THE ANALYSIS OF REPEATED MEASUREMENTS: A COMPARISON OF MIXED-MODEL SATTERTHWAITE F TESTS AND A NONPOOLED ADJUSTED DEGREES OF FREEDOM MULTIVARIATE TEST

H. J. Keselman
University of Manitoba
Winnipeg, Manitoba
Canada R3T 2N2

James Algina
University of Florida
Gainesville, Florida
32611-7047

Rhonda K. Kowalchuk
University of Manitoba
Winnipeg, Manitoba
Canada R3T 2N2

Russell D. Wolfinger
SAS Institute
Cary, North Carolina
27513

***Key Words*: Repeated measurements; Mixed-model analyses; Welch-James adjusted-df test, Unbalanced nonspherical designs;**

ABSTRACT

Mixed-model analysis is the newest approach to the analysis of repeated measurements. The approach is supposed to be advantageous (i.e., efficient and powerful) because it allows users to model the covariance structure of their data prior to assessing treatment effects. The statistics for this method are based on an F-distribution with degrees of freedom often just approximated by the residual degrees of freedom. However, previous results indicated that these statistics can produce biased Type I error rates under conditions believed to characterize behavioral science research. This study investigates a more complex degrees of freedom method based on Satterthwaite's technique of matching moments. The resulting mixed-model F-tests are compared with a Welch-James-type test which has been found to be generally robust to

assumption violations. Simulation results do not universally favor one approach over the other, although additional considerations are discussed outlining the relative merits of each approach.

1. INTRODUCTION

Keselman, Algina, Kowalchuk, and Wolfinger (1999) compared various strategies for analyzing treatment effects in unbalanced (unequal group sizes) repeated measures designs when covariance matrices were unequal and nonspherical and the data were nonnormal as well. In particular, they compared the univariate adjusted degrees of freedom (df) statistics of Greenhouse and Geisser (1959), Huynh and Feldt (1976), and Huynh (1978) with a multivariate corrected df Welch (1947, 1951)-James (1951, 1954)-type statistic (WJ) presented by Johansen (1980), [see Keselman, Carriere & Lix (1993) for the application of the WJ statistic to the analysis of repeated measures effects] and a mixed-model analysis as computed by PROC MIXED (SAS Institute, 1996). The first two approaches correct for violations of the sphericity assumption associated with the conventional F tests, while the Huynh (1978) approach corrects for violations of multisample sphericity, employing the conventional F test. The Johansen (1980) approach, a multivariate extension of the Welch (1951) and James (1951) procedures for completely randomized designs, involves the computation of a statistic that does not pool across heterogeneous sources of variation and estimates error df from sample data. On the other hand, the mixed-model approach allows users to model the covariance structure of the data prior to computing tests of significance of the repeated measures effects.

For completeness we note that sphericity (denoted by ϵ) refers to homogeneity of the treatment-difference variances among the levels of the repeated measures variable (Huynh & Feldt, 1970; Rogan et al., 1979;

Rouanet & Lepine, 1970). When the design contains a between-subjects grouping factor, then an additional requirement for valid conventional F-tests is that the covariance matrices of these treatment-difference variances are the same for all levels of this grouping factor. Jointly, these two assumptions have been referred to as multisample sphericity (Mendoza, 1980).

Keselman et al. (1999) found that only the procedures presented by Johansen (1980) and Huynh (1978) were able to generally control rates of Type I error under assumption violation conditions. Based on power results presented by Algina and Keselman (1998), Keselman et al. recommended the WJ statistic since it typically is more powerful than the Huynh procedure.

However, in Keselman et al. (1999), the df for the mixed-model analyses were based on the residual variance and are equal to the error df in the conventional F test (see SAS Institute, 1996, pp. 565-566). When data are not spherical it is known that these df are too large when the conventional F test is used and one might expect they will be too large when a mixed-model analysis is used. An alternative to the df used in the Keselman et al. study is the Satterthwaite df.

The Satterthwaite approximation method has traditionally been applied in variance component (random-effects) models and in some alternatives to fixed-effects analysis of variance such as the Welch (1938) t and Brown and Forsythe (1974) tests, and has a rich history dating back to Satterthwaite (1941). Khuri, Mathew, and Sinha (1998) provide a recent review of many of its uses. However, the application of the Satterthwaite method to general covariance structures is a recent development and merits consideration because of its success in other arenas. The Satterthwaite df tend to be much more conservative than the residual df and thus offer the hope of providing more accurate F-tests since the study by Keselman et al. (1999) showed the residual df were often too liberal.

Therefore, the purpose of the current study was to compare the mixed-model procedure using the Satterthwaite df to the WJ procedure in terms of control of Type I error rates and power.

