\boldsymbol{\mu}$ is a J x K column vector obtained by vertically stacking the rows of $\mathbf{B}$. That is, $\boldsymbol{\mu} = (\mathbf{B}_1^T, \ldots, \mathbf{B}_J^T)$, where $\mathbf{B}_j = (\mu_{j1}, \ldots, \mu_{jK})^T$.

The test statistic is defined as

$$T_{WJ} = (\mathbf{C}\overline{\mathbf{Y}})^T\,(\mathbf{CSC}^T)^{-1}\,(\mathbf{C}\overline{\mathbf{Y}})\,. \tag{26}$$

where $\overline{\mathbf{Y}}$ estimates $\boldsymbol{\mu}$ and $\mathbf{S} = \text{diag}(\mathbf{S}_1/n_1, \ldots, \mathbf{S}_J/n_J)$, a block diagonal matrix with elements $\mathbf{S}_j/n_j$. This statistic divided by a constant c, is approximately distributed as an F-statistic with $v_1 = (df_D \times df_U)$ and $v_2 = (v_1(v_1 + 2))/(3A)$ df, where $c = v_1 + 2A - (6A)/(v_1 + 2)$ and A is given by

$$\frac{1}{2}\sum_{j=1}^{J}\left[\text{tr}\left\{\mathbf{SC}^T(\mathbf{CSC}^T)^{-1}\mathbf{CQ}_j\right\}^2 + \left\{\text{tr}\,(\mathbf{SC}^T(\mathbf{CSC}^T)^{-1}\mathbf{CQ}_j)\right\}^2\right]\Big/(n_j - 1)\,. \tag{27}$$

where $\mathbf{Q}_j$ is a block diagonal matrix of dimension JK x JK such that the (s,t)th diagonal block of $\mathbf{Q}_j = \mathbf{I}_K$ if s=t=j and is $\mathbf{0}$ otherwise.

Keselman et al. (1993) found the WJ approach provided reasonable control of Type I error given certain sample size requirements. Specifically, to test the within-subjects main effect the smallest group size should be 2 to 3 times larger than the number of RM minus one and if the data are likely to violate the normality assumption, the ratio increases to 3 or 4 to one. To test a within-subjects interaction effect the smallest group size should be 3 or 4 times larger than the number of RM minus one, while this ratio increases to 5 or 6 to one if the assumption of normality is unlikely to be satisfied. Provided these sample size requirements are adhered to, the WJ test is superior with regard to power compared to the univariate adjusted and multivariate tests when both covariance matrices and group sizes are unequal (Keselman et al., 1995). Lix and Keselman (1995) provide a SAS/IML (SAS Institute, 1989) program that can be used to compute the WJ test for any RM design.

An investigation (Algina & Keselman, 1997) of the generalizability of the sample size requirements given by Keselman et al. (1993) for the WJ test include the following modifications: (a) for a test of the RM main effect, the sample size requirements can be reduced as the number of levels of the grouping factor increase, and (b) for a test of the RM interaction effect, the sample size requirements should be increased as the number of levels of the grouping factor increase. When the sample size requirements of the WJ test cannot be obtained an alternative is the CIGA test. However, the WJ test is more powerful than the CIGA test when sample sizes allow adequate control of Type I error for the WJ test (Algina & Keselman, 1998).

Empirical Bayes Approach. Boik (1997) proposed a hybrid analysis for a RM design based on the univariate and multivariate approaches that uses a two stage model. In contrast to the general linear model for RM data, the first stage model of the present approach is

$$\mathbf{YU} = \mathbf{X\Theta} + \mathbf{E}, \tag{28}$$

where $\mathbf{\Theta} = \mathbf{BU}$ and $\mathbf{E} = \xi\mathbf{U}$. The rows of $\mathbf{E}$ are assumed to be independently and identically distributed as N_{K-1} ($\mathbf{0}$, $\mathbf{\Phi}$), where $\mathbf{\Phi} = \mathbf{U}^T\mathbf{\Sigma}\mathbf{U}$. The second stage model assumes prior distributions on $\mathbf{\Theta}$ and $\mathbf{\Phi}$. Specifically, $\mathbf{\Theta}$ and $\mathbf{\Phi}$ are assumed to be independently distributed and $\mathbf{\Theta}$ is uniformly distributed over a J(K-1) dimensional space and $\mathbf{\Phi}$ follows a spherical inverted Wishart distribution

$$\mathbf{\Phi}^{-1} \approx \mathbf{W}_{K-1}(f, \tau^{-1}\mathbf{I}). \tag{29}$$

This implies that

$$E(\mathbf{\Phi}) = \sigma^2 \mathbf{I}_{K-1}, \tag{30}$$

where

$$\sigma^2 = \frac{\tau}{f - (K-1) - 1}. \tag{31}$$

This is referred to as second stage sphericity, that is sphericity is satisfied on average but not necessarily for any given covariance matrix. To quantify the prior belief in sphericity the hyperparameter f is computed ($(K-1)-1 < f < \infty$) such that "Small values of f correspond to a belief that departure from sphericity will be large, whereas large values of f correspond to the belief that departure from sphericity will be small." (Boik, 1997 p. 160). The conventional multivariate test statistics can be used to test hypotheses. The hypothesis matrix is the same as the conventional multivariate approach (i.e., $\mathbf{S}_h$) and the error matrix is given by

$$\mathbf{S}_{eb} = \tau \mathbf{I}_{K-1} + \mathbf{S}_e, \tag{32}$$

with (N - J + f) df. To obtain an empirical Bayes (EB) solution, the hyperparameters of f and τ are estimated from the observed data (see Boik, 1997 for formulas).

Boik (1997) found the EB approach adequately controlled Type I error and was more powerful than multivariate and adjusted univariate approaches. Keselman, Kowalchuk, and Boik (in press) further investigated the robustness of the EB procedure comparing it to adjusted df univariate, multivariate, WJ, and CIGA methods. As expected, the EB approach was sensitive to the same conditions that affect the robustness of the approaches that comprise this method (i.e., covariance heterogeneity combined with unequal group sizes). Thus, the EB approach is only recommended when data are normally distributed and group sizes are equal.

Mixed Model Approach. An analysis strategy for RM designs recommended in the literature is based on a mixed model approach which allows users to model the covariance structure of their data rather than presume certain structures as is the case with conventional univariate and multivariate test statistics (Jennrich & Schluchter, 1986; Liang & Zeger, 1986; Wolfinger, 1993, 1996). Being able to specify the structure of the covariance matrix should lead to a more parsimonious model of the data and as a result more powerful tests of the fixed-effect parameters (Wright & Wolfinger, 1996). This mixed model approach is now available through SAS' (1996, 1999) PROC MIXED procedure.

The general linear mixed model is (SAS Institute, 1996)

$$\mathbf{Y}_u = \mathbf{X}\beta + \mathbf{Z}\gamma + \xi_u \,, \tag{33}$$

where $\mathbf{Y}_u$ is an m x 1 vector of measured responses (i.e., univariate data, m = N x K), $\mathbf{X}$ is an m x p known design matrix, $\boldsymbol{\beta}$ is a p x 1 vector of unknown fixed-effects, $\mathbf{Z}$ is an m x q known design matrix, γ is a q x 1 vector of unknown random-effects, and ξ_u is an m x 1 unknown error vector. Both γ and ξ_u have expectations $\mathbf{0}$ and variances $\mathbf{G}$ and $\mathbf{R}$, respectively. The variance of $\mathbf{Y}_u$ is therefore equal to $\mathbf{V} = \mathbf{ZGZ}^T + \mathbf{R}$, where $\mathbf{R}$ is an m x m block diagonal matrix with blocks corresponding to the individual units of analysis with each block having a specified variance-covariance structure. In contrast to the previous general linear model, $\mathbf{Y}_u$ is a univariate representation of multivariate data (i.e., the multiple responses of each unit of analysis are stacked into a single vector). The name mixed model refers to the fact that both fixed-effect ($\boldsymbol{\beta}$) and random-effect (γ) parameters are contained in the model. If $\mathbf{Z} = \mathbf{0}$ and $\mathbf{R} = \sigma\mathbf{I}_N$, then the mixed model reduces to the general linear model.

An initial step is to estimate $\mathbf{G}$ and $\mathbf{R}$. PROC MIXED uses two likelihood based methods; maximum likelihood (ML) and restricted/residual maximum likelihood (REML). The details which are beyond the scope of this paper can be found in Wolfinger, Tobias, and Sall (1994). Based on simulation studies, REML is recommended (Wright, 1995; Wright & Wolfinger, 1996). To obtain estimates for $\boldsymbol{\beta}$ and γ, the solution to the following mixed model equations (Searle, 1971) is needed

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}\begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Y}_u \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Y}_u \end{pmatrix}. \tag{34}$$

The solutions can also be written as

$$\hat{\beta} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^- \mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}_u , \tag{35}$$

and

$$\hat{\gamma} = \mathbf{G}\mathbf{Z}^1\mathbf{V}^{-1}(\mathbf{Y}_u - \mathbf{X}\hat{\beta}) . \tag{36}$$

The covariance matrix of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\gamma}}$ is

$$\mathbf{W} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^- , \tag{37}$$

where "-" denotes a generalized inverse (Searle, 1971). With only fixed-effects included in the model, the variance-covariance matrix of $\boldsymbol{\beta}$ is reduced to $(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X})^-$. Statistical inference on fixed-effects of the model are obtained by testing the following null hypothesis

$$H_o: \mathbf{L}\hat{\beta} = \mathbf{0} , \tag{38}$$

where $\mathbf{L}$ is a df_L x p contrast matrix, with rank($\mathbf{L}$) = $df_L \le p$. A general F-statistic is given by

$$F = \frac{\hat{\beta}^T\mathbf{L}^T(\mathbf{L}\mathbf{W}\mathbf{L}^T)^- \mathbf{L}\hat{\beta}}{\text{rank}(\mathbf{L})} , \tag{39}$$

which has an approximate F-distribution with df rank($\mathbf{L}$) and v. The MIXED procedure provides various options for denominator df for tests of fixed-effects (see SAS Institute, 1996, pp. 565-566). For example, one can select a Satterthwaite solution described by

Giesbrecht and Burns (1985), McLean and Sanders (1988), and Fai and Cornelius (1996).

Selecting a CS or HF covariance structure to model the data in the MIXED procedure gives the results for a conventional univariate analysis of RM. When an UN covariance matrix is fit to the data, the F-statistics are a scalar multiple of the multivariate Lawley-Hotelling trace statistic (Kleinbaum, 1973; Wright, 1995) but not one of the usual F-approximations of multivariate tests reported by the SAS GLM procedure. In addition to the covariance structures previously mentioned (i.e., CS, HF, and UN), other structures that could be selected to model RM data include First-Order Autoregressive (AR1) structure, Heterogeneous First-Order Autoregressive (ARH1) structure, Heterogeneous Compound Symmetric (CSH) structure, and a linear Random Coefficient (RC) structure (Wolfinger, 1996). The autoregressive and random coefficient structures model data such that measurements taken closer together are more highly correlated than measurements taken farther apart, which is characteristic of RM data. The AR1 structure has an additional property such that points that are a fixed distance apart have a consistent correlation pattern.

The covariance structures not previously defined have the following form (when K = 4);

(a) AR1

$$\mathrm{AR1} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & \rho \\ & & & 1 \end{bmatrix}, \tag{40}$$

(b) ARH1

$$\text{ARH1} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho \\ & & & \sigma_4^2 \end{bmatrix}. \quad (41)$$

(c) CSH

$$\text{CSH} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho \\ & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho \\ & & & \sigma_4^2 \end{bmatrix}. \quad (42)$$

and (d) RC

$$\text{RC} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}^T + \begin{bmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \sigma^2 & \\ & & & \sigma^2 \end{bmatrix}. \quad (43)$$

The CS and AR1 structures are considered homogeneous structures since the variances along the main diagonal are equal, however they differ in terms of their off-diagonal elements. That is, the covariances of the AR1 structure decrease exponentially, whereas the covariances of the CS structure remain constant. Furthermore, both the CS and AR1 structures require only two parameters to be estimated. Generalizations of CS and AR1 that allow distinct variances along the main diagonal are considered heterogeneous structures and are known as CSH and ARH1, respectively. Both CSH and

ARH1 structures require (K+1) parameters to be estimated. The HF structure is similar to the CSH structure, that is, it has the same variance structure along the main diagonal and the same number of unknown parameters. The UN structure is considered the most general heterogeneous structure with unequal variances and covariances and requires K(K+1)/2 parameters to be estimated. The RC structure is a linear model that allows random modeling of an intercept and slope with time as the independent variable.

An advantage of the MIXED procedure is that it allows the user to specify, separately and jointly, between-subjects and within-subjects heterogeneity. Between-subjects heterogeneity occurs when subjects exhibit different variance patterns across groups but are similar within a group. On the other hand, within-subjects heterogeneity occurs when the data from the same subject does not exhibit a constant variance across measurement occasions. Between-subjects heterogeneity is specified by the option GROUP=effect on the REPEATED statement in SAS. This results in all observations from a single level of a grouping variable having the same estimated covariance parameters with each group level having different parameters but the same covariance structure. Within-subjects heterogeneity is specified by the choice of variance-covariance structure fit to the data.

An important consideration when using this approach is choosing the covariance structure that best describes or models the data. Wolfinger (1993, 1996) presented a three stage approach based on an adaptation of a method presented by Diggle (1988) to select the most appropriate covariance structure. The third stage relies on formal statistical techniques to compare variance-covariance structures. Specifically, two model fit criteria from the MIXED procedure can be used to select the "best" covariance structure; Akaike's Information Criterion (AIC) (Akaike, 1974) and Schwarz' Bayesian Criterion (SBC) (Schwarz, 1978) (Littell et al., 1996). The form of the two criteria are

$$AIC = l_R(\hat{\theta}) - q , \tag{44}$$

and

$$SBC = l_R(\hat{q}) - q/2 \log(n-p) . \tag{45}$$

where $\hat{\theta}$ is the restricted/residual maximum likelihood estimate of the unknown variance-covariance parameter θ, q is the number of unknown elements of θ, n = N x K, and p = J x K (Wolfinger, 1996). The respective values of these two criteria are compared across various covariance structures with the rule that larger-is-better (i.e., the structure with the largest criterion value is the best covariance structure for the data and therefore one should interpret the fixed-effect tests associated with this particular structure). The two criteria may not necessarily agree on the best structure since the SBC has a stronger penalty which is a function of the number of unknown parameters and sample size. Therefore, the SBC will likely favor more parsimonious models compared to AIC. Keselman, Algina, Kowalchuk, and Wolfinger (1998) found this to be true in the context of a simulation study. The authors investigated the Akaike (1974) and Schwarz (1978) criteria in an unbalanced nonspherical heterogeneous RM design in which the true covariance structure of the data, the distributional form of the data, as well as sample size was varied. Their results indicated that neither criterion uniformly performed well and in particular the Schwarz criterion more frequently picked the wrong covariance structure. That is, a more parsimonious covariance structure, one with fewer unknown parameters, was more often picked by the Schwarz criterion than the true structure.

In addition to the above mentioned model fit criteria, Dawson, Gennings, and Carter (1997) presented two graphical techniques (i.e., draftsman's display plots and parallel axis plots) that can be used to determine the variance-covariance structure of RM data before using the MIXED procedure. Examining graphical plots of one's data is a useful technique, however the recommendation of the authors to also use the Akaike and Schwarz criteria to make the final selection of the best fitting covariance structure is not without problems as previously noted by Keselman et al. (1998). Furthermore, the default F-tests from the MIXED procedure can be biased under certain conditions (Wright, 1995; Wright & Wolfinger, 1996).

Given the availability of this approach to the analysis of RM designs through PROC MIXED, Keselman, Algina, Kowalchuk, and Wolfinger (1999b) compared the mixed model approach to the WJ test, the CIGA test, a multivariate test, and Greenhouse and Geisser (1959) and Lecoutre (1991) modified Huynh and Feldt (1976) adjusted df methods. Specifically, rates of Type I error control were investigated in unbalanced nonspherical RM designs having one between-subjects and one within-subjects variable when covariance homogeneity and normality assumptions were violated separately and jointly. The covariance structures investigated were UN, ARH1, and RC; heterogeneous within-subjects and heterogeneous within- and between-subjects structures. The default F-tests available with the MIXED procedure generally became conservative or liberal when unequal group sizes were paired either positively or negatively with unequal covariance matrices. The rates of Type I error for the RM effects were not controlled even when the correct covariance structure was fit to the data. As expected the WJ and CIGA approaches were able to control their Type I error rates, however, sample size requirements enumerated by Keselman et al. (1993) were necessary in order to obtain a robust WJ test.

The mixed model analyses of RM effects investigated by Keselman et al. (1999b) were based on default F-tests available through PROC MIXED in which the error df corresponded to that of the conventional univariate F-test with the exception of tests based on RC and UN covariance structures (see SAS Institute, 1996, pp. 565-566). However, a user has the option of requesting F-approximations based on Satterthwaite's df solution to test for RM fixed-effects. The conjecture that better Type I error control may be achieved with this option was investigated by Keselman, Algina, Kowalchuk, and Wolfinger (1999a). The results showed that if the correct covariance structure was selected, PROC MIXED Satterthwaite F-tests which allow for within- and between-subjects heterogeneity can in most cases effectively control their rates of Type I error when the data are nonnormal in form, covariance matrices are unequal, and the design is unbalanced. An important caveat however is that the user must know the true covariance structure to obtain robust tests of the RM fixed-effects with PROC MIXED.

Multiple Comparison Procedures

The words comparison and contrast are used interchangeably to refer to a linear contrast of means. The term multiple implies that there can be many different comparisons among a set of means. However, this discussion will be restricted to procedures that test all possible pairwise comparisons among a set of means.

Control of Type I error is an important consideration when doing multiple testing. An error rate per-contrast or per-comparison (α_{PC}) is the probability of committing a Type I error on a single contrast (i.e., falsely declaring a comparison significant). A problem with adopting this form of Type I error control when conducting multiple tests is that the probability of at least one Type I error increases exponentially as the number of comparisons increase. Another approach is to use a MCP that limits the overall (familywise) error rate to a nominal alpha level. Familywise error rate (α_{FW}) is the probability that one or more Type I errors will be made on a set (i.e., family) of comparisons. Authors typically agree that Type I error control is of primary importance followed by power (e.g., Keselman, 1994; Kirk, 1995; Seaman, Levin, & Serlin, 1991). That is, only those procedures that are robust with respect to Type I error are further evaluated in terms of power.