2. THE MIXED-MODEL APPROACH

We first present a brief discussion of the mixed-model approach to the analysis of repeated measurements. The linear model underlying the mixed-model approach can be written as follows:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E}, \tag{2.1}$$

where **Y** is a vector of response scores, **X** and **Z** are known design matrices, **B** is a vector of unknown fixed-effects parameters, and **U** is a vector of unknown random-effects. In the mixed-model **E** is an unknown error vector whose elements need not be independent and homogeneous. The name for this approach to the analysis of repeated measurements stems from the fact that the model contains both unknown fixed- and random-effects. We assume that **U** and **E** are normally distributed with

$$\mathrm{E}\begin{bmatrix}\mathbf{U}\\ \mathbf{E}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

and

$$\mathrm{VAR}\begin{bmatrix}\mathbf{U}\\ \mathbf{E}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}.$$

Thus, the variance of the response measure is given by

$$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}. \tag{2.2}$$

Accordingly, one can model **V** by specifying **Z** and covariance structures for **G** and **R**. Note that the usual general linear model is arrived at by letting $\mathbf{Z} = \mathbf{0}$ and $\mathbf{R} = \sigma^2\mathbf{I}$. A mixed-model approach to the analysis of repeated measurements has been discussed by Jennrich and Schluchter (1986), Littell, Milliken, Stroup and Wolfinger (1996), and Wolfinger (1993). The choice of estimation procedure for mixed-model analysis and the formation of test statistics is described in Littell et al. (1996, pp.498-502). Results can be obtained with several software programs; we use PROC MIXED (SAS Institute, 1996).

Advocates of this approach suggest that it provides the "best" approach to the analysis of repeated measurements since it can, among other things, handle missing data and also allows users to model the covariance structure of the data. (For discussions regarding the analysis of repeated measurements with missing data the reader should see Little, 1994, 1995) The first of these advantages would be important to those researchers who do not have complete measurements on all subjects across time; however, in many applied areas, though not all, which involve controlled experiments, this advantage is not likely to be relevant since data in these contexts are rarely missing. (In our study we only consider repeated measures designs in which there are complete measurements on each subject.) The second consideration, however, could be most relevant to experimenters since, according to the developers of mixed-model analyses, modeling the correct covariance structure of the data should result in more powerful tests of the fixed-effects parameters. As Wolfinger (1996, p. 208) notes, "A basic rationale here is 'parsimony means power'; that is, to obtain the most efficient inferences about the mean model, one selects the most parsimonious covariance structure possible that still reasonably fits the data."

The mixed-model approach allows researchers to examine a number of different covariance structures that could possibly describe their particular data [e.g., compound symmetric (CS), spherical (HF) (the structure

assumed by many programs for valid univariate tests; see Huynh and Feldt, 1970), unstructured (UN) (the structure assumed by many programs for valid multivariate tests), first-order autoregressive (AR), random coefficients (RC), etc.]. AR and RC structures reflect that measurement occasions that are closer in time are more highly correlated than those farther apart in time. The program allows even greater flexibility to the user by allowing him/her to model covariance structures that have within-subjects and/or between-subjects heterogeneity. Within-subjects heterogeneity is characterized by unequal variances across repeated measures while with between-subjects heterogeneity the covariance matrices differ across groups (e.g., $\mathbf{\Sigma}_1 = \frac{1}{3}\mathbf{\Sigma}_2$, and $\mathbf{\Sigma}_3 = \frac{5}{3}\mathbf{\Sigma}_2$).

Table 1 contains a number of different covariance structures that could represent the covariance structure for repeated measures data. The CS and AR structures are homogeneous structures because the variances along the main diagonal are constant. The covariances for the CS form remain constant whereas for the AR form they decline exponentially. On the other hand, the UN covariance matrix represents a heterogeneous form because the variances (for each observation) vary. These three structures illustrate, in general, the advantage/disadvantage of adopting different covariance structures when modeling repeated measures data. The UN structure allows the best possible variance-covariance fit, however at a cost of estimating many parameters [$K(K+1)/2$, where K stands for the number of repeated measurements]. On the other hand, AR is an example of a very parsimonious structure (i.e., only 2 parameters), however, there is greater potential for lack-of-fit to the data. An intermediary form, ARH, permits the variances of the AR structure to change at each time point. The RC structure is also intermediary.

TABLE I
Covariance Structures

Compound Symmetric (CS)

(Parameters-2)

$$\begin{bmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 \end{bmatrix}$$

Firstorder autoregressive[AR]

(Parameters-2)

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Random Coefficients (RC)

(Parameters-4)

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} + \begin{bmatrix} \sigma^2 & & & \\ & \sigma^2 & & \\ & & \sigma^2 & \\ & & & \sigma^2 \end{bmatrix}$$

Unstructured (UN)

(Parameters-10)

$$\begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22}^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33}^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44}^2 \end{bmatrix}$$

Heterogeneous AR (ARH)

(Parameters-5)

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 \end{bmatrix}$$

Another degree of complexity can be added by considering heterogeneity between groups. That is, mixed-model analysis not only alows users to model the covariance structure for subjects for a repeated measures variable but also allows them to fit different covariance structures to each group of subjects for each level of a between-subjects grouping variable.

With such a wide variety of covariance structures available under the mixed-model approach, selecting can become a difficult task. Keselman, Algina, Kowalchuk, and Wolfinger (1998) compared two popular selection criteria due to Akaike (1974) and Schwarz (1978). These criteria were compared for various between- by within-subjects repeated measures designs in which the true covariance structure of the data was varied as well as the distributional form of the data, group size, and between-subjects covariance equality/inequality. Their data indicated that neither criterion uniformly selected the correct covariance structure. Indeed, for most of the structures investigated, both criteria, and particularly the Schwarz (1978) criteria, more frequently picked the wrong covariance structure. Thus, though the mixed-model approach allows users to model the covariance structure, two popular criteria for selecting the "best" structure performed poorly.