Three definitions of power commonly used are (Einot & Gabriel, 1975; Ramsey, 1978): (a) any-pairs power, (b) average per-pair power, and (c) all-pairs power. Any-pairs power is the probability of rejecting at least one true pairwise difference. Average per-pair power is the average probability of rejecting true pairwise differences. Lastly, all-pairs power is the probability of rejecting all true pairwise differences. With regards to MCP testing, only average per-pair and all-pairs power are meaningful, whereas if a researcher is interested in detecting any difference among means, then the power of the omnibus (F) test would be most relevant (Keselman, 1994; Ramsey, 1978).

There are two main types of MCPs that limit their familywise error rate to alpha: (a) simultaneous procedures, and (b) stepwise or sequential procedures. Simultaneous procedures use one critical value for all pairwise comparisons, whereas stepwise procedures require a sequence of critical values. Examples of popular simultaneous MCPs include the Bonferroni (Dunn, 1961) and Tukey (1953) procedures, while well known stepwise MCPs include the Newman (1939)-Keuls (1952) and Fisher's (1935) least significant difference procedures. However, these stepwise procedures cannot control α_{FW} when $K > 3$. As a result, Ryan (1960), Einot and Gabriel (1975), and Welsch (1977) proposed modifications to the Newman-Keuls method (known in SAS as the REGW method) and Hayter (1986) proposed a modification to Fisher's procedure.

With numerous MCPs to choose from, a researcher is presented with a difficult task of choosing the "best" method. Numerous Monte Carlo studies (e.g., Keselman, Keselman, & Shaffer, 1991; Keselman & Lix, 1995; Keselman, Lix, & Kowalchuk, 1998; Seaman et al., 1991) provide information that can be used to judge which method is most appropriate under certain conditions. A review of the educational and psychological literature (Keselman et al., 1998; Kowalchuk et al., 1996) found almost half of the articles incorporating a mixed design used multiple pairwise comparisons of RM means. The most popular method used was Tukey's (1953) procedure followed by Newman-Keuls (Keuls, 1952; Newman, 1939), multiple t-tests, and Bonferroni (Dunn, 1961) procedures. The selection of MCPs chosen by researchers has not changed since an earlier review by Jacard, Becker, and Wood (1984). The popularity of these procedures is likely based on their availability in statistical packages. In the present study, the MCPs examined included simultaneous procedures most commonly used by researchers (e.g., Tukey and Bonferroni) and stepwise procedures which have been found, through Monte Carlo studies (see Keselman, 1994; Keselman & Lix, 1995) to be robust with regards to Type I error control and powerful to detect true differences (e.g., Shaffer's (1986) sequentially rejective Bonferroni procedure and Welsch's (1977) step-up range procedure).

Contrast (Pairwise) Test Statistic. A contrast among levels of the RM marginal means is given by $\psi = \mathbf{c}_k^T \boldsymbol{\mu}_k$, where $\mathbf{c}_k^T$ is a coefficient vector of weights, $\Sigma_k \mathbf{c}_k^T = 0$, and $\boldsymbol{\mu}_k$ is the vector of K population means. The notation for a contrast is also given by $\psi = \Sigma_j a_j (\mathbf{c}_k^T \boldsymbol{\mu}_{jk})$, where $a_j = 1/J$ for all $j = 1, \ldots, J$ (i.e., an unweighted means analysis) and is estimated by $\hat{\psi} = \Sigma_j a_j (\mathbf{c}_k^T \mathbf{X}_{jk})$. An estimate of the variance of a contrast is given by $\sigma^2_j(\hat{\psi}) = \Sigma_j a_j^2 (\mathbf{c}_k^T \mathbf{S}_j \mathbf{c}_k)/n_j$. A general form of the test statistic for the hypothesis H_o: $\psi = 0$ is given by

$$\frac{\hat{\psi}}{\sigma(\hat{\psi})} . \tag{46}$$

The conventional test of a contrast uses a pooled estimate of error variance based on the interaction mean square (e.g., $MS_{KxS/J}$). The t-ratio is given by

$$\frac{(\hat{\psi} - \psi_o)}{\sqrt{MS_{KxS/J} \sum_{k=1}^{K} \mathbf{c}_k^T \mathbf{c}_k \sum_{j=1}^{J} \frac{a_j^2}{n_j}}} , \tag{47}$$

where $MS_{KxS/J}$ is the within-subjects error term from a J (between) x K (within) univariate ANOVA. This t-statistic provides an exact test of the null hypothesis, H_o: $\psi = 0$, if and only if the assumption of sphericity (multisample sphericity) is satisfied (Keselman, 1982; Mitzel & Games, 1981). Maxwell (1980) found the use of the Tukey procedure with a common estimate of variance in a simple RM design under violation of sphericity lead to inflated Type I error rates under a complete null hypothesis. The use of a constant estimate of variance assumes that the variance of each contrast is equal. However, this is not likely to occur with real data and as a result the pooled estimate of variance may be too large or too small for some contrasts. Therefore, individual estimates of variance that vary from contrast to contrast should be used (Keselman, 1982; Keselman, Rogan, & Games, 1981; Maxwell, 1980; Mitzel & Games, 1981). Thus, an alternative form of the *t*-ratio has an error term based on an estimate of variance that pools across the levels of the between-subjects variable but considers only the levels of the RM variable of interest. The *t*-ratio is expressed as

$$\frac{(\hat{\psi} - \psi_{o})}{\sqrt{\sum_{j=1}^{J} \frac{a_j^2}{n_j} \frac{\sum_j (n_j - 1)\, \mathbf{c}_k^T \mathbf{S}_j \mathbf{c}_k}{N - J}}}, \tag{48}$$

where $\mathbf{c}_k$ is a coefficient vector representing a pairwise contrast among the RM marginal means. The validity of this test is therefore based on satisfying the between-subjects condition of the multisample sphericity assumption (Keselman, 1982). That is, the value of the variance estimate for each contrast is constant across the levels of the between-subjects variable.

Keselman and Keselman (1988) compared four simultaneous MCPs for testing pairwise RM marginal means under violation of multisample sphericity in an unbalanced design containing one between-subjects and one within-subjects variable. A Tukey approach with a pooled estimate of error variability ($MS_{K \times S/J}$) and three approaches based on a pooled sample covariance matrix (i.e., pooled across the levels of the between-subjects variable); a modified Tukey approach, a Bonferroni approach, and an approach using a multivariate critical value. Results indicated that for tests of unweighted means all four simultaneous procedures failed to consistently provide Type I error control under violation of multisample sphericity. Therefore, Maxwell's (1980) recommendation of a Bonferroni approach based on separate variance estimates for each pairwise comparison in a one-way (simple) RM design cannot be extended to RM designs containing a between-subjects variable.

Keselman et al. (1991) presented a statistic (KKS) based on a variance estimate that is not dependent on multisample sphericity constraints. That is, the error term does not pool across the levels of the between-subjects factor and considers only the levels of the RM variable of interest. The test statistic is expressed as

$$\frac{(\hat{\psi} - \psi_0)}{\sqrt{\sum_{j=1}^{J} \frac{a_j^2}{n_j} (\mathbf{c}_k^T \mathbf{S}_j \mathbf{c}_k)}}, \tag{49}$$

which can be approximated as a t-variable with Satterthwaite (1941, 1946) df given by

$$v_s = \frac{\left[\sigma(\hat{\psi})\right]^4}{\sum_{j=1}^{J} \left(\frac{\left(\frac{a_j^2}{n_j} \right)^2 (\mathbf{c}_k^T \mathbf{S}_j \mathbf{c}_k)^2}{n_j - 1} \right)}. \tag{50}$$

Keselman et al. (1991) compared four approaches for pairwise comparisons among RM means using the KKS statistic in an unbalanced J x K design. Type I error control was investigated under conditions of nonnormality and variance-covariance heterogeneity. Results indicated that the nonpooled statistic based on Satterthwaite df with a Studentized range, a Studentized maximum modulus, or a Bonferroni critical value provided adequate α_{FW} control under most conditions.

Using the KKS statistic, Keselman (1993) investigated several stepwise MCPs under violation of multisample sphericity in a RM design containing one between- and one within-subjects variable. Based on Type I error control, the author recommended the following three procedures; Welsch's (1977) step-up procedure, Hayter's (1986) modified two-stage least significant difference procedure, and Shaffer's (1986) sequentially rejective Bonferroni procedure which begins with an omnibus test. Either a corrected df univariate F-test or a multivariate test was used as the omnibus test in those procedures requiring this first step. Furthermore, Keselman (1994) compared previously investigated stepwise and simultaneous MCPs (Keselman, 1993; Keselman et al., 1991) in a J x K design under conditions of nonnormality and variance-covariance heterogeneity. Welsch's (1977) step-up procedure was considered "superior to all of the other MCPs" (Keselman, 1994, p. 154) in terms of Type I error control and power to detect nonnull pairwise differences.

Because of the sensitivity of the univariate F-test, the adjusted df univariate approaches, and the multivariate approach to violation of multisample sphericity given unequal group sizes, a robust alternative was investigated as the omnibus test for those procedures that require an omnibus test at stage one (Keselman & Lix, 1995). The WJ procedure was used as the omnibus test along with the KKS statistic with numerous stepwise MCPs for RM means under nonnormality and variance-covariance heterogeneity. The authors recommended the Welsch (1977) step-up procedure, Hayter's (1986) two-stage modified least significant difference procedure, Shaffer's (1986) sequentially rejective Bonferroni procedure that begins with an omnibus test, and the following procedures modified by a technique described by Duncan (1957) including the Peritz (1970) procedure, Ryan-Welsch (Ryan, 1960; Welsch, 1977) multiple range procedure, and a multiple range procedure that begins with an omnibus test (Shaffer, 1979, 1986).

In summary, robust MCPs have been identified (Keselman 1994; Keselman & Lix, 1995; Keselman et al., 1991) based on a test statistic that uses a nonpooled error term and a between-subjects heterogeneous UN form of the variance-covariance matrix with df adjusted by Satterthwaite's (1941, 1946) solution. Furthermore, Lix and Keselman (1995) provide a SAS/IML (SAS Institute, 1989) program to compute these tests in RM designs. However, the availability of PROC MIXED now allows researchers to select among various forms of variance-covariance matrices to model their RM data. As well, the option of Satterthwaite df hypothesis testing for fixed-effect omnibus tests and pairwise comparisons on RM marginal means is also now available through the MIXED procedure. Should researchers adopt a mixed model methodology, in particular PROC MIXED, for testing pairwise multiple comparisons of RM means? This question bears investigation because currently researchers need not adopt mixed model methodology to obtain robust tests of pairwise multiple comparisons of RM means.

Therefore, the purpose of the present study was to investigate Type I error control for simultaneous and stepwise MCPs using PROC MIXED in an unbalanced between- by within-subjects design under violation of normality and variance-covariance homogeneity (i.e., multisample sphericity). Thus, the potential advantage of PROC MIXED with its ability to specify various covariance structures such as AR1, ARH1, and RC was compared to a known robust procedure (i.e., KKS statistic) based on an UN covariance structure. Furthermore, MCPs that adequately controlled their rates of Type I error were then compared for their sensitivity to detect true pairwise differences.

Monte Carlo Study

A Monte Carlo study was used to investigate the robustness of selected MCPs available through SAS' (1996, 1999) PROC MIXED procedure. To investigate α_{FW} Type I error control, a simulation study was designed with a true null hypothesis. A set of pseudorandom numbers were generated using a computer algorithm to sample from a population with known characteristics. The experiment was replicated numerous times and for each replication, a test statistic was computed from the generated data, and was compared to a theoretically known critical value. Based on this result, the null hypothesis was either rejected or retained. Thus, an empirical estimate of Type I error was obtained.

Design

A RM design containing one between-subjects and one within-subjects factor with the number of levels of the between-subjects factor equal to three and the number of levels of the within-subjects factor equal to four and eight was investigated.

Study Variables

The following variables were manipulated: (a) population covariance structure, (b) covariance structures, (c) group sizes, (d) pairings of covariance matrices and group sizes, (e) shape of the data, (f) the covariance structure fit to the data, (g) type of null hypothesis, and (h) type of nonnull mean configuration.

Because published research does not contain enough information to determine the extent to which assumptions are satisfied (e.g., normality, sphericity, variance-covariance homogeneity across groups) it is difficult to know the type of data likely to be encountered by educational and psychological researchers. Therefore, the conditions investigated were selected to represent a range of possibilities that may occur in applied settings.

Population Covariance Structures. The following types of covariance structures were used to generate simulated data: (a) ARH1, (b) RC, and (c) UN. Each of these structures models data that exhibits within-subjects heterogeneity (i.e., variances along the main diagonal were unequal) and violates the sphericity assumption (i.e., epsilon = 0.75).[1] See Appendix A for element values of the population covariance structures for K = 4 and K = 8.

Group Covariance Structures. Homogeneous (i.e., equal across groups) and heterogeneous (i.e., unequal across groups) covariance structures were investigated. Specifically, the unequal group covariance matrices were in a 1:3:5 ratio, that is $\boldsymbol{\Sigma}_1 = 1/3\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3 = 5/3\boldsymbol{\Sigma}_2$. This ratio was chosen because previous studies (e.g., Keselman, 1994; Keselman & Lix, 1995; Keselman et al., 1993; Keselman et al., 1999a, 1999b) have found it to have a negative effect on Type I error control and therefore presents a condition that may also affect the validity of the tests examined in this investigation.

Group Sizes. Equal and unequal group sizes were investigated. Total sample size (N) was set equal to 45, and 60. These sample sizes were chosen because a review of the empirical literature (Keselman et al., 1998; Kowalchuk et al., 1996) found that more than half of the articles containing a mixed design reported a total sample size of 60 or less. For each value of N, two conditions of group size inequality were examined, a moderate degree and a substantial degree of inequality. A coefficient of sample size variation (C_v) was set equal to approximately .16 for the moderate condition and .33 for the more disparate condition. C_v is defined as

$$C_v = \frac{\sqrt{\sum_{j=1}^{J} \frac{(n_j - \bar{n})^2}{J}}}{\bar{n}} . \tag{51}$$

where $\bar{n}$ is the average group size. The two unequal sample size cases for each total sample size were respectively, (a) 12, 15, 18 and 9, 15, 21 (N = 45), and (b) 16, 20, 24 and 12, 20, 28 (N = 60).

Pairings of Covariance Matrices and Group Sizes. Positive and negative pairings of covariance matrices and group sizes were investigated. These pairings are known to produce conservative and liberal rates of Type I error, respectively. For each total sample size condition there were six pairings of covariance matrices and group sizes investigated: (a) equal n_j, equal Σ_j; (b) equal n_j, unequal Σ_j; (c/c') unequal n_j, unequal Σ_j (positively paired); and (d/d') unequal n_j, unequal Σ_j (negatively paired). The c'/d' condition corresponds to the more disparate unequal group size cases and the c/d condition corresponds to the less disparate unequal group size cases.

Data Generation. Although the test procedures investigated are based on the assumption of multivariate normality, this condition is unlikely to be satisfied when working with real data. For example, Micceri (1989) examined 440 measures characteristic of psychological and educational data and found none even approximated a normal distribution. Furthermore, the two sample independent t- test is sensitive with respect to Type I error control when data are sampled from distributions with extreme degrees of skewness (e.g., 1.64; Sawilowsky & Blair, 1992). Therefore, data were generated from multivariate normal and nonnormal population distributions in order to provide conditions in which the tests may not perform favorably.

To generate multivariate (i.e., K-variate) normal data, pseudorandom observation vectors $\mathbf{Y}_{ij}^T = [Y_{ij1}, Y_{ij2}, \ldots, Y_{ijK}]$ with a mean vector $\boldsymbol{\mu}_j^T = [\mu_{j1}, \mu_{j2}, \ldots, \mu_{jK}]$ and covariance matrix $\boldsymbol{\Sigma}_j$ were obtained by a triangular decomposition of $\boldsymbol{\Sigma}_j$ (referred to as the Cholesky factorization or square root method):

$$\mathbf{Y}_{ij} = \boldsymbol{\mu}_j + \mathbf{L}\mathbf{Z}_{ij}\,, \tag{52}$$

where $\mathbf{L}$ is an upper triangular matrix satisfying the equality $\mathbf{L}^T\mathbf{L} = \boldsymbol{\Sigma}_j$, and $\mathbf{Z}_{ij}$ is an independent normally distributed vector. The vectors of observations were obtained by the RANNOR function in SAS (1989).

The nonnormal data was a multivariate lognormal distribution with marginal distributions based on $Y_{ijk} = \exp(X_{ijk})$, where $X_{ijk} \sim N(0, .25)$. Skewness and kurtosis values are 1.75 and 5.90, respectively. Algina and Oshima (1994, 1995) provide details of the steps involved to generate multivariate lognormal data.

Covariance Structures Fit to the Data. In addition to the true population covariance structure, other selected covariance structures that model between-subjects and within-subjects heterogeneity, separately and jointly were also fit to the data. The following 12 covariance structures were fit to the data: (a) UN, (b) UN-H, (c) ARH1, (d) ARH1-H, (e) RC, (f) RC-H, (g) HF, (h) HF-H, (i) CSH, (j) CSH-H, (k) AR1, (l) AR1-H, where the "-H" corresponds to the between-subjects heterogeneous version of the covariance structure. The AIC and SBC criteria from SAS' (1996, 1999) PROC MIXED were used to select the best covariance structure among the 12 possible structures. That is, the pairwise test statistics for the MCPs were based on the covariance structure selected by AIC or SBC. In addition, the test statistics for the MCPs were based on an UN-H covariance structure and the correct covariance structure. Therefore, four testing strategies were compared. That is, one approach based on always assuming an UN-H covariance structure, a second approach based on prior knowledge of the true population covariance structure, a third approach based on using the AIC criterion to select the best covariance structure, and a fourth approach based on using the SBC criterion to select the best covariance structure.