A possible explanation for the low percentages of correct structure selection for the Akaike criterion is that structures other than the true ones may provide adequate approximations. For example, heterogeneous AR and RC structures may provide adequate surrogates for more complex UN structures. Most likely larger sample sizes would provide more power to distinguish these structures. However, sample sizes in the Keselman et al. (1998) study were chosen to reflect sizes typically employed in applied settings and thus their findings would be relevant to the majority of applied researcher endeavors, at least according to the survey results provided by Kowalchuk, Lix, and Keselman (1996) and Keselman et al. (1998). Keselman et al. also speculated that the Schwarz (1978) criterion may have performed poorly because of the way in which PROC MIXED

computes its penalty criterion. Release 7.01 of PROC MIXED (not yet available) will compute the Schwarz (1978) criterion with a less stringent penalty.

Because of these results, in the current study we compared the mixed-model and WJ procedure both when the correct model was fit to the data and when the model was selected by using the Akaike (1974) information criteria. Thus we compared the procedures under conditions that are optimal for the mixed-model (correct structural model is known) and under conditions that are more realistic (the structural model is selected based on the data).

Once a covariance structure has been determined, the mixed-model approach proceeds by estimating the parameters in $\mathbf{V}$ using restricted maximum likelihood (REML, refer to Little et al., 1996 for details). This produces an estimated variance matrix $\widehat{\mathbf{V}}$, which is then used to estimate $\mathbf{B}$ using the well-known generalized least squares formula

$$\widehat{\mathbf{B}} = (\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-}\,\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{Y}$$

where $^{-}$ denotes a generalized inverse used to account for a typically less-than-full-rank $\mathbf{X}$ matrix. Inferences for the fixed-effects (specifically a test of H_0: $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$) are conducted by forming different versions of an estimable contrast matrix $\mathbf{L}$ and forming the F-statistic

$$F = \frac{\widehat{\boldsymbol{B}}'\boldsymbol{L}'(\boldsymbol{L}\widehat{\boldsymbol{C}}\boldsymbol{L}')^{-1}\boldsymbol{L}\widehat{\boldsymbol{B}}}{q} \qquad (2.3)$$

where $\mathbf{L}$ has q rows and rank q and $\widehat{\mathbf{C}} = (\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-}$. The sampling distribution of F is approximated by an F-distribution with q numerator and ν denominator df.

In Keselman et al. (1999), ν was set equal to the residual df $n - \text{rank}(\mathbf{X}||\mathbf{Z})$, where n is the total sample size. Since this usually resulted in liberal tests, we consider a Satterthwaite approximation, computed as

follows. First, compute the spectral decomposition $\mathbf{L}\widehat{\mathbf{C}}\mathbf{L}' = \mathbf{P}'\mathbf{D}\mathbf{P}$, where **P** is an orthogonal matrix of eigenvectors and **D** is a diagonal matrix of eigenvalues, both of dimension $q \times q$. Define l_{m} to be the mth row of **PL**, and let

$$\nu_{\mathrm{m}} = \frac{2(D_{\mathrm{m}})^2}{g'_{\mathrm{m}}\mathbf{A}\, g_{\mathrm{m}}}$$

where

- D_{m} is the mth diagonal element of **D**.
- g_{m} is the gradient of $l_{\mathrm{m}}\mathbf{C}\, l'_{\mathrm{m}}$ with respect to θ evaluated at $\widehat{\theta}$, where

$\mathbf{C} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}$, θ is the vector of unknown parameters in **V** and $\widehat{\theta}$ is its

REML estimate.

- **A** is the asymptotic variance-covariance matrix of $\widehat{\theta}$ obtained from the

second derivative matrix of the likelihood equations.

Then let

$$S = \sum_{m=1}^{q} \frac{\nu_{\mathrm{m}}}{\nu_{\mathrm{m}} - 2}\, I\,(\nu_{\mathrm{m}} > 2)$$

where the indicator function eliminates terms for which $\nu_{\mathrm{m}} \leq 2$. The df for F are then computed as

$$\nu = \frac{2S}{S - q}$$

provided $S > q$; otherwise ν is set to zero. This procedure is a generalization of the Satterthwiate methods described by Giesbrecht and Burns (1985), McLean and Sanders (1988), and Fai and Cornelius (1996).

3. METHODS OF THE SIMULATION

The simplest of the higher-order repeated measures designs involves a single between-subjects factor and a single within-subjects factor, in which subjects ($i = 1, \ldots, n_j$, $\Sigma n_j = N$) are selected randomly 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 procedures investigated in this paper assume the Y_{ijk} are normally distributed. For this repeated measures design, researchers are typically interested in testing for main (K) and interaction ($J \times K$) effects.

3.1 Test Statistics

Since we will compare our results with the generally robust WJ procedure found by Keselman et al. (1999) we, for completeness, present this procedure here as well. The test statistics examined therefore included:

(a) WJ (Johansen, 1980; Keselman et al., 1993). Consider the repeated measures design just described, however, allow $\boldsymbol{\Sigma}_j \neq \boldsymbol{\Sigma}_{j'}$, $j \neq j'$. Suppose under these model assumptions that we wish to test the hypothesis:

$$H_o: \mathbf{C}\boldsymbol{\mu} = \mathbf{0} \tag{3.1.1}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_J')'$, $\boldsymbol{\mu}_j = (\mu_{j1}, \ldots, \mu_{jK})'$, $j = 1, \ldots, J$, and $\mathbf{C}$ is a full rank contrast matrix of dimension $r \times JK$. Then an approximate df multivariate Welch (Welch, 1947, 1951)-James (James, 1951, 1954)-type statistic according to Johansen (1980) and Keselman et al. is

$$T_{WJ} = (\mathbf{C}\overline{\mathbf{Y}})'(\mathbf{C}\mathbf{S}\mathbf{C}')^{-1}(\mathbf{C}\overline{\mathbf{Y}}), \tag{3.1.2}$$

where $\overline{\mathbf{Y}} = (\overline{\mathbf{Y}}_1', \ldots, \overline{\mathbf{Y}}_J')'$, with $E(\overline{\mathbf{Y}}) = \boldsymbol{\mu}$, and the sample covariance matrix of $\overline{\mathbf{Y}}$ is $\mathbf{S} = \text{diag}(\mathbf{S}_1/n_1, \ldots, \mathbf{S}_J/n_J)$, where $\mathbf{S}_j$ is the sample variance-covariance matrix of the j-th grouping factor. T_{WJ}/c is

distributed, approximately, as an F variable with df $f_1 = r$, and $f_2 = r(r+2)/(3A)$, and c is given by $r + 2A - 6A/(r + 2)$, with

$$A = \frac{1}{2}\sum_{j=1}^{J}[\mathrm{tr}\{\mathbf{S}\,\mathbf{C}'(\mathbf{C}\mathbf{S}\,\mathbf{C}')^{-1}\,\mathbf{C}\,\mathbf{Q}_j\}^2 + \{\mathrm{tr}\,(\mathbf{S}\,\mathbf{C}'(\mathbf{C}\mathbf{S}\,\mathbf{C}')^{-1}\,\mathbf{C}\,\mathbf{Q}_j)\}^2]/(n_j - 1)\,. \tag{3.1.3}$$

The matrix $\mathbf{Q}_j$ is a block diagonal matrix of dimension $JK \times JK$, corresponding to the j-th group. The (s,t)-th block of $\mathbf{Q}_j = \mathbf{I}_{K\times K}$ if $s = t = j$ and is $\mathbf{0}$ otherwise. In order to obtain the main and interaction tests with the WJ procedure let $\mathbf{C}'_{K-1}$ be a $[(K-1)\times K]$ contrast matrix and let $\mathbf{C}_{J-1}$ be similarly defined. A test of the main effect can be obtained by letting $\mathbf{C} = \mathbf{1}_J \otimes \mathbf{C}_{K-1}$, where $\mathbf{1}_J$ is the $(j \times 1)$ unit vector and $\otimes$ denotes the Kronecker product. The contrast matrix for a test of the interaction effect is $\mathbf{C} = \mathbf{C}_{J-1} \otimes \mathbf{C}_{K-1}$. Lix and Keselman (1995) present a SAS/IML (SAS Institute, 1989) program that can be used to compute the WJ test for any repeated measures design that does not contain quantitative covariates nor has missing values.

(b) Mixed-model Analyses. Several PROC MIXED Satterthwaite F tests were investigated. We obtained an F test (AIC) in which the covariance structure was selected according to the best Akaike value and two F tests where a prespecified covariance structure was fit to the data. Specifically, for each structure investigated we obtained an F test based on the within heterogeneous version of that structure (WH) and an F test based on the correct combined within and between-heterogeneous form of the investigated structure (WBH). For example, when sampling data from an AR structure, we fit both a within-subjects heterogeneous version of AR [with SAS syntax TYPE = ARH(1)] and a within and between heterogeneous version of AR [with SAS syntax TYPE = ARH(1) GROUP = J)]. The first of these tests was performed in order to assess how rates of error are affected when the correct structure is selected but between group heterogeneity is not accounted for when the F test is

computed. In addition, we obtained three Satterthwaite F tests assuming particular covariance structures, ARH_j, RC_j and UN_j where the subscript $_j$ denotes between-subjects heterogeneity (i.e., allowing for unequal covariance matrices across groups). These three F tests presumed both within-subjects and between-subjects heterogeneity.

3.2 Study Variables

The test statistics for testing repeated measures main and interaction effect hypotheses were examined for balanced and unbalanced designs containing one between-subjects and one within-subjects factor; there were three and four levels of these factors, respectively. Because substantial computing time was required for data simulation and analysis, selected combinations of six factors (we ran enough conditions to determine the effects of the manipulated variables) were investigated which included: (a) type of population covariance structure, (b) equal and unequal covariance structures, (c) various cases of total sample size, (d) equal and unequal group sizes, (e) positive and negative pairings of covariance matrices and group sizes, and (f) normal and nonnormal data.