Null Hypothesis. Empirical Type I error rates were collected when the population mean vectors reflected a complete null hypothesis (e.g., $\mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$; $\mu_1 = \mu_2 = \ldots = \mu_7 = \mu_8 = 0$) and a partial null hypothesis (e.g., $\mu_1 = \mu_2 \neq \mu_3 = \mu_4$; $\mu_1 = \ldots = \mu_4 \neq \mu_5 = \ldots = \mu_8$). A partial null hypothesis occurs when some population means are equal but the overall null hypothesis is not true (Toothaker, 1991, p. 13). Although a partial null hypothesis could be represented by many nonnull configurations only one was chosen for each level of K. Because a researcher never knows the true state of a null hypothesis, it is important that a MCP control α_{FW} under both complete and partial null hypotheses.

Nonnull Mean Configuration. Three types of mean configurations were investigated (Ramsey, 1978): (a) a minimum range configuration, where the first half of the means in the range are equal and the second half are also equal but different from the first half; (b) a maximum range configuration, where the first and last mean represent two extremes and the remaining means are the average of these two extreme values; and (c) an equally spaced range configuration, where the means are equally spaced across the range. Two definitions of power were investigated: (a) all-pairs power and (b) average per-pair power.

Description of Multiple Comparison Procedures

The following MCPs available through SAS' (1996) PROC MIXED procedure were investigated: (a) Bonferroni, (b) Sidák, (c) Tukey, and (d) Studentized maximum modulus (SMM/GT2). In addition, the following MCPs, not available through SAS were also investigated: (a) Shaffer's (1986) sequentially rejective Bonferroni procedure, (b) Hochberg's (1988) sequentially acceptive Bonferroni procedure, and (c) Welsch's (1977) step-up range procedure. The MCPs available through SAS are simultaneous procedures, whereas the additional procedures investigated are stepwise procedures that can easily be evaluated using the statistical output provided by SAS. The inclusion of stepwise procedures is based on their generally superior performance compared to simultaneous procedures with regard to power rates (Keselman, 1994).

The following simultaneous procedures are available by default through SAS (1996). The Bonferroni procedure (Dunn, 1961) tests each comparison at a α/c level of significance, where c is the number of pairwise comparisons (i.e., $c = K(K-1)/2$) and $\alpha = \alpha_{FW}$. The Sidák (1967) procedure tests each comparison at a $1 - (1 - \alpha)^{1/c}$ level of significance. The Sidák procedure is based on the multiplicative inequality in contrast to the additive Bonferroni inequality and therefore is slightly more powerful. The Tukey (1953) procedure tests each comparison by comparing the observed test statistic to a critical value from the Studentized range distribution ($q_{\alpha, K, v} / \sqrt{2}$) (see Scheffé, 1959 p. 28 for the specification of a Studentized range variable). Lastly, the Studentized maximum modulus procedure tests each contrast by comparing the observed test statistic to a critical value from the Studentized maximum modulus distribution ($M_{\alpha, c, v}$) (see Scheffé, 1959 p. 78 for the specification of a Studentized maximum modulus variable).

The stepwise procedures examined in this study were limited to those that do not require an omnibus test as a first step because of the lack of robustness of omnibus RM fixed-effects with PROC MIXED (Keselman et al., 1999a, 1999b). The three procedures chosen have shown promising results in previous simulation studies (Keselman, 1994; Keselman & Lix, 1995).

Shaffer's (1986) sequentially rejective Bonferroni procedure is a modification of Holm's (1979) procedure. Holm modified the Bonferroni procedure such that testing is done in a stepwise fashion with successively higher significance levels, thus the procedure has greater power than the Bonferroni procedure. Shaffer's modification further improves the power of Holm's procedure. Shaffer's procedure begins by arranging the p-values associated with the test statistics of the c-pairwise comparisons from smallest to largest (i.e., $p_1 \leq p_2 \leq, \ldots, \leq p_c$ corresponding to hypotheses $H_1, \ldots, H_c$). The smallest p-value (p_1) is compared to α/c. If $p_1 \geq \alpha/c$, then statistical testing stops and all remaining pairwise contrast hypotheses are retained. Otherwise, if $p_1 < \alpha/c$, then H_1 is rejected and one proceeds to test the remaining pairwise hypotheses in a stepwise fashion by comparing the associated p-values to α/c^*, where c^* is equal to the maximum number of true null hypotheses, given the number of hypotheses rejected at previous steps. The values for c^* can be obtained from Shaffer's (1986, p. 828) Table 2.

Hochberg's (1988) sequentially acceptive Bonferroni procedure is based on the same critical values as Holm's (1979) procedure but testing proceeds from the largest p-value and rejection of an hypothesis implies rejection of all hypotheses with equal to or smaller p-values. This procedure rejects all hypotheses ($H_{m'}$, where $m' \leq m$, and $m = c, c-1, \ldots, 1$) if $p_m \leq \alpha/(c - m + 1)$. Thus, one begins by examining the largest p-value (p_c). If $p_c \leq \alpha$ then all hypotheses are rejected. If $p_c > \alpha$ then H_c is accepted, and one proceeds to compare p_{c-1} to $\alpha/2$. If $p_{c-1} \leq \alpha/2$ then all hypotheses ($m = c-1, \ldots, 1$) are rejected, otherwise H_{c-1} is retained and one proceeds to compare p_{c-2} to $\alpha/3$, and the process (i.e., steps) continues.

Welsch (1977) proposed a step-up range procedure that begins by rank ordering the means and examining adjacent means (i.e., two-range tests) first. If any two-range test(s) are significant then any larger set of means that contain the significant subset(s) are declared significant by implication. If any two-range test(s) are found nonsignificant, then one proceeds to test larger range tests (e.g., three-range). If a three-range test is significant then all larger sets of means that contain this subset are declared significant by implication. The range tests for a set of means vary from r = 2 (two-range test) to r = K (K-range test). The table of critical values can be found in Keselman (1994).

Pairwise Test Statistic. The form the of the t-test statistic from PROC MIXED (SAS Institute, 1996) is

$$t = \frac{\mathbf{L}\hat{\beta}}{\sqrt{\mathbf{LWL}^T}}, \quad (53)$$

where **L** consists of a single row contrast vector, $\boldsymbol{\beta}$ is a p x 1 vector of fixed-effect parameters, and **W** is the variance-covariance matrix associated with $\boldsymbol{\beta}$. The df were estimated using two options available through PROC MIXED, the default option (i.e., BETWITHIN or CONTAIN) and Satterthwaite's solution (see SAS Institute, 1996, pp. 565-566).

The value obtained from this statistic is similar to using the KKS (Keselman et al., 1991) statistic (see equation 49) with $\mathbf{S}_j$ replaced by the estimated variance-covariance matrix selected to model the data through SAS' (1996, 1999) PROC MIXED. The flexibility of the MIXED procedure is that it allows a user to specify various variance-covariance structures rather than always assuming an Unstructured between-subjects heterogeneous structure (i.e., UN-H). Therefore, selecting the UN-H covariance structure with the Satterthwaite df option with the MIXED procedure is equivalent to the KKS statistic with Satterthwaite df.[2]

Simulation Program

The program was written in SAS MACRO (SAS Institute, 1997) and SAS/IML (SAS Institute, 1989) languages and was run on release version 8.0 of SAS. Because the computational time required for PROC MIXED (SAS Institute, 1996) is substantial only selected combinations of the eight study variables were examined using 1000 simulations or replications with a .05 level of significance.

Results

Type I Error Rates

To evaluate the robustness of a MCP, Bradley's (1978) liberal criterion was used. That is, if an empirical estimate of Type I error ($\hat{\alpha}$) was contained within the interval of $.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$, then the test procedure was considered robust. For an alpha level of .05 the interval is $.025 \leq \hat{\alpha} \leq .075$. If Type I error was not contained in this interval then a test procedure was considered nonrobust. In the tables, bold entries correspond to these latter values. Other quantitative measures of robustness reported in the literature include Bradley's stringent criterion (i.e., $.9\alpha \leq \hat{\alpha} \leq 1.1\alpha$) and a binomial standard error approach [e.g., plus or minus two or three times $(\alpha\,(1-\alpha)/N)^{1/2}$, where N is the number of simulations; see Wright & Wolfinger, 1996]. The choice of robustness criterion may lead to different interpretations of Type I error results. Although no universal standard is available, Bradley's liberal criterion provides an acceptable range of values to judge robustness. That is, an applied researcher should feel comfortable with a procedure that controls Type I error within these bounds, if the procedure limits the rate across a wide range of conditions in which assumptions are violated.

Four testing strategies were compared to evaluate the operating characteristics of seven MCPs. That is, one approach based on always assuming a UN-H variance-covariance structure (i.e., KKS approach), a second approach based on prior knowledge of the true population variance-covariance structure, a third approach based on using the AIC criterion to select the best structure, and a fourth approach based on using the SBC criterion to select the best structure. The seven MCPs investigated included four simultaneous procedures [Bonferroni (Bon), Sidak, Tukey, and Studentized maximum modulus (SMM)] and three stepwise procedures [Shaffer's sequentially rejective Bonferroni (SRB), Hochberg's sequentially acceptive Bonferroni (Hoch), and Welsch's step-up range]. The MCPs were computed using a test statistic with a nonpooled error term (i.e., does not pool across the between- and within-subjects factors). In addition, two ways of estimating the df for the test statistic were examined. One approach based on a Satterthwaite df solution and the other approach based on the default df option available through PROC MIXED.

The results presented are for selected combinations of the eight variables investigated which included (a) type of population variance-covariance structure, (b) homogeneous and heterogeneous group variance-covariance structures, (c) equal and unequal group sizes, (d) positive and negative pairings of variance-covariance matrices and group sizes, (e) multivariate normal and nonnormal data, (f) type of variance-covariance structure fit to the data, (g) type of null hypothesis, and (h) type of nonnull mean configuration. The combinations investigated were chosen to demonstrate differences among the MCPs with regard to their error rates.

Normally Distributed Data. Table 1 contains the study conditions collected when data were obtained from a normal distribution for K = 4 and N = 45. Positive and negative pairings of group sizes and variance-covariance matrices were investigated only for the more disparate unequal sample size condition. These conditions represent cases when multisample sphericity is not satisfied (i.e., when sphericity is not equal to one and the variance-covariance matrices are unequal across the levels of the grouping variable). Tables 2 through 5 contain Type I error rates for the four testing strategies, respectively based on Satterthwaite df and Tables 6 through 9 contain error rates for the four testing strategies, respectively based on default df. Tables 10 and 11 contain the percentages with which the AIC and SBC criteria selected the correct variance-covariance structure from among 12 possible structures.

Table 1

Study Conditions (Normal Distribution, K = 4, N = 45)

Condition	Pop Cov Str[a]	Sample Sizes	Cov Mat[b]	Pairing	Null Hypothesis
c'	ARH1-H	9,15,21	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d'	ARH1-H	9,15,21	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c'	RC-H	9,15,21	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d'	RC-H	9,15,21	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c'	UN-H	9,15,21	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d'	UN-H	9,15,21	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c'	ARH1-H	9,15,21	1:3:5	+	$\mu_1=\mu_2\neq\mu_3=\mu_4$
d'	ARH1-H	9,15,21	1:3:5	-	$\mu_1=\mu_2\neq\mu_3=\mu_4$
c'	RC-H	9,15,21	1:3:5	+	$\mu_1=\mu_2\neq\mu_3=\mu_4$
d'	RC-H	9,15,21	1:3:5	-	$\mu_1=\mu_2\neq\mu_3=\mu_4$
c'	UN-H	9,15,21	1:3:5	+	$\mu_1=\mu_2\neq\mu_3=\mu_4$
d'	UN-H	9,15,21	1:3:5	-	$\mu_1=\mu_2\neq\mu_3=\mu_4$

Note. [a] Population Covariance Structure, [b] Covariance Matrix.

Table 2

Empirical Type I Error Rates (%) when Fitting a UN-H Covariance Structure

(Normal Distribution, K = 4, N = 45, Satterthwaite df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	4.50	4.60	5.90	4.60	4.50	4.50	**7.70**
d'	4.60	4.60	5.10	4.80	4.80	4.80	**7.90**
RC							
c'	2.80	2.80	3.70	3.00	2.90	2.90	6.20
d'	4.10	4.40	5.40	4.60	4.60	4.60	**8.50**
UN							
c'	3.80	4.00	4.90	4.00	3.80	3.90	**8.30**
d'	3.60	3.60	5.30	3.90	4.50	4.50	7.50
Partial Null Hypothesis							
ARH1							
c'	**1.50**	**1.50**	**1.80**	**1.50**	4.00	3.80	4.80
d'	**1.30**	**1.30**	**1.80**	**1.40**	3.30	3.20	4.50
RC							
c'	**1.70**	**1.90**	**2.30**	**1.90**	3.60	3.50	4.70
d'	**1.60**	**1.70**	2.50	**1.70**	3.70	3.30	5.00
UN							
c'	**0.70**	**0.70**	**0.70**	**0.70**	**1.80**	**1.90**	3.70
d'	**1.20**	**1.20**	**1.30**	**1.20**	2.90	2.70	3.90

Note. Bon = Bonferroni; SMM = Studentized maximum modulus; SRB = Shaffer's (1986) sequentially rejective Bonferroni; Hoch = Hochberg's (1988) sequentially acceptive Bonferroni.

Table 3

Empirical Type I Error Rates (%) when Fitting the True Covariance Structure

(Normal Distribution, K = 4, N = 45, Satterthwaite df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	4.90	4.90	5.50	4.90	4.70	4.90	**8.30**
d'	3.50	3.60	4.90	3.80	3.80	3.80	7.30
RC							
c'	3.10	3.10	3.70	3.10	3.10	3.10	6.60
d'	3.60	3.80	5.00	4.20	3.40	3.40	6.10
UN							
c'	3.80	4.00	4.90	4.00	3.80	3.90	**8.30**
d'	3.60	3.60	5.30	3.90	4.50	4.50	7.50
Partial Null Hypothesis							
ARH1							
c'	**1.40**	**1.40**	**1.50**	**1.40**	3.80	3.50	4.60
d'	**1.30**	**1.30**	**1.70**	**1.30**	2.70	2.70	4.20
RC							
c'	**2.10**	**2.10**	2.60	**2.10**	4.50	4.30	5.20
d'	**1.80**	**1.80**	**2.10**	**1.80**	4.10	3.50	5.20
UN							
c'	**0.70**	**0.70**	**0.70**	**0.70**	**1.80**	**1.90**	3.70
d'	**1.20**	**1.20**	**1.30**	**1.20**	2.90	2.70	3.90

Note. See note from Table 2.

Table 4

Empirical Type I Error Rates (%) with Akaike Criterion Selecting the Best Covariance Structure (Normal Distribution, K = 4, N = 45, Satterthwaite df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	5.20	5.20	6.00	5.20	5.20	5.20	**8.60**
d'	3.60	3.70	4.60	3.70	3.80	3.90	**7.60**
RC							
c'	3.20	3.20	3.80	3.20	3.20	3.20	6.70
d'	3.60	3.70	5.10	4.10	3.30	3.30	6.60
UN							
c'	4.70	4.90	5.50	4.90	4.70	4.70	7.50
d'	4.10	4.10	4.90	4.20	4.60	4.60	**7.60**
Partial Null Hypothesis							
ARH1							
c'	**1.30**	**1.30**	**1.90**	**1.30**	3.40	3.10	4.40
d'	**1.20**	**1.20**	**1.60**	**1.30**	2.60	2.80	4.10
RC							
c'	**2.20**	**2.20**	2.60	**2.20**	4.40	4.30	5.20
d'	**1.90**	**1.90**	**2.20**	**1.90**	4.20	3.60	5.20
UN							
c'	**2.00**	**2.20**	2.50	**2.20**	4.30	4.00	6.30
d'	**2.30**	**2.40**	3.00	**2.40**	4.70	4.40	6.90

Note. See note from Table 2.

Table 5

Empirical Type I Error Rates (%) with Schwarz Criterion Selecting the Best Covariance Structure (Normal Distribution, K = 4, N = 45, Satterthwaite df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	5.10	5.10	5.70	5.10	5.10	5.10	**8.30**
d'	3.80	3.80	4.70	3.80	3.90	3.90	**7.70**
RC							
c'	**2.30**	**2.30**	2.90	**2.30**	**2.30**	**2.30**	4.80
d'	5.30	5.50	6.80	5.90	5.40	5.50	**8.70**
UN							
c'	4.80	4.90	5.40	5.00	4.90	4.90	**7.70**
d'	4.30	4.30	4.80	4.30	4.40	4.40	**7.80**
Partial Null Hypothesis							
ARH1							
c'	**1.30**	**1.30**	**1.90**	**1.30**	3.30	2.90	4.20
d'	**1.50**	**1.50**	**1.90**	**1.70**	3.10	3.10	4.50
RC							
c'	**1.10**	**1.10**	**1.50**	**1.10**	3.40	3.40	4.50
d'	**2.20**	**2.20**	2.60	**2.20**	4.90	4.40	6.10
UN							
c'	2.70	2.80	3.20	2.80	4.80	4.60	6.50
d'	2.70	2.70	3.50	2.70	5.30	5.00	7.20

Note. See note from Table 2.

Table 6

Empirical Type I Error Rates (%) when Fitting a UN-H Covariance Structure

(Normal Distribution, K = 4, N = 45, Default df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	5.10	5.10	5.80	5.10	5.10	5.10	7.50
d'	7.20	7.20	**8.40**	7.30	7.20	7.20	**11.20**
RC							
c'	5.30	5.40	5.90	5.40	5.30	5.30	**8.50**
d'	6.70	6.70	**7.90**	6.70	6.70	6.70	**12.90**
UN							
c'	4.40	4.40	5.00	4.40	4.40	4.50	7.40
d'	6.70	6.70	**7.80**	6.70	6.70	6.70	**10.70**
Partial Null Hypothesis							
ARH1							
c'	**2.30**	**2.30**	2.60	**2.30**	4.40	4.40	5.10
d'	3.50	3.60	4.30	3.60	6.50	5.70	7.40
RC							
c'	**1.80**	**1.80**	2.50	**1.80**	4.90	5.20	6.30
d'	3.80	3.80	4.50	3.80	6.80	6.70	**8.50**
UN							
c'	**2.30**	**2.30**	2.70	**2.30**	4.10	3.80	5.80
d'	3.20	3.20	3.80	3.20	6.00	5.40	7.40

Note. See note from Table 2.

Table 7

Empirical Type I Error Rates (%) when Fitting the True Covariance Structure

(Normal Distribution, K = 4, N = 45, Default df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	4.50	4.70	5.20	4.70	4.50	4.50	6.80
d'	4.90	4.90	5.40	4.90	4.90	5.00	**8.20**
RC							
c'	3.80	3.90	4.80	4.10	3.80	3.80	**7.80**
d'	4.20	4.30	4.80	4.30	4.20	4.20	**8.90**
UN							
c'	4.40	4.40	5.00	4.40	4.40	4.50	7.40
d'	6.70	6.70	**7.80**	6.70	6.70	6.70	**10.70**
Partial Null Hypothesis							
ARH1							
c'	**1.80**	**1.80**	**2.10**	**1.80**	3.70	3.70	4.50
d'	3.10	3.10	3.20	3.10	5.60	5.50	6.80
RC							
c'	**1.40**	**1.50**	**1.90**	**1.50**	4.50	4.40	5.40
d'	**2.40**	**2.40**	3.00	**2.40**	5.90	5.20	7.30
UN							
c'	**2.30**	**2.30**	2.70	**2.30**	4.10	3.80	5.80
d'	3.20	3.20	3.80	3.20	6.00	5.40	7.40

Note. See note from Table 2.

Table 8

Empirical Type I Error Rates (%) with Akaike Criterion Selecting the Best Covariance Structure (Normal Distribution, K = 4, N = 45, Default df)

Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
Complete Null Hypothesis							
ARH1							
c'	4.60	4.60	5.40	4.60	4.60	4.60	7.10
d'	4.10	4.20	5.10	4.20	4.10	4.20	**8.40**
RC							
c'	4.10	4.20	4.90	4.30	4.10	4.10	**8.10**
d'	4.60	4.70	5.30	4.70	4.60	4.60	**9.40**
UN							
c'	3.60	3.60	4.20	3.60	3.60	3.60	7.30
d'	6.20	6.20	7.30	6.20	6.20	6.20	**9.20**
Partial Null Hypothesis							
ARH1							
c'	**1.70**	**1.70**	**1.80**	**1.70**	3.60	3.40	4.60
d'	3.00	3.00	3.40	3.00	5.40	5.00	6.50
RC							
c'	**1.40**	**1.50**	**2.00**	**1.50**	4.90	4.50	6.10
d'	2.50	2.60	3.20	2.60	5.80	5.30	7.50
UN							
c'	2.50	2.50	3.30	2.50	6.10	5.40	**8.00**
d'	4.20	4.20	4.80	4.20	6.60	6.20	**8.10**

Note. See note from Table 2.

Table 9

Empirical Type I Error Rates (%) with Schwarz Criterion Selecting the Best Covariance Structure (Normal Distribution, K = 4, N = 45, Default df)

Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
Complete Null Hypothesis							
ARH1							
c'	3.80	3.80	4.60	3.80	3.80	3.80	6.50
d'	4.10	4.20	5.00	4.30	4.10	4.10	**8.10**
RC							
c'	2.90	3.00	3.60	3.10	2.90	2.90	4.70
d'	5.40	5.50	6.40	5.50	5.40	5.40	**11.00**
UN							
c'	3.40	3.40	4.10	3.40	3.40	3.40	6.60
d'	5.90	5.90	6.60	5.90	5.90	5.90	**8.50**
Partial Null Hypothesis							
ARH1							
c'	**1.60**	**1.60**	**1.70**	**1.60**	3.80	3.50	4.50
d'	2.80	2.90	3.50	3.00	5.70	5.40	6.70
RC							
c'	**0.70**	**0.80**	**1.10**	**0.80**	3.10	2.90	3.70
d'	2.70	2.80	3.50	2.80	6.60	5.90	**8.00**
UN							
c'	2.90	3.00	3.70	3.00	5.90	5.30	7.40
d'	4.50	4.50	5.00	4.50	6.50	5.90	**8.20**

Note. See note from Table 2.

Table 10

Percentage of Time Akaike Criterion Selected the Correct Covariance Structure

(rounded to whole numbers) (Normal Distribution, K = 4, N = 45, Satterthwaite df)

	Covariance Structure											
Cond	CSH	CSH-H	HF	HF-H	AR1	AR1-H	ARH1	ARH1-H	RC	RC-H	UN	UN-H
c'						81		12		4		
d'						78		12		8		
c'		2								92		3
d'		2								93		3
c'						56		15				13
d'						56		13				15
c'						81		10		5		
d'						80		11		6		
c'		3		2						91		2
d'		2		2						93		2
c'						52		16				12
d'						56		16				14

Note. Cond = Condition; Shading represents the true covariance structure; Only the three most frequently selected covariance structures for each condition are reported.

Table 11

Percentage of Time Schwarz Criterion Selected the Correct Covariance Structure (rounded to whole numbers) (Normal Distribution, K = 4, N = 45, Satterthwaite df)

	Covariance Structure											
Cond	CSH	CSH-H	HF	HF-H	AR1	AR1-H	ARH1	ARH1-H	RC	RC-H	UN	UN-H
c'					6	92	1					
d'					1	98				1		
c'	1								38	59		
d'						3			11	84		
c'					4	87	4					
d'						95		1		1		
c'					5	92	2					
d'					1	98				1		
c'						1			32	64		
d'						3			10	86		
c'					3	85	4					
d'						94				2	1	

Note. See note from Table 10.

The values from Tables 2 and 3 indicate that for both complete and partial null hypotheses always fitting a UN-H variance-covariance structure performs similarly to fitting the true variance-covariance structure for each of the population variance-covariance structures investigated (i.e., ARH1, RC, and UN) when df were based on Satterthwaite's solution. Under a complete null hypothesis, all MCPs control Type I error rates within Bradley's (1978) liberal criterion with the exception of Welsch's (1977) step-up range procedure. Under a partial null hypothesis, the simultaneous MCPs were typically conservative with Type I error rates as low as 0.70%, whereas the stepwise MCPs were well controlled with only the occasional conservative rate of 1.80% and 1.90% for the SRB and Hoch procedures, respectively when unequal groups sizes were positively paired with unequal variance-covariance matrices (condition c').

When multisample sphericity is violated, a test statistic that does not pool across between- and within-subjects factors and is based on Satterthwaite df is a robust approach to examine all possible pairwise comparisons among the levels of a RM factor using either a simultaneous or stepwise MCP. Furthermore, there does not appear to be an advantage to fitting the true variance-covariance structure compared to always fitting a UN-H structure to the data. For example, averaged across the investigated conditions, the SRB and Hoch procedures had error rates equal to 3.59% and 3.52%, respectively (see Table 3) when the true variance-covariance structure was fit to the data, while average error rates were 3.70% and 3.63%, respectively (see Table 2) when a UN-H structure was always fit to the data.

Tables 4 and 5 contain rates of error when the AIC and SBC criteria, respectively were used to select the best variance-covariance structure from among 12 possible structures when df were based on Satterthwaite's solution. Under a complete null hypothesis, relying on either of these two model selection criterion will generally provide good Type I error control for all MCPs except Welsch's procedure. On the other hand, given a partial null hypothesis, only the stepwise MCPs provided robust error rates across the investigated conditions.

Available through SAS' (1996, 1999) PROC MIXED is a nonpooled test statistic with df based on the default option. The error rates of MCPs based on this approach were investigated under violation of multisample sphericity to examine whether it is the form of the test statistic that provides robust procedures regardless of the estimation of the df.

Table 6 presents the Type I error rates of the MCPs based on the default df option available through PROC MIXED when always fitting a UN-H variance-covariance structure to the data. Under a complete null hypothesis, error control was within Bradley's (1978) limits except for the Tukey and Welsch procedures when unequal group sizes were negatively paired with unequal variance-covariance matrices (condition d'). The error rates averaged across population variance-covariance structures were 8.03% for the Tukey procedure and 11.60% for the Welsch procedure. Under a partial null hypothesis, empirical estimates of Type I error were conservative for the Bon, Sidak, and SMM procedures averaging 2.13% for the positively paired conditions (c') across population variance-covariance structures. A similar pattern of results is evident in Table 7 when the true variance-covariance structure was fit to the data and df were based on the default option. However, a notable difference was that liberal values for condition d' for

Tukey's procedure under ARH1 and RC population variance-covariance structures became robust when fitting the true variance-covariance structure rather than fitting a UN-H structure to the data. Furthermore, error rates in general were smaller when fitting the true population variance-covariance structure compared to always fitting a UN-H structure. For example, averaged across investigated conditions the SRB and Hoch procedures had error rates equal to 4.86% and 4.73%, respectively (see Table 7) when the true variance-covariance structure was fit to the data, while average error rates were 5.68% and 5.56%, respectively (see Table 6) when a UN-H structure was always fit to the data.

Based on default df, allowing either the AIC or SBC criterion (see Tables 8 and 9, respectively) to select the best variance-covariance structure from among 12 possible structures provided similar error control under complete and partial null hypotheses. Specifically, Type I error rates were well controlled with the following exceptions. Welsch's procedure had liberal rates as high as 11.00% under a complete null hypothesis and the simultaneous procedures under a partial null hypothesis had conservative rates ranging between 0.70% and 2.00% when unequal group sizes were positively paired with unequal variance-covariance matrices (condition c') for the ARH1 and RC population structures.

Although Type I error is generally controlled by allowing either the AIC or SBC criterion to select the best variance-covariance structure, their accuracy at picking the correct variance-covariance structure is typically poor. Tables 10 and 11 give percentages reflecting the frequency with which each criterion selected the correct variance-covariance structure from among 12 possible structures for each investigated condition. The shading in the table represents the true population variance-covariance structure for each condition. The rates were similar across the two df options investigated, therefore only the conditions based on Satterthwaite df solution were tabled. When the true population variance-covariance structure was RC-H, both the AIC criterion (see Table 10) and the SBC criterion (see Table 11) selected the correct structure with an accuracy rate between 91% to 93% (average=92%) and 59% to 86% (average=73%), respectively. However, when the true population variance-covariance structure was either ARH1-H or UN-H, both criteria selected the wrong variance-covariance structure with the greatest frequency. That is, a between-subjects heterogeneous version of a First-Order Autoregressive (AR1-H) structure was selected with rates between 52% to 81% (average=68%) for AIC and 85% to 98% (average=93%) for SBC. Interestingly, when the true population variance-covariance structure was ARH1-H or UN-H, the SBC criterion never selected the correct structure, whereas the AIC criterion selected the correct structure with an average accuracy rate of only 12%. The accuracy rates (%) for AIC and SBC were similar for complete and partial null hypotheses.

In summary, when normality is satisfied but multisample sphericity is violated in an unbalanced RM design, two MCPs were robust regardless of the method of determining df for the nonpooled pairwise test statistic. Averaged across investigated conditions, the error rates for SRB and Hoch when fitting a UN-H variance-covariance structure based on default df were 5.68% and 5.56%, respectively (see Table 6), whereas, the average error rates based on Satterthwaite df were 3.70% and 3.63%, respectively (see

Table 2). Furthermore, there was no Type I error advantage to fitting the true population variance-covariance structure compared to always fitting a UN-H structure to the data.

Lognormal Distributed Data. Table 12 contains the study conditions collected when data were obtained from a lognormal distribution for K = 4 and N = 45 with df based on Satterthwaite's solution. Type I error rates for the four testing strategies are contained in Tables 13 through 16, respectively. Fewer study conditions were examined when data were obtained from a lognormal distribution for K = 4 and N = 45 with df based on the default option and are given in Table 17 with Tables 18 through 21 containing error rates for the four testing strategies, respectively. Tables 22 and 23 contain percentages that the AIC and SBC criteria, respectively selected the best variance-covariance structure from among 12 possible structures. The empirical error rates from these investigated conditions provide information on the robustness properties of the MCPs when normality and multisample sphericity were violated separately and jointly.

Table 12

Study Conditions (Lognormal Distribution, K = 4, N = 45, Satterthwaite df)

Condition	Pop Cov Str	Sample Sizes	Cov Mat	Pairing	Null Hypothesis
a	ARH1	15,15,15	1:1:1	NA	$\mu_1=\mu_2=\mu_3=\mu_4=0$
b	ARH1-H	15,15,15	1:3:5	NA	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c	ARH1-H	12,15,18	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c'	ARH1-H	9,15,21	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d	ARH1-H	12,15,18	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d'	ARH1-H	9,15,21	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
a	RC	15,15,15	1:1:1	NA	$\mu_1=\mu_2=\mu_3=\mu_4=0$
b	RC-H	15,15,15	1:3:5	NA	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c	RC-H	12,15,18	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c'	RC-H	9,15,21	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d	RC-H	12,15,18	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d'	RC-H	9,15,21	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
a	UN	15,15,15	1:1:1	NA	$\mu_1=\mu_2=\mu_3=\mu_4=0$
b	UN-H	15,15,15	1:3:5	NA	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c	UN-H	12,15,18	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c'	UN-H	9,15,21	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d	UN-H	12,15,18	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d'	UN-H	9,15,21	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c'	ARH1-H	9,15,21	1:3:5	+	$\mu_1=\mu_2\neq\mu_3=\mu_4$
d'	ARH1-H	9,15,21	1:3:5	-	$\mu_1=\mu_2\neq\mu_3=\mu_4$
c'	RC-H	9,15,21	1:3:5	+	$\mu_1=\mu_2\neq\mu_3=\mu_4$
d'	RC-H	9,15,21	1:3:5	-	$\mu_1=\mu_2\neq\mu_3=\mu_4$
c'	UN-H	9,15,21	1:3:5	+	$\mu_1=\mu_2\neq\mu_3=\mu_4$
d'	UN-H	9,15,21	1:3:5	-	$\mu_1=\mu_2\neq\mu_3=\mu_4$

Note. See note from Table 1.

Table 13

Empirical Type I Error Rates (%) when Fitting a UN-H Covariance Structure

(Lognormal Distribution, K = 4, N = 45, Satterthwaite df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
a	3.80	3.90	5.10	4.20	3.80	3.90	7.30
b	3.00	3.20	4.40	3.20	3.10	3.10	7.10
c	2.80	2.90	4.10	3.10	2.90	3.00	7.30
c'	3.20	3.30	4.10	3.30	3.20	3.20	7.10
d	2.60	2.80	4.10	2.80	2.90	2.90	7.30
d'	3.30	3.30	4.70	3.60	3.40	3.50	**7.60**
RC							
a	4.40	4.40	5.60	4.60	4.30	4.30	**9.00**
b	4.90	5.00	6.60	5.10	5.10	5.10	**9.30**
c	5.60	5.60	7.20	5.60	5.60	5.60	**9.20**
c'	5.30	5.30	6.10	5.50	5.40	5.40	**9.20**
d	5.40	5.60	6.80	5.70	5.50	5.70	**9.50**
d'	6.30	6.30	7.30	6.50	6.50	6.60	**10.40**
UN							
a	3.80	3.80	4.40	3.90	3.80	3.80	7.20
b	3.80	3.80	4.10	3.90	3.60	3.60	7.10
c	**2.30**	2.50	3.10	2.60	**2.40**	**2.40**	6.50
c'	3.70	3.70	4.40	3.90	3.50	3.50	6.90
d	3.90	3.90	4.60	3.90	3.90	3.90	7.30
d'	3.30	3.30	4.30	3.70	3.50	3.50	6.40

Table 13 continued

Partial Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	**1.00**	**1.10**	**1.50**	**1.20**	2.90	2.80	3.80
d'	**0.90**	**1.00**	**1.70**	**1.20**	3.20	2.60	4.10
RC							
c'	**2.10**	**2.10**	2.60	**2.10**	4.60	4.40	5.50
d'	**1.40**	**1.40**	**1.90**	**1.50**	3.00	2.80	4.80
UN							
c'	**1.50**	**1.50**	**1.90**	**1.60**	3.10	3.00	4.20
d'	**1.00**	**1.00**	**1.40**	**1.10**	2.70	2.50	3.50

Note. See note from Table 2.

Table 14

Empirical Type I Error Rates (%) when Fitting the True Covariance Structure

(Lognormal Distribution, K = 4, N = 45, Satterthwaite df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
a	3.60	3.70	4.50	3.70	3.70	3.70	7.20
b	**2.30**	2.70	3.70	2.90	**2.40**	**2.40**	6.40
c	3.40	3.50	4.10	3.60	3.30	3.40	6.70
c'	3.10	3.20	3.90	3.20	3.20	3.20	6.40
d	2.70	2.70	3.30	2.70	2.80	2.80	5.50
d'	2.80	2.80	3.60	3.20	2.70	2.70	5.50
RC							
a	4.30	4.50	5.00	4.50	4.30	4.30	**8.70**
b	2.90	3.10	4.20	3.20	2.90	2.90	6.70
c	5.20	5.30	5.90	5.30	5.30	5.30	**8.20**
c'	3.80	3.90	4.50	4.00	3.70	3.80	7.00
d	5.30	5.30	6.10	5.40	5.10	5.40	**9.20**
d'	3.70	3.90	4.20	3.90	3.60	3.60	6.70
UN							
a	3.90	3.90	4.50	3.90	3.90	3.90	7.30
b	3.80	3.80	4.10	3.90	3.60	3.60	7.10
c	**2.30**	2.50	3.10	2.60	**2.40**	**2.40**	6.50
c'	3.70	3.70	4.40	3.90	3.50	3.50	6.90
d	3.90	3.90	4.60	3.90	3.90	3.90	7.30
d'	3.30	3.30	4.30	3.70	3.50	3.50	6.40

Table 14 continued

Partial Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	**0.90**	**1.00**	**1.30**	**1.00**	2.70	**2.40**	3.50
d'	**0.90**	**0.90**	**1.00**	**0.90**	2.60	**2.20**	3.50
RC							
c'	**1.90**	**1.90**	**2.10**	**2.00**	4.10	3.90	5.20
d'	**1.30**	**1.30**	**1.60**	**1.30**	3.30	2.60	4.30
UN							
c'	**1.50**	**1.50**	**1.90**	**1.60**	3.10	3.00	4.20
d'	**1.00**	**1.00**	**1.40**	**1.10**	2.70	2.50	3.50

Note. See note from Table 2.