The three types of covariance structures we used to generate simulated data were: (a) UN, (b) ARH, and (c) RC, all of which exhibited within-subject heterogeneity (see Littell et al., 1996; SAS, 1995; Wolfinger, 1993, 1996). The ARH models were heterogeneous first-order autoregressive models, while in the RC models the intercepts and slopes in a simple regression (time point as the independent variable) were random. The variance-covariance values for the three structures examined were:
(a) UN:

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 8. & 6.5243544 & 5.8554383 & 3.3929186 \\ & 10. & 7.087869 & 4.0195836 \\ & & 10. & 5.4032272 \\ & & & 12. \end{bmatrix}; \epsilon = .75$$

(b) ARH:

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 8. & 6.5293185 & 4.7664025 & 3.8115726 \\ & 10. & 7.3 & 5.837627 \\ & & 10. & 7.9967493 \\ & & & 12. \end{bmatrix}; \epsilon = .75$$

(c) RC:

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 2.7301562 & 2.3002343 & 2.8703124 & 3.4403904 \\ & 4.2803124 & 4.2603903 & 5.2404688 \\ & & 6.6504686 & 7.0405469 \\ & & & 9.8406248 \end{bmatrix}; \epsilon = .75$$

For each of the preceding structures, equal as well as unequal between-subjects (groups) covariance matrices were investigated. When unequal, the matrices were multiples of one another, namely $\boldsymbol{\Sigma}_1 = \frac{1}{3}\boldsymbol{\Sigma}_2$, and $\boldsymbol{\Sigma}_3 = \frac{5}{3}\boldsymbol{\Sigma}_2$. This degree and type of covariance heterogeneity was selected because Keselman and Keselman (1990) found that, of the conditions they investigated, it resulted in the greatest discrepancies between the empirical and nominal rates of Type I error and, therefore, was a condition under which the effects of covariance heterogeneity could readily be examined.

Based on the belief that applied researchers work with data that is characterized by both within and between-subjects heterogeneity, eleven covariance structures were fit with PROC MIXED for the AIC test; that is, we allowed AIC to select a structure from among eleven possible structures. These structures were: (a) CS, (b) UN, (c) AR, (d) HF (spherical, Huynh & Feldt, 1976) (e) CSH (heterogeneous CS), (f) ARH, (g) RC, (h) UN_j, (i) HF_j, (j) ARH_j, and (k) RC_j, where the j subscript indicates that covariance matrices are not equal across groups. Thus, we

allowed PROC MIXED to select from among homogeneous, within heterogeneous, and within and between heterogeneous structures.

The test statistics were investigated when the number of observations across groups were equal or unequal. Based on the findings reported by Keselman et al. (1993) and Wright (1995), we considered a number of cases of total sample size: $N = 30$, $N = 45$, and $N = 60$. These values were also chosen because sample sizes of 60 or less were common (53%) in the 226 between by within-subjects repeated measures design studies published in 1994 in prominent educational and psychological journals according to the survey conducted by Kowalchuk et al. (1996). For each value of N, both a moderate and substantial degree of group size inequality were typically investigated. The moderately unbalanced group sizes had a coefficient of sample size variation (C) equal to $\simeq .16$, while for the more disparate cases $C \simeq .33$, where C is defined as $(\Sigma_j(n_j - \overline{n})^2/J)^{\frac{1}{2}}/\overline{n}$, and $\overline{n}$ is the average group size. The $C \simeq .16$ and $C \simeq .33$ unequal group sizes cases were respectively equal to: (a) 8, 10, 12 and 6, 10, 14 ($N = 30$), (b) 12, 15, 18 and 9, 15, 21 ($N = 45$), and (c) 16, 20, 24 and 12, 20, 28 ($N = 60$).

Positive and/or negative pairings of these unequal group sizes and unequal covariance matrices were investigated. A positive pairing referred to the case in which the largest n_j was associated with the covariance matrix containing the largest element values; a negative pairing referred to the case in which the largest n_j was associated with the covariance matrix with the smallest element values. In short, for each value of N, four pairings of unequal covariance matrices and unequal group sizes were investigated: moderately and very unequal n_js which were both positively and negatively paired with the unequal $\boldsymbol{\Sigma}_j$s. We chose to investigate these pairings of group sizes and covariance matrices because they, for many test statistics, typically result in conservative and liberal rates of Type I error, respectively, and thus could also produce these same effects for the statistics investigated in our study.

Rates of Type I error were collected when the simulated data were obtained from multivariate normal or multivariate nonnormal distributions. The algorithm for generating the multivariate normal data can be found in Keselman et al (1993). The nonnormal distribution was a multivariate lognormal distribution with marginal distributions based on $Y_{ijk} = d_j\{\exp(X_{ijk}) - E[\exp(X_{ijk})]\} + \mu_{ij}$ $(i = 1, \ldots, n_j)$ where X_{ijk} is distributed as N(0, .25) and $d_j = 1$, 1/3, or 5/3; this distribution has skewness (γ_1) and kurtosis (γ_2) values of 1.75 and 5.90, respectively. The procedure for generating the multivariate lognormal data is based on Johnson, Ramberg, and Wang (1982) and is presented in Algina and Oshima (1994). This particular type of nonnormal distribution was selected since applied data, particularly in the behavioral sciences, typically have skewed distributions (Micceri, 1989; Wilcox, 1994). Furthermore, Sawilowsky and Blair (1992) found in their Monte Carlo investigation of the two independent samples t test that only distributions with extreme degrees of skewness (e.g., $\gamma_1 = 1.64$) affected Type I error control. In addition, Algina and Oshima (1995) found that tests for mean equality are affected when distributions are lognormal and homogeneity assumptions are not satisfied. Thus, we felt that our approach to modeling skewed data would adequately reflect conditions in which the tests might not perform optimally.

Finally, an issue considered in the current investigation, and not by Keselman et al. (1993) nor Wright (1995), but was addressed by Keselman et al. (1999), was nonsphericity. In our investigation the sphericity index (ϵ) was set at 0.75. When $\epsilon = 1.0$, sphericity is satisfied and for the J $\times$ K design the lower bound of $\epsilon = 1/(K - 1)$. We chose to set the elemental values of $\boldsymbol{\Sigma}_j$ such that $\epsilon = 0.75$ because we wanted a value that is considered realistic for applied data sets (see Huynh & Feldt, 1976). We also wanted $\epsilon \neq 1.0$ because Algina and Keselman (1997) found that the rate of Type I error for the WJ test does vary somewhat with the value of ϵ

even though it is a multivariate statistic and should not be dependent upon the data conforming to sphericity.

One thousand replications of each condition were performed using a .05 significance level. It is important to note that due to extremely demanding processing time requirments the number of simulations per condition was limited to 1000.

4. RESULTS

4.1 Type I Error Rates

To evaluate the particular conditions under which a test was insensitive to assumption violations, Bradley's (1978) liberal criterion of robustness was employed. According to this criterion, in order for a test to be considered robust, its empirical rate of Type I error ($\widehat{\alpha}$) must be contained in the interval $0.5\alpha \leq \widehat{\alpha} \leq 1.5\alpha$. Therefore, for the five percent level of significance used in this study, a test was considered robust in a particular condition if its empirical rate of Type I error fell within the interval $.025 \leq \widehat{\alpha} \leq .075$. Correspondingly, a test was considered to be nonrobust if, for a particular condition, its Type I error rate was not contained in this interval. In the tables, bolded entries are used to denote these latter values. We chose this criterion since we feel that it provides a reasonable standard by which to judge robustness. Readers can choose to interpret the empirical values by seeing if they are above or below 2 standard deviations of the level of significance by using the usual binomial formula; we, however, chose to use the simpler Bradley rule. In our opinion, applied researchers should be comfortable working with a procedure that controls the rate of Type I error within these bounds, if the procedure limits the rate across a wide range of assumption violation conditions.

The results presented are for selected combinations of the six factors investigated which included (a) type of population covariance structure,

(b) equal and unequal covariance structures, (c) various cases of total sample size, (d) equal and unequal group sizes, (e) positive and negative pairings of covariance matrices and group sizes, and (f) normal and nonnormal data. The conditions presented, however, sufficiently demonstrate differences in the operating characteristics of the test statistics. Since the Type I error rates were well controlled when group sizes were equal, the selected combinations of conditions that are presented in this section are for cases of unequal group sizes.

Normally Distributed Data. Table 2 contains rates (%) of error for the test of the repeated measures main and interaction effect when data were obtained from a normal distribution. The tabled values indicate that the Satterthwaite F tests which were based on a covariance structure selected by AIC were adversely affected when covariance matrices and group sizes were unequal. For example, when testing the repeated measures main effect, the empirical percentage of Type I error attained a value of 8% for the AIC based Satterthwaite F test. The empirical rates were as large as 9.9% for the interaction test. On the other hand, the WJ results reported by Keselman et al. (1999), and reproduced in Table 2, indicate that this procedure was less affected by heterogeneity. In particular, the WJ rates were on occasion liberal for the repeated measures interaction effect test, though controlled, when sample size was larger.

The remaining values in this table indicate that: (a) if the PROC MIXED Satterthwaite F tests assumed the correct covariance structure but only assumed within-subjects heterogeneity (WH conditions), the rates were distorted; that is, the WH F test values, which were based on tests that do not allow for between-subjects heterogeneity, ranged from 0.9% to 17.7% across the two repeated measures effects, and (b) if the Satterthwaite F tests assumed the correct covariance structure and also allowed for both within- and between-subjects heterogeneity (WBH conditions), the rates of error, except under the UN covariance structure, were controlled. In addition, the rates were not only generally well

controlled when the correct WBH structure was used to compute the F tests, but as well, controlled when either a ARH_j or RC_j structure was assumed. On the other hand, assuming a UN_j structure did not generally provide Type I error protection. Three of the nine main effect rates were liberal while all interaction rates exceeded Bradley's (1978) upper liberal bound, attaining values as large as 13.9%.