Table 15

Empirical Type I Error Rates (%) with Akaike Criterion Selecting the Best Covariance Structure (Lognormal Distribution, K = 4, N = 45, Satterthwaite df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
a	4.20	4.40	5.10	4.40	4.30	4.30	**8.20**
b	2.70	2.90	4.10	3.10	2.80	2.80	6.60
c	3.60	3.70	4.50	3.80	3.60	3.70	7.40
c'	3.10	3.20	4.10	3.30	3.20	3.20	6.50
d	3.20	3.20	4.20	3.20	3.40	3.40	7.00
d'	3.50	3.50	4.60	3.80	3.70	3.70	6.00
RC							
a	4.90	5.00	6.00	5.10	4.90	4.90	**9.50**
b	4.20	4.50	5.50	4.50	4.10	4.10	**8.20**
c	5.50	5.60	6.80	5.60	5.70	5.70	**8.60**
c'	4.70	4.80	5.60	4.90	4.80	4.80	**7.80**
d	6.20	6.20	6.80	6.30	6.20	6.20	**9.60**
d'	5.10	5.10	5.70	5.20	5.30	5.30	**9.30**
UN							
a	3.30	3.30	3.90	3.30	3.30	3.30	6.90
b	3.60	3.60	4.40	3.60	3.50	3.50	**7.70**
c	2.80	2.90	3.60	3.00	2.90	2.90	6.00
c'	3.30	3.30	4.00	3.40	3.00	3.10	6.50
d	3.50	3.50	4.40	3.60	3.60	3.60	7.40
d'	3.70	3.80	4.40	3.90	3.80	3.80	6.60

Table 15 continued

Partial Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	**0.80**	**0.90**	**1.30**	**1.00**	**2.30**	**2.40**	3.50
d'	**0.90**	**0.90**	**1.20**	**0.90**	2.80	**2.30**	3.90
RC							
c'	**2.10**	**2.10**	2.50	**2.10**	4.50	4.20	5.10
d'	**1.60**	**1.60**	**2.20**	**1.70**	3.00	2.80	4.30
UN							
c'	**2.30**	**2.30**	2.50	**2.30**	3.90	3.60	4.70
d'	**1.20**	**1.20**	**1.70**	**1.20**	3.70	3.10	5.20

Note. See note from Table 2.

Table 16

Empirical Type I Error Rates (%) with Schwarz Criterion Selecting the Best Covariance Structure (Lognormal Distribution, K = 4, N = 45, Satterthwaite df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
a	3.80	4.00	4.50	4.00	4.00	4.00	7.50
b	3.20	3.30	4.00	3.30	3.20	3.20	6.40
c	3.60	3.60	4.60	3.70	3.60	3.60	7.40
c'	2.80	2.90	3.70	3.00	2.80	2.80	5.90
d	3.00	3.00	3.60	3.00	3.10	3.10	6.70
d'	3.60	3.60	4.50	3.70	3.70	3.70	6.70
RC							
a	4.60	4.70	5.00	4.70	4.50	4.50	**8.90**
b	3.40	3.80	4.50	3.90	3.60	3.60	7.40
c	4.90	5.00	5.60	5.00	4.70	4.70	**7.80**
c'	4.10	4.30	5.00	4.30	4.10	4.20	6.40
d	5.90	6.00	7.20	6.10	5.90	6.10	**10.00**
d'	4.40	4.70	5.10	4.70	4.70	4.70	**8.40**
UN							
a	3.10	3.10	4.00	3.10	3.10	3.10	6.60
b	3.30	3.30	4.50	3.40	3.40	3.40	**7.80**
c	3.60	3.80	4.40	3.80	3.60	3.60	6.50
c'	3.00	3.20	3.90	3.20	2.80	2.90	6.70
d	3.20	3.20	3.80	3.30	3.30	3.30	6.80
d'	3.90	4.00	4.60	4.00	3.90	3.90	6.50

Table 16 continued

Partial Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	**0.80**	**0.90**	**1.10**	**1.00**	**2.30**	2.50	3.60
d'	**1.40**	**1.40**	**1.70**	**1.40**	3.30	2.90	4.50
RC							
c'	**2.10**	**2.10**	**2.40**	**2.20**	3.80	3.70	4.50
d'	**1.90**	**1.90**	**2.40**	**1.90**	3.90	3.50	5.00
UN							
c'	**2.10**	**2.10**	**2.30**	**2.10**	3.90	3.40	4.50
d'	**2.00**	**2.00**	2.80	**2.10**	4.80	4.20	6.40

Note. See note from Table 2.

Table 17

Study Conditions (Lognormal Distribution, $K = 4$, $N = 45$, Default df)

Condition	Pop Cov Str[a]	Sample Sizes	Cov Mat[b]	Pairing	Null Hypothesis
c'	ARH1-H	9,15,21	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d'	ARH1-H	9,15,21	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c'	RC-H	9,15,21	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d'	RC-H	9,15,21	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c'	UN-H	9,15,21	1:3:5	+	$\mu_1=\mu_2=\mu_3=\mu_4=0$
d'	UN-H	9,15,21	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=0$
c'	ARH1-H	9,15,21	1:3:5	+	$\mu_1=\mu_2\neq\mu_3=\mu_4$
d'	ARH1-H	9,15,21	1:3:5	-	$\mu_1=\mu_2\neq\mu_3=\mu_4$
c'	RC-H	9,15,21	1:3:5	+	$\mu_1=\mu_2\neq\mu_3=\mu_4$
d'	RC-H	9,15,21	1:3:5	-	$\mu_1=\mu_2\neq\mu_3=\mu_4$
c'	UN-H	9,15,21	1:3:5	+	$\mu_1=\mu_2\neq\mu_3=\mu_4$
d'	UN-H	9,15,21	1:3:5	-	$\mu_1=\mu_2\neq\mu_3=\mu_4$

Note. See note from Table 1.

Table 18

Empirical Type I Error Rates (%) when Fitting a UN-H Covariance Structure

(Lognormal Distribution, K = 4, N = 45, Default df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	3.80	3.90	4.10	4.00	3.80	3.80	6.80
d'	5.80	5.80	7.20	5.80	5.80	5.80	**10.90**
RC							
c'	6.80	6.80	**7.90**	6.90	6.80	6.80	**11.20**
d'	**8.70**	**8.70**	**9.20**	**8.70**	**8.70**	**8.70**	**14.00**
UN							
c'	4.30	4.50	5.00	4.50	4.30	4.30	**8.70**
d'	4.70	4.80	5.60	4.80	4.70	4.70	**9.80**
Partial Null Hypothesis							
ARH1							
c'	**1.60**	**1.70**	**1.70**	**1.70**	3.50	3.50	4.90
d'	2.70	2.80	3.30	2.80	5.70	5.60	7.30
RC							
c'	**2.20**	**2.20**	2.50	**2.20**	5.20	5.00	6.20
d'	4.30	4.40	4.60	4.40	7.50	7.10	**9.10**
UN							
c'	**1.60**	**1.70**	**1.90**	**1.70**	5.30	4.50	7.20
d'	2.90	2.90	3.40	2.90	5.70	5.40	7.40

Note. See note from Table 2.

Table 19

Empirical Type I Error Rates (%) when Fitting the True Covariance Structure

(Lognormal Distribution, K = 4, N = 45, Default df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	2.70	2.80	3.70	2.80	2.70	2.70	6.70
d'	3.70	3.80	4.80	3.90	3.70	3.70	**8.10**
RC							
c'	4.90	5.00	5.60	5.00	4.90	4.90	**8.70**
d'	6.00	6.10	7.10	6.20	6.00	6.00	**9.80**
UN							
c'	4.30	4.50	5.00	4.50	4.30	4.30	**8.70**
d'	4.70	4.80	5.60	4.80	4.70	4.70	**9.80**
Partial Null Hypothesis							
ARH1							
c'	**1.70**	**1.70**	**2.00**	**1.70**	3.40	3.60	4.20
d'	**1.90**	**1.90**	**2.30**	**1.90**	4.00	3.80	5.40
RC							
c'	**1.70**	**1.80**	**2.00**	**1.80**	4.50	4.50	5.50
d'	2.90	2.90	3.40	2.90	5.30	4.90	6.70
UN							
c'	**1.60**	**1.70**	**1.90**	**1.70**	5.30	4.50	7.20
d'	2.90	2.90	3.40	2.90	5.70	5.40	7.40

Note. See note from Table 2.

Table 20

Empirical Type I Error Rates (%) with Akaike Criterion Selecting the Best Covariance Structure (Lognormal Distribution, K = 4, N = 45, Default df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	2.60	2.80	3.60	2.80	2.60	2.70	6.40
d'	4.40	4.50	5.50	4.50	4.40	4.40	**9.20**
RC							
c'	6.10	6.20	6.70	6.20	6.10	6.20	**9.60**
d'	7.40	7.40	**8.50**	7.40	7.40	7.40	**11.90**
UN							
c'	4.10	4.40	4.90	4.50	4.10	4.20	**9.00**
d'	4.50	4.50	5.40	4.50	4.50	4.50	**8.70**
Partial Null Hypothesis							
ARH1							
c'	**1.40**	**1.50**	**2.00**	**1.60**	3.40	3.40	4.10
d'	**2.00**	**2.00**	**2.20**	**2.00**	4.50	4.40	6.00
RC							
c'	**2.20**	**2.20**	2.50	**2.20**	4.90	5.10	6.10
d'	3.60	3.60	4.10	3.60	6.40	6.10	**7.80**
UN							
c'	**1.80**	**1.80**	**2.10**	**1.80**	4.80	3.80	7.10
d'	2.80	2.80	3.20	2.80	6.20	5.60	**7.90**

Note. See note from Table 2.

Table 21

Empirical Type I Error Rates (%) with Schwarz Criterion Selecting the Best Covariance Structure (Lognormal Distribution, K = 4, N = 45, Default df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	**2.40**	2.50	3.60	2.50	**2.40**	**2.40**	6.30
d'	4.40	4.40	5.20	4.50	4.40	4.40	**9.20**
RC							
c'	4.50	4.50	5.00	4.50	4.50	4.50	**8.00**
d'	6.50	6.60	**8.30**	6.60	6.50	6.50	**10.80**
UN							
c'	4.20	4.30	4.70	4.40	4.20	4.40	7.40
d'	4.00	4.00	4.50	4.00	4.00	4.00	7.20
Partial Null Hypothesis							
ARH1							
c'	**1.10**	**1.10**	**1.70**	**1.20**	3.60	3.60	4.60
d'	**2.20**	**2.20**	2.60	**2.20**	5.80	4.90	7.10
RC							
c'	**1.50**	**1.60**	**1.80**	**1.60**	3.80	4.00	5.00
d'	3.40	3.40	3.80	3.40	5.90	5.60	7.30
UN							
c'	**2.40**	**2.40**	2.70	**2.40**	4.50	4.20	6.70
d'	3.90	3.90	4.50	3.90	7.00	6.30	**8.40**

Note. See note from Table 2.

Table 22

Percentage of Time Akaike Criterion Selected the Correct Covariance Structure (rounded to whole numbers) (Lognormal Distribution, K = 4, N = 45, Satterthwaite df)

	Covariance Structure											
Cond	CSH	CSH-H	HF	HF-H	AR1	AR1-H	ARH1	ARH1-H	RC	RC-H	UN	UN-H
a							19	32				13
b						29		44				18
c						30		45				17
c'						28		41				19
d						27		44				18
d'						22		42				23
a									16	28		30
b		8								48		35
c		8								46		35
c'		8								46		35
d		7								46		38
d'		9								47		36
a								20			29	26
b						13		31				38
c						15		27				41
c'						12		28				43
d						13		28				41
d'						11		29				43
c'						29		41				18
d'						24		44				20
c'		10								46		32
d'		7								45		37
c'						14		29				39
d'						13		29				39

Note. See note from Table 10.

Table 23

Percentage of Time Schwarz Criterion Selected the Correct Covariance Structure (rounded to whole numbers) (Lognormal Distribution, K = 4, N = 45, Satterthwaite df)

	Covariance Structure											
Cond	CSH	CSH-H	HF	HF-H	AR1	AR1-H	ARH1	ARH1-H	RC	RC-H	UN	UN-H
a					49	11	**25**					
b					5	71		**14**				
c					6	69		**16**				
c'					9	66		**13**				
d					6	69		**15**				
d'						67		**18**		4		
a	6								**65**	14		
b		6							13	**67**		
c		6							15	**67**		
c'		5							17	**65**		
d		5							12	**70**		
d'		5							13	**71**		
a					29		26				**15**	
b		5				56		20			5	
c						57		18			5	
c'						55		14			6	
d		6				59		17				
d'		5				57		18				
c'					7	70		**13**				
d'						66		**17**		6		
c'		6							21	**59**		
d'		6							13	**68**		
c'						52	9	14				
d'						58	4	18				

Note. See note from Table 10.

Table 13 contains Type I error rates when a UN-H variance-covariance structure was always fit to the data and df were based on Satterthwaite's solution. Rates of error for each MCP were generally well controlled under a complete null hypothesis with the following exceptions: the Welsch procedure under a ARH1 variance-covariance structure with a rate of 7.60% (condition d'), the Welsch procedure under a RC variance-covariance structure with rates ranging between 9.00% and 10.40%, and the Bon, SRB, and Hoch procedures under a UN variance-covariance structure with rates of 2.30%, 2.40%, and 2.40%, respectively for a positive pairing of unequal variance-covariance matrices and unequal group sizes (condition c). Under a partial null hypothesis only the SRB, Hoch, and Welsch procedures were able to control Type I error with rates ranging between 2.50% and 5.50%, whereas the simultaneous MCPs typically had conservative rates ranging between 0.90% and 2.60%.

Rates of Type I error control when the true population variance-covariance structure was fit to the data and df were based on Satterthwaite's solution are contained in Table 14. The tendency across conditions was for rates to be slightly smaller when fitting the true population variance-covariance structure compared to rates when a UN-H structure was always fit to the data. Under a complete null hypothesis all MCPs generally provided error rates within Bradley's (1978) limits with the occasional liberal value for the Welsch procedure and the occasional conservative value for the Bon, SRB, and Hoch procedures. While under a partial null hypothesis only the three stepwise MCPs provided robust error control across investigated conditions.

Once again, there was no Type I error advantage to fitting the true variance-covariance structure compared to always fitting a UN-H structure. For example, averaged across investigated conditions the SRB and Hoch procedures had error rates equal to 3.43% and 3.37%, respectively (see Table 14) when the true variance-covariance structure was fit to the data, while average error rates were 3.83% and 3.80%, respectively (see Table 13) when a UN-H structure was always fit to the data.

Allowing either the AIC or SBC criterion to choose the best variance-covariance structure (see Table 15 and 16, respectively) provided similar error control across the investigated MCPs. Under a complete null hypothesis all procedures except Welsch, were able to maintain error rates within Bradley's (1978) limits, however only the stepwise procedures were robust under a partial null hypothesis with an occasional conservative rate of 2.30% for the SRB and Hoch procedures.

When the df were based on the default option available through PROC MIXED, rates of error based on always fitting a UN-H variance-covariance structure to the data (see Table 18) were not well controlled across all population variance-covariance structures. Specifically, under a complete null hypothesis, rates were liberal for all MCPs when the population variance-covariance structure was RC and unequal group sizes were negatively paired with unequal variance-covariance matrices (e.g., rates ranged between 8.70% to 14.00% across the MCPs for condition d'). Empirical rates of Type I error improved when fitting the true variance-covariance structure to the data (see Table 19), however the Welsch procedure remained nonrobust under a complete null hypothesis. Similar to the results based on Satterthwaite df, only the stepwise MCPs provided robust error rates under a partial null hypothesis across investigated conditions. It is important to note that the SRB and Hoch procedures based on default df were only robust across all conditions when the true population variance-covariance structure was fit to the data.

Relying on either the AIC or SBC criterion (see Tables 20 and 21, respectively) to select the best variance-covariance structure provided similar rates of error when df were based on the default option. Only the SRB and Hoch procedures provided empirical error rates within Bradley's limits for complete and partial null hypotheses across the population variance-covariance structures investigated.

Tables 22 and 23 contain the frequency with which the AIC and SBC criteria, respectively selected the correct variance-covariance structure from among 12 possible structures. The values were similar regardless of the choice of df option, therefore only the conditions where Satterthwaite df were used are tabled. The AIC criterion (see Table 22) selected the correct variance-covariance structure with the greatest frequency across most conditions, however the percentages ranged from 16% to 48% with an average accuracy rate of only 41%. In comparison, the SBC criterion (see Table 23) selected the correct variance-covariance structure with rates between 59% to 71% when the true population structure was RC but when the true population structure was either ARH1 or UN, the SBC criterion selected the wrong variance-covariance structure (i.e., AR1 or AR1-H) with the greatest frequency (i.e., rates between 29% to 71%).

A limited number of conditions were examined when K = 8 to determine whether the results for the MCPs based on the four testing strategies for K = 4 would extend to a larger number of levels of the RM factor (i.e., an increase in the number of levels of the RM factor from 4 to 8 increases the number of pairwise tests from 6 to 28). The total sample size was increased to N = 60 from N = 45 and only the negative pairing of unequal group sizes with unequal variance-covariance matrices were examined for the more disparate sample size case when the data was lognormally distributed (see Appendix B for Tables B1 to B11). The results indicate that all MCPs except Welsch's procedure had error rates within Bradley's (1978) limits for a complete null hypothesis when always fitting a UN-H structure with Satterthwaite df (see Table B2). When the true population structure was fit to the data, the error rates became smaller and under certain conditions conservative (see Table B3). Relying on either the AIC or SBC criterion to select the best variance-covariance structure provided similar robust error control except for the occasional conservative rate for the SRB and Hoch procedures and liberal error rates for the Welsch procedure across population structures (see Tables B4 and B5). The accuracy with which the AIC criterion selected the correct variance-covariance structure improved to an average of 74% across the investigated conditions (see Table B6), while the SBC criterion was only highly accurate when the true population structure was RC-H (i.e., 95%) and otherwise selected the wrong structure with the greatest frequency (see Table B7). Consistent with the K = 4 results is the tendency for error rates for MCPs to become smaller under a partial null hypothesis regardless of the testing strategy adopted and in particular conservative for the simultaneous MCPs. The pattern of results when using the default df option with K = 8 is similar to when K = 4 (see Tables B8 to B11). That is, error rates were liberal for all MCPs when fitting a UN-H structure to the data for a RC population variance-covariance structure and improved when fitting the true population structure. In general, the results for K = 4 extend to K = 8 with the exception of the need for a larger total sample size in order for the convergence criteria to be met for the REML estimation.