Lognormal Distributed Data. Table 3 contains Type I error rates (%) for the tests of the repeated measures main and interaction effect when data were obtained from the lognormal distribution. Because Keselman et al. (1993) and Algina and Keselman (1998) indicated that sample sizes must be larger than those required when data are normally distributed to obtain a robust test with the WJ procedure, the reported rates are for $N = 45$ and $N = 60$. The rates of error were similar to those reported in Table 2. That is, the AIC based Satterthwaite F test again had distorted rates of error; AIC based F tests, though never resulting in conservative values, frequently resulted in liberal values. Once again the WJ rates were generally well controlled, particularly for the larger sample size case investigated. It is important to note, that WJ ceased to be liberal [i.e., 5.9%] for the test of the interaction effect under an RC structure when the smallest of the group sizes was increased from $n_1 = 12$ to $n_1 = 16$ (i.e., $n_1 = 16$, $n_2 = 20$, $n_3 = 24$). Satterthwaite F tests that did not allow for between-subjects heterogeneity resulted in conservative or liberal rates of error while F tests which allowed for both within and between heterogeneity had rates which were well controlled. Lastly, always fitting an ARH_j or RC_j structure provided good Type I error control while always fitting an UN_j structure generally did not.

Summary of Type I Error Results. Our results indicate that, if the correct covariance structure is used, 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 data are nonnormal in form, covariance matrices are unequal, and the design is unbalanced. The disadvantage however is that researchers should not expect to find the correct covariance structure with the AIC criterion. On the other hand, our results indicate that, for the conditions we investigated, Satterthwaite F tests that allow within- and between-subjects heterogeneity, specifically those that fit either a random coefficients or autoregressive covariance structure, always provided effective Type I error protection. It could be argued that researchers could expect Type I error protection if they used PROC MIXED Satterthwaite F tests and always fit either of these two structures. However, one really does not know how such a strategy would work under conditions not investigated in this study.

Interestingly, though the mixed-model procedure with the UN_j structure for the dispersion matrices and the WJ procedure are based on the same statistical model, they provided very different Type I error control; this finding is based on the fact that they differ in terms of the distribution used to approximate the sampling distribution of the statistics. In order to shed further light on the relative advantages/disadvantages of mixed-model analyses versus a WJ analysis we decided to compare the statistical power of the two approaches. This phase of the study was intended to see whether it would be necessary to probe further into Type I error comparisons. That is, it would not be necessary to further explore Type I error differences if it could be shown that the WJ test is comparable in power to the PROC MIXED Satterthwaite F tests.

4.2 Power

In this phase of the investigation we simulated power rates for the repeated measures main effect. In particular, we examined a maximum range configuration of nonnull main effect means. That is, for each of the three groups in our 3×4 repeated measures design, one element of the mean vector was $-\mu$, a second element was μ, and the remaining two elements were zero. With the maximum range mean configuration there are six possible permutations defined by the pattern of zero and nonzero elements: (a) ($-\mu$, 0, 0, μ), (b) ($-\mu$, 0, μ, 0), (c) ($-\mu$, μ, 0, 0), (d) (0, $-\mu$, 0, μ), (e) (0, $-\mu$, μ, 0), and (f) (0, 0, $-\mu$, μ). Values of μ were selected to give a .50 a priori target power value for WJ across the six maximum range configurations based on a noncentrality parameter postulated by Algina and Keselman (1998). We varied the permutation of the means because for a fixed sample size and level of significance the relative power of univariate and multivariate approaches depends upon the specific nonnull vector of repeated measures means, the covariance matrix that describes the variances and covariances among the scores for the levels of the repeated measures factor, and the *relationship* between these two characteristics of the population.

As in the Type I error phase of this investigation, we also varied (a) sample size equality/inequality (20, 20, 20 and 16, 20, 24), (b) covariance equality/inequality, (c) pairing of covariances and group sizes, (d) form of covariance structure, and (d) shape of underlying distribution.

Since each condition of the simulation took from 12 to 72 hours to execute we collected only enough conditions to confidently determine a pattern of results. Tables 4 and 5 contain power rates for WJ and the PROC MIXED Satterthwaite F test which always was based on the correct population covariance structure, for normal and lognormal data, respectively. For example, when sampling from a population where the variances and covariances were UN in form, the Satterthwaite F test was

always based on a UN structure, where as when the correct structure was ARH the F test was always based on an ARH structure.

As expected the PROC MIXED F test was, with one exception, more powerful than the WJ test. However, the power advantage never exceeded 6 percentage points and on average equalled 3.1 percentage points (modal advantage = 3 percentage points). The largest discrepancies (the three cases of 6 percentage points) occurred when the true covariance structure was given by a random coefficients arrangement and data were lognormal. We also examined our data to check whether the power advantage for the mixed-model analysis would remain when the wrong structure was fit to the data. Interestingly, for 16 of the 37 conditions investigated, another within and between heterogeneous structure resulted in a more powerful test than when the correct structure was always fit to the data. However, in these 16 conditions, the power of this alternative structure still never exceeded the WJ power by more than 7 percentage points, and the average superiority was just 4 percentage points in the remaining 21 conditions. WJ, however, was more powerful than the incorrect structure F test by 8 percentage points on average and on occasion exceeded the incorrect structure F test value by 14 percentage points.