Power Rates

To simulate power rates for the MCPs, three nonnull mean configurations were examined. That is, a minimum range configuration, a maximum range configuration, and an equally spaced range configuration. Effect sizes varied between 0.50 and 1.25 and were selected to avoid floor and ceiling effects (see Appendix C for population means). All MCPs were compared in terms of average per-pair and all-pairs power with the testing strategies of always fitting a UN-H variance-covariance structure or the true variance-covariance structure. Guidelines provided by Einot and Gabriel (1975) were used to evaluate differences in power values across procedures. That is, power differences greater than 20% were considered substantial while those less than 10% were considered negligible.

Normally Distributed Data. Table 24 contains all-pairs and average per-pair power rates for data obtained from a normal distribution for K = 4 and N = 45 for a minimum range configuration when df were based on Satterthwaite's solution and a UN-H variance-covariance structure was always fit to the data. As expected, the stepwise MCPs were more powerful than the simultaneous MCPs, however this power advantage was only negligible (i.e., less than 10%). For all-pairs power, the average power rates across investigated conditions for the simultaneous MCPs (Bon, Sidak, Tukey, and SMM) were 35.25%, 35.58%, 39.08%, and 36.05%, respectively and for the stepwise MCPs (SRB, Hoch, and Welsch) were 45.07%, 45.28%, and 59.57%, respectively. For average per-pair power, the average power rates across investigated conditions for the simultaneous and stepwise MCPs (Bon, Sidak, Tukey, SMM, SRB, Hoch, and Welsch) were 64.72%, 65.02%, 68.11%, 65.53%, 71.49%, 69.41%, and 78.83%, respectively. Although the Welsch procedure was most powerful, the result is illusory because this procedure had liberal error rates under a complete null hypothesis. When the true population variance-covariance structure was fit to the data, the MCPs were more powerful, however this power advantage never exceeded 2 percentage points (see Table 25). The rates averaged across investigated conditions for the simultaneous and stepwise MCPs were 37.42%, 37.78%, 41.08%, 38.05%, 46.37%, 46.52%, and 59.80%, respectively for all-pairs power and 66.02%, 66.27%, 69.11%, 66.66%, 71.07%, 69.28%, and 78.82%, respectively for average per-pair power.

Power rates for MCPs based on default df are contained in Table 26 when always fitting a UN-H variance-covariance structure and Table 27 when fitting the true population variance-covariance structure to the data. The power advantage of fitting the correct variance-covariance structure compared to always fitting a UN-H structure never exceeded 2 percentage points when the rates were averaged across conditions separately for all-pairs and average per-pair power. Specifically, the average across conditions for all-pairs power for the simultaneous and stepwise MCPs when a UN-H variance-covariance structure was fit to the data were 39.45%, 39.75%, 42.27%, 39.80%, 50.13%, 50.30%, and 62.77%, respectively and when the true variance-covariance structure was fit to the data the average rates were 40.92%, 41.25%, 43.58%, 41.32%, 51.17%, 51.28%, and 62.80%, respectively. For average per-pair power, the average across conditions

Table 24

Power Rates (Minimum Mean Configuration) when Fitting a UN-H Covariance Structure

(Normal Distribution K = 4, N = 45, Satterthwaite df)

All-Pairs Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	49.90	50.10	53.80	50.30	61.30	61.50	72.50
d'	26.60	26.70	29.80	27.70	34.80	35.00	49.50
RC							
c'	52.80	53.60	57.30	53.90	63.40	63.70	76.20
d'	24.10	24.40	28.20	25.10	33.60	33.70	48.60
UN							
c'	41.80	42.00	44.60	42.30	51.70	51.90	66.20
d'	16.30	16.70	20.80	17.00	25.60	25.90	44.40

Average Per-Pair Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	76.58	76.78	79.18	76.93	83.13	81.58	87.48
d'	53.70	54.00	57.88	54.93	61.15	58.30	70.55
RC							
c'	79.78	80.10	82.08	80.33	85.28	84.10	89.93
d'	54.03	54.50	58.75	55.35	61.70	59.15	70.98
UN							
c'	72.98	73.23	75.03	73.38	78.90	77.53	85.05
d'	51.23	51.50	55.75	52.28	58.78	55.80	69.00

Note. c' = positive pairing of unequal covariance matrices and unequal group sizes; d' = negative pairing of unequal covariance matrices and unequal group sizes (n_j = 9, 15, 21); Bon = Bonferroni; SMM = Studentized maximum modulus; SRB = Shaffer's (1986) sequentially rejective Bonferroni; Hoch = Hochberg's (1988) sequentially acceptive Bonferroni.

Table 25

Power Rates (Minimum Mean Configuration) when Fitting the True Covariance Structure (Normal Distribution K = 4, N = 45, Satterthwaite df)

All-Pairs Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	53.10	53.50	57.00	53.60	64.40	64.50	74.30
d'	28.10	28.80	32.10	29.20	40.00	40.10	51.00
RC							
c'	56.70	56.70	59.70	56.80	65.20	65.30	76.40
d'	28.50	29.00	32.30	29.40	31.30	31.40	46.50
UN							
c'	41.80	42.00	44.60	42.30	51.70	51.90	66.20
d'	16.30	16.70	20.80	17.00	25.60	25.90	44.40

Average Per-Pair Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	78.58	78.85	80.88	78.95	84.65	83.63	88.78
d'	55.68	56.05	59.33	56.53	64.05	61.68	71.33
RC							
c'	80.90	81.03	82.88	81.15	85.53	84.38	90.15
d'	56.73	56.98	60.78	57.68	54.50	52.63	68.63
UN							
c'	72.98	73.23	75.03	73.38	78.90	77.53	85.05
d'	51.23	51.50	55.75	52.28	58.78	55.80	69.00

Note. See note from Table 24.

Table 26

Power Rates (Minimum Mean Configuration) when Fitting a UN-H Covariance Structure

(Normal Distribution K = 4, N = 45, Default df)

All-Pairs Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	53.20	53.70	56.80	53.70	66.40	66.50	74.00
d'	30.20	30.40	33.40	30.50	42.70	42.70	55.40
RC							
c'	57.50	57.70	59.80	57.70	66.60	66.80	75.20
d'	31.30	31.50	33.40	31.50	40.30	40.70	56.40
UN							
c'	42.80	43.30	45.90	43.50	54.60	54.80	67.10
d'	21.70	21.90	24.30	21.90	30.20	30.30	48.50

Average Per-Pair Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	77.73	78.00	79.95	78.03	84.23	83.08	87.93
d'	61.38	61.45	63.78	61.48	68.85	66.60	75.48
RC							
c'	82.28	82.40	83.63	82.40	86.80	85.88	89.73
d'	63.45	63.70	65.80	63.78	70.00	68.20	77.23
UN							
c'	74.85	75.05	77.00	75.13	81.23	79.88	86.08
d'	58.15	58.45	60.43	58.48	65.00	62.80	73.70

Note. See note from Table 24.

Table 27

Power Rates (Minimum Mean Configuration) when Fitting the True Covariance Structure

(Normal Distribution K = 4, N = 45, Default df)

All-Pairs Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	56.90	57.20	59.40	57.30	66.20	66.30	74.40
d'	33.40	33.80	36.50	33.80	46.30	46.30	56.80
RC							
c'	57.60	57.90	59.50	58.00	67.30	67.40	75.20
d'	33.10	33.40	35.90	33.40	42.40	42.60	54.80
UN							
c'	42.80	43.30	45.90	43.50	54.60	54.80	67.10
d'	21.70	21.90	24.30	21.90	30.20	30.30	48.50

Average Per-Pair Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c'	79.35	79.48	81.08	79.58	84.65	83.28	88.30
d'	63.08	63.40	65.63	63.48	70.80	68.73	76.28
RC							
c'	81.88	82.00	83.38	82.13	86.40	85.68	89.40
d'	63.10	63.40	65.28	63.45	69.60	67.40	76.30
UN							
c'	74.85	75.05	77.00	75.13	81.23	79.88	86.08
d'	58.15	58.45	60.43	58.48	65.00	62.80	73.70

Note. See note from Table 24.

when fitting a UN-H variance-covariance structure for each of the MCPs were 69.64%, 69.84%, 71.77%, 69.88%, 76.02%, 74.41%, and 81.69%, respectively and when fitting the true variance-covariance structure, the average rates were 70.07%, 70.30%, 72.13%, 70.38%, 76.28%, 74.63%, and 81.68%, respectively. The MCPs were more powerful when the df were based on the default option, however the difference in average rates across conditions was negligible (i.e., never greater than 6 percentage points) compared to average power rates for MCPs based on Satterthwaite df.

Lognormal Distributed Data. All-pairs and average per-pair power rates for data obtained from a lognormal distribution for K = 4 and N = 45 for a minimum, maximum, and equally spaced mean range configuration were investigated. Tables 28 and 29 contain all-pairs and average per-pair power rates, respectively when the MCPs were based on Satterthwaite df and a UN-H variance-covariance structure was always fit to the data and Tables 30 and 31, contain power rates when the true variance-covariance structure was fit to the data for all-pairs and average per-pair power rates, respectively. Because of the lack of robustness for MCPs based on default df for the combined violation of normality and homogeneity of variance-covariance matrices, power rates are not reported.

Table 28

All-Pairs Power Rates when Fitting a UN-H Covariance Structure

(Lognormal Distribution, K = 4, N = 45, Satterthwaite df)

All-Pairs Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c' (min)	53.30	53.60	56.80	54.00	62.70	62.90	73.40
d' (min)	28.20	28.50	33.20	29.50	38.10	38.10	53.40
c' (max)	49.80	50.40	53.40	50.70	68.00	68.50	70.60
d' (max)	18.10	18.40	22.40	19.20	34.60	36.10	42.10
c' (eq)	37.20	37.60	42.00	38.00	70.40	71.60	59.10
d' (eq)	9.40	9.60	12.90	10.40	35.50	38.30	23.10
RC							
c' (min)	56.30	56.70	60.40	56.90	70.00	70.00	82.40
d' (min)	22.20	22.40	27.30	23.20	33.40	33.40	57.20
c' (max)	83.10	83.70	85.90	84.30	94.40	94.90	94.40
d' (max)	41.50	42.10	47.90	43.30	63.60	65.60	68.70
c' (eq)	18.50	18.90	23.40	19.30	60.70	62.70	44.10
d' (eq)	3.30	3.30	5.60	3.50	25.70	29.80	15.10
UN							
c' (min)	38.80	39.20	44.20	39.60	52.90	52.90	66.80
d' (min)	20.90	21.00	24.10	21.50	29.30	29.40	47.90
c' (max)	36.20	36.40	40.30	36.60	54.50	55.70	56.40
d' (max)	12.30	12.50	15.10	13.00	25.80	27.00	28.40
c' (eq)	17.70	17.90	20.30	18.10	50.70	51.70	36.10
d' (eq)	5.10	5.20	7.10	5.60	23.40	25.30	14.60

Note. c' = positive pairing of unequal covariance matrices and unequal group sizes; d' = negative pairing of unequal covariance matrices and unequal group sizes (n_j = 9, 15, 21); min = minimum, max = maximum, eq= equally spaced mean configuration; Bon = Bonferroni; SMM = Studentized maximum modulus; SRB = Shaffer's (1986) sequentially rejective Bonferroni; Hoch = Hochberg's (1988) sequentially acceptive Bonferroni.

Table 29

Average Per-Pair Power Rates when Fitting a UN-H Covariance Structure

(Lognormal Distribution, K = 4, N = 45, Satterthwaite df)

Average Per-Pair Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c' (min)	76.78	76.93	79.38	77.13	82.58	81.23	87.25
d' (min)	56.40	56.60	60.78	57.60	63.40	61.13	72.70
c' (max)	84.74	84.90	86.44	84.98	90.52	90.60	92.02
d' (max)	64.88	65.16	68.72	65.86	73.06	72.30	78.62
c' (eq)	85.55	85.72	87.25	85.88	93.77	94.20	92.07
d' (eq)	68.07	68.32	71.17	69.02	78.68	79.68	79.05
RC							
c' (min)	82.88	83.05	85.30	83.25	89.35	88.33	93.35
d' (min)	58.63	59.05	64.40	60.18	68.00	64.63	79.53
c' (max)	96.02	96.20	96.76	96.32	98.72	98.86	98.80
d' (max)	80.92	81.24	84.42	81.96	88.88	89.32	91.60
c' (eq)	80.20	80.35	82.35	80.60	91.45	92.23	88.90
d' (eq)	64.30	64.43	68.05	65.27	76.17	77.62	76.67
UN							
c' (min)	72.83	73.08	75.93	73.40	80.40	78.55	85.90
d' (min)	53.73	53.98	57.58	54.63	60.18	58.05	71.28
c' (max)	78.24	78.34	80.34	78.52	84.96	85.20	86.30
d' (max)	60.48	60.74	64.12	61.54	68.80	67.90	72.88
c' (eq)	78.75	78.98	80.52	79.08	88.40	88.75	86.30
d' (eq)	63.17	63.37	66.42	64.05	73.40	73.78	74.85

Note. See note from Table 28.

Table 30

All-Pairs Power Rates when Fitting the True Covariance Structure

(Lognormal Distribution, K = 4, N = 45, Satterthwaite df)

All-Pairs Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c' (min)	56.10	56.30	58.50	56.50	66.80	66.90	75.10
d' (min)	32.70	33.00	35.90	33.20	42.20	42.30	55.60
c' (max)	53.00	53.30	56.60	53.40	71.60	72.30	73.40
d' (max)	23.70	23.90	27.00	24.30	41.50	42.60	46.30
c' (eq)	39.90	40.40	44.60	41.00	74.00	75.00	61.50
d' (eq)	12.60	12.80	16.40	13.30	39.80	42.20	27.50
RC							
c' (min)	58.60	58.80	61.30	59.00	67.00	67.20	79.90
d' (min)	27.40	27.60	31.80	27.90	35.20	35.40	54.00
c' (max)	82.60	83.10	86.20	83.30	91.90	92.40	92.80
d' (max)	47.10	47.40	52.20	48.20	64.90	65.90	71.30
c' (eq)	20.80	21.30	25.30	21.40	60.80	61.90	45.30
d' (eq)	4.90	5.30	6.70	5.30	27.30	28.90	17.80
UN							
c' (min)	38.80	39.20	44.20	39.60	52.90	52.90	66.80
d' (min)	20.90	21.00	24.10	21.50	29.30	29.40	47.90
c' (max)	36.20	36.40	40.30	36.60	54.50	55.70	56.40
d' (max)	12.30	12.50	15.10	13.00	25.80	27.00	28.40
c' (eq)	17.70	17.90	20.30	18.10	50.70	51.70	36.10
d' (eq)	5.10	5.20	7.10	5.60	23.40	25.30	14.60

Note. See note from Table 28.

Table 31

Average Per-Pair Power Rates when Fitting the True Covariance Structure

(Lognormal Distribution, K = 4, N = 45, Satterthwaite df)

Average Per-Pair Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
c' (min)	78.80	78.95	80.80	79.13	84.43	83.30	88.23
d' (min)	59.35	59.60	63.13	60.00	67.30	64.75	74.50
c' (max)	86.30	86.42	87.64	86.46	91.74	91.88	92.94
d' (max)	68.36	68.64	71.12	68.96	76.88	76.08	80.74
c' (eq)	86.68	86.83	88.23	86.97	94.47	94.82	92.52
d' (eq)	71.43	71.58	73.83	71.88	81.62	82.33	80.97
RC							
c' (min)	83.45	83.70	85.25	83.88	87.80	86.73	92.70
d' (min)	62.40	62.63	66.18	63.10	64.78	62.63	76.88
c' (max)	95.28	95.42	96.46	95.64	96.90	96.98	98.40
d' (max)	82.38	82.68	85.42	83.32	83.30	83.50	91.48
c' (eq)	81.00	81.15	82.93	81.47	90.80	91.20	89.12
d' (eq)	65.78	66.18	69.32	66.83	72.40	72.85	77.10
UN							
c' (min)	72.83	73.08	75.93	73.40	80.40	78.55	85.90
d' (min)	53.73	53.98	57.58	54.63	60.18	58.05	71.28
c' (max)	78.24	78.34	80.34	78.52	84.96	85.20	86.30
d' (max)	60.48	60.74	64.12	61.54	68.80	67.90	72.88
c' (eq)	78.75	78.98	80.52	79.08	88.40	88.75	86.30
d' (eq)	63.17	63.37	66.42	64.05	73.40	73.78	74.85

Note. See note from Table 28.

In terms of all-pairs power (see Tables 28 and 30) the stepwise MCPs were more powerful than the simultaneous MCPs across mean configurations and population variance-covariance structures. For the minimum and maximum range configurations the power difference was generally less than 20%, however for the equally spaced range configuration the power difference was generally substantial (i.e., greater than 20%). Among the three stepwise MCPs, the Welsch procedure was most powerful under a minimum range configuration, the SRB and Hoch procedures were most powerful under an equally spaced range configuration, and all three procedures had similar power rates under a maximum range configuration. Among the four simultaneous MCPs, Tukey's procedure was consistently most powerful across mean configurations for the investigated conditions. When always fitting a UN-H variance-covariance structure, the average all-pairs power rates of the simultaneous procedures, (Bon, Sidak, Tukey, and SMM), were 30.66%, 30.97%, 34.57%, and 31.48%, respectively and average rates for the three stepwise procedures (SRB, Hoch, and Welsch) were 49.65%, 50.77%, and 51.88%, respectively (see Table 28). The stepwise MCPs were more powerful than the simultaneous MCPs, with an average difference of 18.85%. Fitting the true population variance-covariance structure does provide a power advantage, however this difference does not exceed 3 percentage points compared to always fitting a UN-H structure (e.g., average all-pairs power rates were 32.80%, 33.08%, 36.31%, 33.40%, 51.09%, 51.94%, and 52.82%, respectively for the simultaneous and stepwise procedures when fitting the true population variance-covariance structure; see Table 30).

In terms of average per-pair power (see Tables 29 and 31), the stepwise procedures were more powerful than the simultaneous procedures across the mean configurations and population variance-covariance structures, however this power advantage was negligible (i.e., less than 10%). In addition, fitting the true population variance-covariance structure was generally more powerful than always fitting a UN-H variance-covariance structure but the difference was typically less than 2 percentage points. For the three stepwise procedures, SRB, Hoch, and Welsch, the average across investigated conditions for average per-pair power were 80.60%, 80.13%, and 83.78%, respectively, when fitting a UN-H variance-covariance structure and the average rates for Bon, Sidak, Tukey, and SMM were 72.59%, 72.80%, 75.55%, and 73.29%, respectively. When fitting the true population variance-covariance structure, the average average per-pair power rates were 73.80%, 74.02%, 76.40%, 74.38%, 80.48%, 79.96%, and 84.06%, respectively for the simultaneous and stepwise procedures.

For K = 8, a minimum range configuration was examined for the negative pairing of unequal groups sizes and variance-covariance matrices with df based on Satterthwaite's solution. In general, the power advantage of the stepwise MCPs over the stepwise procedures was smaller. It is important to note that Tukey's procedure was comparable to the SRB and Hoch procedures in terms of power rates and was at times more powerful than the simultaneous procedures (see Appendix B, Table B12). Likewise, for K = 8, fitting the true variance-covariance matrix was typically more powerful than always fitting a UN-H structure (see Appendix B, Table B13). In terms of all-pairs power, the difference averaged across population variance-covariance structures never exceeded 6 percentage points and for average per-pair power this difference was only 2 percentage points.

Discussion

A mixed model approach which allows a user to model the variance-covariance structure of the data was compared to known robust procedures based on a between-subjects heterogeneous Unstructured form of the variance-covariance matrix (i.e., UN-H) when testing all possible pairwise differences among repeated measures marginal means. Type I error and power results were reported for seven MCPs in a nonspherical repeated measures design containing one between- and one within-subjects variable under violation of normality and variance-covariance homogeneity, separately and jointly. The four simultaneous MCPs investigated are available in SAS' (1996) PROC MIXED and the three stepwise MCPs, although not available in SAS, can easily be computed from the statistical output. In addition to the testing strategies of always assuming a UN-H structure versus fitting the true variance-covariance structure, two model selection criteria were examined as testing strategies to evaluate the operating characteristics of the MCPs.

The testing strategy of always assuming a UN-H variance-covariance structure performed similarly to fitting the true variance-covariance structure across investigated conditions for each of the MCPs. The tendency was for error rates to be smaller when fitting the true structure compared to always fitting a UN-H structure. Furthermore, MCPs that were liberal under a UN-H structure became robust when the true variance-covariance structure was fit to the data. The improved Type I error control when fitting the true structure was more evident when df were based on the default option. For example, under violation of normality and multisample sphericity when the default df option was specified and unequal groups sizes were negatively paired with unequal variance-covariance matrices (condition d') all liberal error rates across MCPs (except Welsch) became robust when the true structure (i.e., RC-H) was fit to the data.

The advantage of always fitting a UN-H variance-covariance structure is that a researcher does not need prior knowledge about the true population variance-covariance structure to provide good Type I error control. An additional benefit of always fitting a UN-H variance-covariance structure as opposed to always fitting another between-subjects heterogeneous structure is that a UN structure is the most general (i.e., allowing the variances/covariances to be unequal and placing no restrictions on the form) and thus can be applied in any situation where a researcher is uncertain about the true nature of the population variance-covariance structure. Therefore, one can obtain robust pairwise comparisons among the levels of the RM main effect without prior knowledge about the true population variance-covariance structure. However, Keselman et al. (1999a) found that one needs prior knowledge about the true population structure in order to obtain robust tests of RM main and interaction effects with PROC MIXED.

Without prior knowledge about the true population variance-covariance structure a researcher has the choice of two model selection criteria available in SAS' (1996) PROC MIXED. The AIC and SBC criteria were investigated as testing strategies where each criterion selected the best variance-covariance structure from among 12 possible structures. Allowing either the AIC or SBC criterion to select the best variance-covariance structure generally provided robust Type I error control, however neither criterion can be relied upon to choose the correct structure with a high accuracy rate.

When multisample sphericity was violated, the accuracy rates for both criteria were highest when the population variance-covariance structure was RC-H. In contrast, when the population structure was either ARH1-H or UN-H, the tendency was for a more parsimonious structure to be selected (i.e., one with fewer parameters to be estimated). In particular a AR1-H structure was selected with the greatest frequency by both criteria. Under the combined violation of normality and multisample sphericity, the AIC criterion generally selected the correct structure with the greatest frequency but this averaged only 41% across conditions. The SBC criterion had the highest accuracy rates when the population structure was RC whereas, when the population structure was ARH1 or UN, a more parsimonious structure was selected with the greatest frequency. That is, the AR1 structure was selected when variance-covariance matrices were equal across the grouping variable (condition a) and the AR1-H structure was selected when variance-covariance matrices were unequal across the grouping variable (conditions b, c, c', d, and d').

The pattern of Type I error control was comparable between AIC and SBC because of the similar error rates across the between-subjects heterogeneous structures selected by either criterion as the best variance-covariance structure. A researcher can rely on either criterion to select a variance-covariance structure that will provide acceptable Type I error control, particularly when the selection is a between-subjects heterogeneous structure. However, if the goal of using either AIC or SBC is to select and examine the true variance-covariance structure, a researcher will likely be mislead because of their low accuracy rates. Therefore, another advantage of always fitting a UN-H structure is that one does not need to fit numerous variance-covariance structures to a set of data and thus compare model selection criterion values across structures to obtain robust pairwise comparisons among the levels of a RM main effect.

When multisample sphericity is violated, robust MCPs using a statistic with a nonpooled error term (i.e., KKS statistic - nonpooled across both between- and within-subjects factors) based on a UN-H variance-covariance structure with Satterthwaite df solution have been suggested (Keselman 1994; Keselman & Lix, 1995; Keselman et al., 1991). The performance of the MCPs based on Satterthwaite df and assuming a UN-H variance-covariance structure through PROC MIXED were consistent with previous research. However, the liberal rates of Type I error for the Welsch procedure across population variance-covariance structures was a new finding in that it indicates the limitation of the results reported by Keselman (1994) and Keselman and Lix (1995). A likely reason for these different results is the various forms of the population variance-covariance matrix and the data generation of a multivariate nonnormal distribution with more extreme degrees of skewness in the present study. A caution with the use of Welsch's procedure is that under certain conditions with Satterthwaite df, a more conservative critical value was used when the computed degrees of freedom for a pairwise test statistic was less than five (i.e., the tabled critical values for Welsch's procedure are not given for degrees of freedom less than five).

The robust performance of a nonpooled pairwise test statistic under violation of multisample sphericity is well known. However, the advantage of adopting a conservative method of estimating df (i.e., Satterthwaite's solution) for this statistic has never been compared to another method of estimating df. SAS' (1996, 1999) PROC MIXED provides a user with the flexibility of a default df option with the use of a nonpooled pairwise statistic. Reported results indicate that when the default degrees of freedom were used Type I error rates were typically larger than when Satterthwaite df was adopted and under certain conditions error rates exceeded Bradley's (1978) upper limit. For example, the combined violation of normality and multisample sphericity resulted in liberal error rates for all MCPs when unequal group sizes were negatively paired with unequal covariance matrices for a RC population variance-covariance structure. Although, fitting the true population variance-covariance structure to the data provided robust rates of error for all MCPs except Welsch's procedure when df were based on the default option. In contrast with the results based on Satterthwaite df, to obtain valid pairwise tests using default df, one needs prior knowledge about the true population variance-covariance structure.

Because a researcher never knows the true state of the population means, it is important that a MCP control Type I error under both complete and partial null hypotheses. The results for the simultaneous procedures under a partial null hypothesis were consistent with previous research (Keselman, 1993, 1994). That is, error rates were typically less than the .05 significance level and most of the time, rates were less than Bradley's (1978) lower limit. Under violation of multisample sphericity and the combined violation of normality and multisample sphericity, only two MCPs were able to maintain Type I error control within Bradley's (1978) limits except for the occasional conservative rate when unequal group sizes were positively paired with unequal variance-covariance matrices. Specifically, the SRB and Hoch procedures based on a UN-H structure with Satterthwaite df.

As expected, the stepwise MCPs were more powerful than the simultaneous MCPs when df were based on Satterthwaite's solution. This power difference was negligible under violation of multisample sphericity, but increased under the combined violation of normality and multisample sphericity in favor of the stepwise procedures. However, increasing the levels of the RM factor from 4 to 8 resulted in less discrepancy between the simultaneous and stepwise procedures in terms of all-pairs and average per-pair power rates. Furthermore, the power advantage of fitting the true population variance-covariance structure compared to always fitting a UN-H structure was negligible across the MCPs for both all-pairs and average per-pair power rates.

In conclusion, either Shaffer's (1986) modified sequentially rejective Bonferroni or Hochberg's (1988) sequentially acceptive Bonferroni procedure based on a UN-H structure with Satterthwaite df are recommended. Not only do these procedures control Type I error across a wide range of conditions but are powerful in detecting true pairwise differences. Lix and Keselman (1995) provide a SAS/IML (SAS Institute, 1989) program that enables a user to calculate nonpooled pairwise test statistics based on Satterthwaite's df solution which can be used in the computation of the previously mentioned MCPs. However, the availability of robust pairwise test statistics in a major statistical package should encourage their adoption by applied researchers because of ease and accessibility considerations. A potential outcome of this research would be the incorporation of robust MCPs such as SRB and Hoch into a future release of the PROC MIXED program given their superior performance over the simultaneous MCPs currently available. A potential advantage of the use of PROC MIXED is the ability of the program to handle missing data. Therefore, a direction for future research is the investigation of the robustness properties of various MCPs through PROC MIXED when data are missing.

References

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

Algina, J. (1994). Some alternative approximate tests for a split plot design. Multivariate Behavioral Research, 29, 365-384.

Algina, J. (1997). Generalization of improved general approximation tests to split-plot designs with multiple between-subjects factors and/or multiple within-subjects factors. British Journal of Mathematical and Statistical Psychology, 50, 243-252.

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

Algina, J., & Keselman, H. J. (1998). A power comparison of the Welch-James and improved general approximation tests in the split-plot design. Journal of Educational and Behavioral Statistics, 23, 152-169.

Algina, J., & Oshima, T. C. (1994). Type I error rates for Huynh's general approximation and improved general approximation tests. British Journal of Mathematical and Statistical Psychology, 47, 151-165.

Algina, J., & Oshima, T. C. (1995). An improved general approximation test for the main effect in a split-plot design. British Journal of Mathematical and Statistical Psychology, 48, 149-160.

Barcikowski, R. S., & Robey, R. R. (1984). Decisions in single group repeated measures analysis: Statistical tests and three computer packages. The American Statistician, 38, 148-150.

Bartlett, M. S. (1939). A note on tests of significance in multivariate analysis. Proceedings of the Cambridge Philosophical Society, 35, 180-185.

Boik, R. J. (1997). Analysis of repeated measures under second-stage sphericity: An empirical Bayes approach. Journal of Educational and Behavioral Statistics, 22, 155-192.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effect of inequality of variances and correlation of errors in the two-way classification. Annals of Mathematical Statistics, 25, 484-498.

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

Chen, R. S., & Dunlap, W. P. (1994). A monte carlo study on the performance of a corrected formula for $\tilde{\varepsilon}$ suggested by Lecoutre. Journal of Educational Statistics, 19, 119-126.

Collier, R. O. Jr., Baker, F. B., Mandeville, G. K., & Hayes, T. F. (1967). Estimates of test size for several test procedures based on conventional variance ratios in the repeated measures design. Psychometrika, 32, 339-353.

Dawson, K. S., Gennings, C., & Carter, W. H. (1997). Two graphical techniques useful in detecting correlation structure in repeated measures data. The American Statistician, 51, 275-283.

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

Duncan, D. B. (1957). Multiple range tests for correlated and heteroscedastic means. Biometrics, 13, 164-176.

Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56, 52-64.

Einot, I., & Gabriel, K. R. (1975). A study of the powers of several methods in multiple comparisons. Journal of the American Statistical Association, 70, 574-583.

Everitt, B. S. (1995). The analysis of repeated measures: A practical review with examples. The Statistician, 44, 113-135.

Fai, A. H., & Cornelius, P. L. (1996). Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. Journal of Statistical Computation and Simulation, 54, 363-378.

Fisher, R. A. (1935). The design of experiments. Edinburgh, Scotland: Oliver & Boyd.

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. Annals of Mathematical Statistics, 29, 855-891.

Giesbrecht, F. G., & Burns, J. C. (1985). Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results. Biometrics, 41, 477-486.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. Psychometrika, 24, 95-112.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's least significant difference test. Journal of the American Statistical Association, 81, 1000-1004.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 75, 800-802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6, 65-70.

Hotelling, H. (1931). The generalization of Student's ratio. Annals of Mathematical Statististics, 2, 360-378.

Hotelling, H. (1951). A generalized T test and measure of multivariate dispersion. In J Neyman (Ed.), Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA: University of California Press.

Huynh, H. (1978). Some approximate tests for repeated measurement designs. Psychometrika, 43, 161-175.

Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. Journal of the American Statistical Association, 65, 1582-1585.

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.

Jacard, J., Becker, M. A., & Wood, G. (1984). Pairwise multiple comparison procedures. Psychological Bulletin, 96, 589-596.

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. Biometrika, 38, 324-329.

James, G. S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. Biometrika, 41, 19-43.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. Biometrics, 42, 805-820.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. Biometrika, 67, 85-92.

Keselman, H. J. (1982). Multiple comparisons for repeated measures means. Multivariate Behavioral Research, 17, 87-92.

Keselman, H. J. (1993). Stepwise multiple comparisons of repeated measures means under violations of multisample sphericity. In F. M. Hoppe (Ed.), Multiple comparisons, selection, and applications in biometry (pp. 167-186). New York: Marcel Dekker.

Keselman, H. J. (1994). Stepwise and simultaneous multiple comparison procedures of repeated measures' means. Journal of Educational Statistics, 19, 127-162.

Keselman, H. J., & Algina, J. (1996). The analysis of higher-order repeated measures designs. In B. Thompson (Ed.), Advances in social science methodology: Vol. 4 (pp. 45-70). Greenwich, CT: JAI Press.

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, 591-604.

Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1999a). The analysis of repeated measurements: A comparison of mixed-model Satterthwaite F tests and a nonpooled adjusted degrees of freedom multivariate test. Communications in Statistics - Theory and Methods, 28, 2967-2999.

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

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

Keselman, H. J., Huberty, C. J, Lix, L. M., Olejnik, S., Cribbie, R. A., Donohue, B., Kowalchuk, R. K., Lowman, L. L., Petosky, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. Review of Educational Research, 68, 350-386.

Keselman, H. J., & Keselman, J. C. (1988). Repeated measures multiple comparison procedures: Effects of violating multisample sphericity in unbalanced designs. Journal of Educational Statistics, 13, 215-226.

Keselman, H. J., & Keselman, J. C. (1993). Analysis of repeated measurements. In L. K. Edwards (Ed.), Applied analysis of variance in behavioral science (pp. 105-145). New York: Marcel Dekker.

Keselman, H. J., Keselman, J. C., & Lix, L. M. (1995). The analysis of repeated measurements: Univariate tests, multivariate tests, or both? British Journal of Mathematical and Statistical Psychology, 48, 319-338.

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

Keselman, H. J., Kowalchuk, R. K., & Boik, R. B. (in press). An investigation of the empirical Bayes approach to the analysis of repeated measurements. British Journal of Mathematical and Statistical Psychology.

Keselman, H. J., & Lix, L. M. (1995). Improved repeated measures stepwise multiple comparison procedures. Journal of Educational and Behavioral Statistics, 20, 83-99.

Keselman, H. J., Lix, L. M., & Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. Psychological Methods, 3, 123-141.

Keselman, H. J., Rogan, J. C., & Games, P. A. (1981). Robust tests of repeated measures means in educational and psychological research. Educational and Psychological Measurement, 41, 163-173.

Keselman, H. J., Rogan, J. C., Mendoza, J. L., & Breen, L. J. (1980). Testing the validity conditions of repeated measures F tests. Psychological Bulletin, 87, 479-481.

Keselman, J. C., & Keselman, H. J. (1990). Analysing unbalanced repeated measures designs. British Journal of Mathematical and Statistical Psychology, 43, 265-282.

Keselman, J. C., Lix, L. M., & Keselman, H. J. (1996). The analysis of repeated measurements: A quantitative research synthesis. British Journal of Mathematical and Statistical Psychology, 49, 275-298.

Keuls, M. (1952). The use of the "Studentized range" in conjunction with an analysis of variance. Euphytica, 1, 112-122.

Kirk, R. E. (1995). Experimental design: Procedures for the behavioral sciences (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kleinbaum, D. G. (1973). Testing linear hypotheses in generalized multivariate linear models. Communications in Statistics, 1, 433-457.

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

Lawley, D. N. (1938). A generalization of Fisher's z test. Biometrika, 30, 180-187.

Lecoutre, B. (1991). A correction for the $\tilde{\varepsilon}$ approximate test in repeated measures designs with two or more independent groups. Journal of Educational Statistics, 16, 371-372.

Lewis, C. (1993). Analyzing means from repeated measures data. In G. Keren & C. Lewis (Eds.), A handbook for data analysis in the behavioral sciences: Statistical issues (pp. 73-94). Hillsdale, NJ: Lawrence Erlbaum.

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

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). SAS system for mixed models. Cary, NC: SAS Institute.

Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. Psychological Bulletin, 117, 547-560.

Looney, S. W., & Stanley, W. B. (1989). Exploratory repeated measures analysis for two or more groups: Review and update. The American Statistician, 43, 220-225.

Lovie, A. D. (1981). On the early history of ANOVA in the analysis of repeated measures designs in psychology. British Journal of Mathematical and Statistical Psychology, 34, 1-15.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. Annals of Mathematical Statistics, 29, 204-209.

Maxwell, S. E. (1980). Pairwise multiple comparisons in repeated measures designs. Journal of Educational Statistics, 5, 269-287.

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

McCall, R. B., & Appelbaum, M. I. (1973). Bias in the analysis of repeated-measures designs: Some alternative approaches. Child Development, 44, 401-415.

McLean, R. A., & Sanders, W. L. (1988). Approximating degrees of freedom for standard errors in mixed linear models. Proceedings of the Statistical Computing Section, American Statistical Association, New Orleans, 50-59.

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

Mitzel, H. C., & Games, P. A. (1981). Circularity and multiple comparisons in repeated measure designs. British Journal of Mathematical and Statistical Psychology, 34, 253-259.

Muller, K. E., LaVange, L. M., Ramey, S. L., & Ramey, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. Journal of the American Statistical Association, 87, 1209-1226

Newman, D. (1939). The distribution of the range in samples from a normal population expressed in terms of an independent estimate of standard deviation. Biometrika, 31, 20-30.

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. Psychological Bulletin, 97, 316-333.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. Journal of the American Statistical Association, 69, 894-908.

Peritz, E. (1970). A note on multiple comparisons. Unpublished manuscript, Hebrew University, Israel.

Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis. Annals of Mathematical Statistics, 26, 117-121.

Quintana, S. M., & Maxwell, S. E. (1994). A monte carlo comparison of seven ε-adjustment procedures in repeated measures designs with small sample sizes. Journal of Educational Statistics, 19, 57-71.

Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. Journal of the American Statistical Association, 73, 479-485.

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

Rouanet, H., & Lepine, D. (1970). Comparison between treatments in a repeated-measures design: ANOVA and multivariate methods. British Journal of Mathematical and Statistical Psychology, 23, 147-163.

Roy, S. N. (1953). On a heuristic method of test construction and it use in multivariate analysis. Annals of Mathematical Statistics, 24, 220-238

Ryan, T. A. (1960). Significance tests for multiple comparison of proportions, variances and other statistics. Psychological Bulletin, 57, 318-328.

SAS Institute. (1989). SAS/IML software: Usage and reference, Version 6. Cary, NC: Author.

SAS Institute. (1996) SAS/STAT software: Changes and enhancements through release 6.11. Cary, NC: Author.

SAS Institute. (1997). SAS Macro language: Reference, First edition. Cary, NC: Author.

SAS Institute. (1999). SAS/STAT user's guide, Version 7. Cary, NC: Author.

Satterthwaite, F. E. (1941). Synthesis of variance. Psychometrika, 6, 309-316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. Biometrics, 2, 110-114.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. Psychological Bulletin, 111, 352-360.

Scheffé, H. (1959). The analysis of variance. New York: Wiley.

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

Seaman, M. A., Levin, J. R., & Serlin R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. Psychological Bulletin, 110, 577-586.

Searle, S. R. (1971). Linear models. New York: John Wiley & Sons.

Shaffer, J. P. (1979). Comparison of means: An F test followed by a modified multiple range procedure. Journal of Educational Statistics, 4, 14-23.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.

Sidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association, 62, 626-633.

Timm, N. H. (1975). Multivariate analysis with applications in education and psychology. Monterey, CA: Brooks-Cole.

Toothaker, L. E. (1991). Multiple comparison procedures. Newbury Park, CA: Sage

Tukey, J. W. (1953). The problem of multiple comparisons. Unpublished manuscript, Princeton University.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. Biometrika, 34, 28-35.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. Biometrika, 38, 330-336.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. Journal of the American Statistical Association, 72, 566-575.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. Biometrika, 24, 471-494.

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

Wolfinger, R. D. (1996). Heterogeneous variance-covariance structures for repeated measures. Journal of Agricultural, Biological, and Environmental Statistics, 1, 205-230.

Wolfinger, R., Tobias, R., & Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. SIAM Journal on Scientific Computing, 15, 1294-1310.

Wright, S. P. (1995). Adjusted F tests for repeated measures with the MIXED procedure. Proceedings of the Twentieth Annual SAS Users Group International Conference (pp. 1154-1159). Cary, NC: SAS Institute.

Wright, S. P., & Wolfinger, R. D. (1996, October). Repeated measures analysis using mixed models: Some simulation results. Paper presented at conference on modelling longitudinal and spatially correlated data: Methods, applications, and future directions. Nantucket, MA.

Footnotes

[1] Sphericity is not a requirement for the statistics investigated, however prior research has indicated that tests of the type investigated were affected when data were not spherical (see Keselman et al., 1993).

[2] An extension of the KKS pairwise statistic to test omnibus fixed-effect tests is the approximate df multivariate approach due to Johansen (1980) and Keselman et al. (1993). However, tests of omnibus fixed-effects with a UN-H covariance structure with Satterthwaite df through the MIXED procedure are not equivalent to this approximate df multivariate WJ approach.

Appendix A

Population Correlation and Covariance Structures

Heterogeneous First-Order Autoregressive (ARH1)

Epsilon = .7490698

Correlation Matrix

1	0.7300	0.5329	0.3890
	1	0.7300	0.5329
		1	0.7300
			1

Covariance Matrix

8.0	6.5293185	4.7664025	3.8115726
	10.0	7.3	5.837627
		10.0	7.9967493
			12.0

Epsilon = .7472201

Correlation Matrix

1	0.46	0.2116	0.097336	0.0447746	0.0205963	0.0094743	0.0043582
	1	0.46	0.2116	0.097336	0.0447746	0.0205963	0.0094743
		1	0.46	0.2116	0.097336	0.0447746	0.0205963
			1	0.46	0.2116	0.097336	0.0447746
				1	0.46	0.2116	0.097336
					1	0.46	0.2116
						1	0.46
							1

Covariance Matrix

8	3.68	1.8926079	0.8705997	0.4004758	0.1842189	0.0928288	0.0427012
	8	4.1143651	1.8926079	0.8705997	0.4004758	0.2018017	0.0928288
		10	4.6	2.116	0.97336	0.4904807	0.2256211
			10	4.6	2.116	1.0662625	0.4904807
				10	4.6	2.3179619	1.0662625
					10	5.0390475	2.3179619
						12	5.52
							12

Random Coefficient (RC)

Epsilon = 0.7521197

Correlation Matrix

1	0.6728846	0.6736112	0.6637471
	1	0.7985195	0.8074601
		1	0.8702999
			1

Covariance Matrix

2.7301562	2.3002343	2.8703124	3.4403904
	4.2803124	4.2603903	5.2404688
		6.6504686	7.0405469
			9.8406248

Epsilon = 7522259

Correlation Matrix

1	0.5544533	0.5603854	0.5589393	0.552914	0.5443524	0.5346061	0.5245107
	1	0.6091684	0.6199025	0.623752	0.6231524	0.6198349	0.6149589
		1	0.6638484	0.6763223	0.6827344	0.6851224	0.6849127
			1	0.7137357	0.7261191	0.7333992	0.7372158
				1	0.7569291	0.7683228	0.7755436
					1	0.7932651	0.8033376
						1	0.8233998
							1

Covariance Matrix

2.1639159	1.2713738	1.3788318	1.4862897	1.5937477	1.7012056	1.8086636	1.9161215
	2.4298318	1.5882897	1.7467477	1.9052056	2.0636636	2.2221215	2.3805795
		2.7977477	2.0072056	2.2166636	2.4261215	2.6355795	2.8450374
			3.2676636	2.5281215	2.7885795	3.0490374	3.3094954
				3.8395795	3.1510374	3.4624954	3.7739533
					4.5134954	3.8759533	4.2384113
						5.2894113	4.7028692
							6.1673272

Unstructured (UN)

Epsilon = .7505351

Correlation Matrix

1	0.729445	0.6546579	0.3462883
	1	0.7087869	0.3669361
		1	0.4932449
			1

Covariance Matrix

8.0	6.5243544	5.8554383	3.3929186
	10.0	7.087869	4.0195836
		10.0	5.4032272
			12.0

Epsilon = 0.7508431

Correlation Matrix

1	0.5943965	0.4785602	0.5034549	0.4738909	0.5157068	0.5507265	0.4286351
	1	0.5073292	0.4995573	0.5354644	0.5217312	0.6097338	0.1999506
		1	0.6014101	0.571214	0.5950098	0.5655234	0.3317410
			1	0.6629326	0.6227244	0.7274189	0.5669606
				1	0.6933965	0.6878822	0.6774241
					1	0.7905351	0.6790915
						1	0.6186267
							1

Covariance Matrix

8	4.7551721	4.2803725	4.5030373	4.2386088	4.6126218	5.3959954	4.1997488
	8	4.5376901	4.4681761	4.7893396	4.6665056	5.9741465	1.9591081
		10	6.0141005	5.7121395	5.9500983	6.1949987	3.6340406
			10	6.6293261	6.2272444	7.9684752	6.2107423
				10	6.9339653	7.5353715	7.4208096
					10	8.6598784	7.4390746
						12	7.4235206
							12

Appendix B

Type I Error and Power Results for K = 8

Table B1

Study Conditions (Lognormal Distribution, K = 8, N = 60 Satterthwaite df)

Condition	Pop Cov Str[a]	Sample Sizes	Cov Mat[b]	Pairing	Null Hypothesis
d'	ARH1-H	12,20,28	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=\mu_5=\mu_6=\mu_7=\mu_8=0$
d'	RC-H	12,20,28	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=\mu_5=\mu_6=\mu_7=\mu_8=0$
d'	UN-H	12,20,28	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4=\mu_5=\mu_6=\mu_7=\mu_8=0$
d'	ARH1-H	12,20,28	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4\neq\mu_5=\mu_6=\mu_7=\mu_8$
d'	RC-H	12,20,28	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4\neq\mu_5=\mu_6=\mu_7=\mu_8$
d'	UN-H	12,20,28	1:3:5	-	$\mu_1=\mu_2=\mu_3=\mu_4\neq\mu_5=\mu_6=\mu_7=\mu_8$

Note. See note from Table 1.

Table B2

Empirical Type I Error Rates (%) when Fitting a UN-H Covariance Structure

(Lognormal Distribution, K = 8, N = 60 Satterthwaite df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
d'	2.50	2.70	4.50	3.20	2.50	2.50	**8.20**
RC							
d'	3.90	3.90	5.00	4.00	3.80	3.80	**9.90**
UN							
d'	3.70	3.70	5.00	3.90	3.70	3.70	**10.00**
Partial Null Hypothesis							
ARH1							
d'	**1.50**	**1.50**	**1.90**	**1.70**	3.10	2.70	6.70
RC							
d'	**1.40**	**1.50**	**2.40**	**1.80**	3.20	3.10	**7.80**
UN							
d'	**1.30**	**1.40**	2.50	**1.70**	3.40	3.20	7.10

Note. See note from Table 2.

Table B3

Empirical Type I Error Rates (%) when Fitting the True Covariance Structure

(Lognormal Distribution, K = 8, N = 60, Satterthwaite df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
d'	2.60	2.60	3.70	2.70	**1.90**	**1.90**	**7.60**
RC							
d'	**2.40**	**2.40**	3.70	2.50	**2.20**	**2.20**	7.50
UN							
d'	3.70	3.70	5.00	3.90	3.70	3.70	**10.00**
Partial Null Hypothesis							
ARH1							
d'	**0.80**	**0.90**	**2.00**	**1.20**	2.50	**2.40**	6.30
RC							
d'	**1.10**	**1.10**	**1.80**	**1.10**	3.30	3.00	5.70
UN							
d'	**1.30**	**1.40**	2.50	**1.70**	3.40	3.20	7.10

Note. See note from Table 2.

Table B4

Empirical Type I Error Rates (%) with Akaike Criterion Selecting the Best Covariance Structure (Lognormal Distribution, K = 8, N = 60, Satterthwaite df)

Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
Complete Null Hypothesis							
ARH1							
d'	2.50	2.50	3.60	2.60	**1.90**	**1.90**	**7.90**
RC							
d'	3.40	3.40	4.40	3.50	3.30	3.30	**8.20**
UN							
d'	3.60	3.70	5.10	4.00	3.60	3.60	**10.10**
Partial Null Hypothesis							
ARH1							
d'	**0.80**	**0.90**	**2.30**	**1.20**	2.80	2.70	6.40
RC							
d'	**1.10**	**1.10**	**2.10**	**1.20**	3.10	2.90	6.40
UN							
d'	**1.50**	**1.60**	2.80	**1.90**	3.60	3.40	7.10

Note. See note from Table 2.

Table B5

Empirical Type I Error Rates (%) with Schwarz Criterion Selecting the Best Covariance Structure (Lognormal Distribution, K = 8, N = 60, Satterthwaite df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
d'	3.20	3.20	4.60	3.30	3.20	3.20	**9.20**
RC							
d'	2.60	2.60	4.00	2.70	**2.40**	**2.40**	**7.90**
UN							
d'	2.70	2.70	4.10	2.80	2.60	2.60	**8.20**
Partial Null Hypothesis							
ARH1							
d'	**1.70**	**1.80**	3.00	**1.90**	4.20	4.00	**8.30**
RC							
d'	**1.40**	**1.40**	**2.20**	**1.40**	3.70	3.40	6.40
UN							
d'	**1.50**	**1.50**	**2.10**	**1.50**	3.50	3.40	6.70

Note. See note from Table 2.

Table B6

Percentage of Time Akaike Criterion Selected the Correct Covariance Structure (rounded to whole numbers) (Lognormal Distribution, K = 8, N = 60, Satterthwaite df)

	Covariance Structure											
Cond	CSH	CSH-H	HF	HF-H	AR1	AR1-H	ARH1	ARH1-H	RC	RC-H	UN	UN-H
d'						13		**84**				3
d'		7								**46**		43
d'				2						6		**90**
d'						14		**83**				2
d'		5								**48**		44
d'				2						6		**90**

Note. See note from Table 10.

Table B7

Percentage of Time Schwarz Criterion Selected the Correct Covariance Structure (rounded to whole numbers) (Lognormal Distribution, K = 8, N = 60, Satterthwaite df)

	Covariance Structure											
Cond	CSH	CSH-H	HF	HF-H	AR1	AR1-H	ARH1	ARH1-H	RC	RC-H	UN	UN-H
d'						84		**16**				
d'		2							2	**95**		
d'		4		2						88		
d'						87		**13**				
d'		3							2	**95**		
d'		7				2				86		

Note. See note from Table 10.

Table B8

Empirical Type I Error Rates (%) when Fitting a UN-H Covariance Structure

(Lognormal Distribution, K = 8, N = 60 Default df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
d'	6.80	6.90	**8.50**	6.90	6.80	6.80	**16.20**
RC							
d'	**8.40**	**8.40**	**10.20**	**8.40**	**8.40**	**8.40**	**17.70**
UN							
d'	6.20	6.20	**7.80**	6.20	6.20	6.20	**16.00**

Note. See note from Table 2.

Table B9

Empirical Type I Error Rates (%) when Fitting the True Covariance Structure

(Lognormal Distribution, K = 8, N = 60, Default df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
d'	4.40	4.50	5.50	4.50	4.40	4.40	**11.80**
RC							
d'	4.60	4.60	5.80	4.60	4.60	4.60	**12.00**
UN							
d'	6.20	6.20	**7.80**	6.20	6.20	6.20	**16.00**

Note. See note from Table 2.

Table B10

Empirical Type I Error Rates (%) with Akaike Criterion Selecting the Best Covariance Structure (Lognormal Distribution, K = 8, N = 60, Default df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
d'	4.50	4.60	5.70	4.60	4.50	4.50	**12.10**
RC							
d'	6.50	6.50	**8.30**	6.50	6.50	6.50	**14.90**
UN							
d'	6.00	6.10	**7.80**	6.10	6.00	6.00	**15.70**

Note. See note from Table 2.

Table B11

Empirical Type I Error Rates (%) with Schwarz Criterion Selecting the Best Covariance Structure (Lognormal Distribution, K = 8, N = 60, Default df)

Complete Null Hypothesis							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
d'	4.30	4.50	5.80	4.50	4.30	4.30	**11.50**
RC							
d'	4.90	4.90	6.30	4.90	4.90	4.90	**12.30**
UN							
d'	4.30	4.40	5.60	4.40	4.30	4.30	**11.70**

Note. See note from Table 2.

Table B12

Power Rates (Minimum Mean Configuration) when Fitting a UN-H Covariance Structure

(Lognormal Distribution K = 8, N = 60, Satterthwaite df)

All-Pairs Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
d'	15.20	15.60	21.50	16.80	22.10	22.20	73.90
RC							
d'	12.00	12.10	17.30	12.70	19.90	20.00	80.90
UN							
d'	23.50	23.70	31.40	25.40	31.70	31.70	86.40

Average Per-Pair Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
d'	69.69	70.03	75.02	71.19	73.65	72.74	92.71
RC							
d'	71.34	71.66	77.13	72.86	75.64	74.81	95.99
UN							
d'	80.81	81.03	84.75	81.74	83.71	83.36	97.48

Note. d' = negative pairing of unequal covariance matrices and unequal group sizes (n_j = 12, 20, 28); Bon = Bonferroni; SMM = Studentized maximum modulus; SRB = Shaffer's (1986) sequentially rejective Bonferroni; Hoch = Hochberg's (1988) sequentially acceptive Bonferroni.

Table B13

Power Rates (Minimum Mean Configuration) when Fitting the True Covariance Structure

(Lognormal Distribution K = 8, N = 60, Satterthwaite df)

All-Pairs Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
d'	24.70	24.90	28.20	25.20	31.30	31.30	77.40
RC							
d'	18.70	18.80	24.30	19.60	21.90	21.90	75.90
UN							
d'	23.50	23.70	31.40	25.40	31.70	31.70	86.40

Average Per-Pair Power							
ARH1							
Condition	Bon	Sidak	Tukey	SMM	SRB	Hoch	Welsch
d'	76.79	77.01	79.88	77.24	80.03	79.51	93.80
RC							
d'	68.51	68.76	74.66	70.57	65.26	64.62	90.46
UN							
d'	80.81	81.03	84.75	81.74	83.71	83.36	97.48

Note. See note from Table B12.

Appendix C

Population Means for Power Conditions

K = 4 Minimum Range Configuration

ARH1: -0.75 -0.75 0.75 0.75 Effect Size = 0.75

RC: -0.50 -0.50 0.50 0.50 Effect Size = 0.50

UN: -0.75 -0.75 0.75 0.75 Effect Size = 0.75

K = 4 Maximum Range Configuration

ARH1: -1.41421356 0.00 0.00 1.41421356 Effect Size = 1.00

RC: -1.06066017 0.00 0.00 1.06066017 Effect Size = 0.75

UN: -1.41421356 0.00 0.00 1.41421356 Effect Size = 1.00

K = 4 Equally Spaced Range Configuration

ARH1: -1.67705098 -0.55901699 0.55901699 1.67705098
Effect Size = 1.25

RC: -1.00623059 -0.33541020 0.33541020 1.00623059
Effect Size = 0.75

UN: -1.67705098 -0.55901699 0.55901699 1.67705098
Effect Size = 1.25

K = 8 Minimum Range Configuration

ARH1: -1.25 -1.25 -1.25 -1.25 1.25 1.25 1.25 1.25
Effect Size = 1.25

RC: -0.50 -0.50 -0.50 -0.50 0.50 0.50 0.50 0.50
Effect Size = 0.50

UN: -1.00 -1.00 -1.00 -1.00 1.00 1.00 1.00 1.00
Effect Size = 1.00