5. DISCUSSION

Our investigation compared two of the newer approaches to the analysis of repeated measures effects: the Welch-James multivariate approach that does not require homogeneity of between groups covariance matrices and a mixed-model approach that allows users to select a covariance structure of the data and as well uses Satterthwaite type F tests. Neither approach requires that the data conform to the multisample sphericity assumption associated with the conventional univariate test of significance and as such should offer applied researchers more robust methods of analysis for analyzing within-subjects effects in repeated measures designs.

These two approaches to the analysis of repeated measures effects were compared for their Type I error and power rates for different covariance structures when between group covariance matrices were unequal in balanced and unbalanced designs when data were either normal or nonnormal.

Type I error results indicated that the mixed-model Satterthwaite F tests, as computed by SAS' (1996) PROC MIXED program, resulted in liberal tests when the covariance structure was determined by the Akaike (1974) Information Criterion. The Satterthwaite F tests were robust to covariance heterogeneity and nonnormality in unbalanced designs when they were based on fitting the correct covariance structure. Since the WJ test was also typically robust, we then compared the two approaches for their statistical power. The reader should remember that the PROC MIXED Satterthwaite F tests should have been more powerful because they were based on fitting the correct covariance structure of the data. However, if the power values for WJ were not sufficiently smaller than these correctly fit tests this would suggest that it could also be adopted since applied researchers would not know the correct covariance structure of their data. Furthermore, applied researchers would not necessarily be able to find the correct structure with the Akaike criterion.

As expected Satterthwaite F tests based on fitting the correct covariance structure were, with one exception, always more powerful to detect a repeated measures main effect, however, the power advantages were small, on average they were only larger by 3.1 percentage points. Since the power advantage was so small we feel that the Welch-James approach to the analysis of repeated measures effects is a viable procedure worth adopting particularly when designs are unbalanced and data are not likely to be spherical nor homogeneous, since applied users need not know the correct covariance structure of their data nor can they rely on currently avaliable methods to determine this structure. However, we also ackowledge that adopting a mixed-model analysis always fitting a heterogeneous first-order autoregressive or random coefficients model might be a reasonable strategy as well to the analysis of repeated measurements. Readers should also keep in mind however, that the WJ can, on occassion, be nonrobust when covariance matrices and group sizes are negatively paired.

Using either of these procedures addresses deviations from multisample sphericity. Users must still examine the distribution of their data and be concerned about the effects of influential data points. In addition, if interpretation of the structural model for the dispersion matrix will advance the goals of the research, the researcher may want to report parameter estimates for the model, even if the tests on means are conducted. However, the researcher should keep in mind that, at least with the sample sizes included in our simulation, there will be some question about the adequacy of the structural model that is fit with the data.

Finally, though our results relate to the conditions we varied in our investigation, there is no reason to believe, particularly given the findings reported by Algina and Keselman (1998), that they would not apply to other mean configurations as well.

ACKNOWLEDGEMENTS

This research was supported by a Social Sciences and Humanities Research Council grant (#410-95-0006).

BIBLIOGRAPHY

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

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

Algina, J., and 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., and 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., and 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.

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

Brown, M.B., and Forsythe, A.B. (1974). "The small sample behavior of some statistics which test the equality of several means," *Technometrics*,16, 129-132.

Bryk, A. S., and Raudenbush, S. W. (1992). "*Hierarchial linear models*," Newbury Park, CA: Sage Publications, Inc.

Fai, A. H. T., and Cornelius, P. L. (1996). "Approximate F-tests of

multiple degree of freedom hypotheses in genarlized least squares analyses of unbalanced split-plot experiments," *Journal of Statistical Computation and Simulation*, 54, 363-378.

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

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

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

Huynh, H., and Feldt, L. (1970). "Conditions under which mean square ratios in repeated measurements designs have exact F distributions," *Journal of the American Statistical Association*, 65,
1582-1589.

Huynh, H., and 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.

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., and Schluchter, M. D. (1986). "Unbalanced repeated -measures models with structured covariance matrices," *Biometrics*, 42, 805-820.

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

Johnson, M. F., Ramberg, J. S., and Wang, C. (1982). "The Johnson

translation system in Monte Carlo studies," *Communications in Statistics-Simulation and Computation*, 11, 521-525.

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

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

Keselman, H. J., Carriere, K. C., and 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., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, 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(3), 350-386.

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

Khuri, A. I., Mathew, T., and Sinha, B. K. (1998). "*Statistical tests for mixed linear models*," New York: Wiley.

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

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

Little, R.J.A. (1994). "A class of pattern-mixture models for normal incomplete data," *Biometrics*, 81, 471-483.

Little, R.J.A. (1995). "Modeling the drop-out mechanism in repeated measures studies," *Journal of the American Statistical Association*, 90, 1112-1121.

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

McLean, R. A., and 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.

Mendoza, J. L. (1980). "A Significance test for multisample sphericity," *Psychometrika*, 45, 495-498.

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

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

Rouanet, H., and 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.

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

SAS Institute. (1996). "*SAS/STAT Software: Changes and Enhancements through Release 6.11*," Cary, NC: Author.

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

Sawilowsky, S. S., and Blair, R. C. (1992). "A more realistic look at the robustness and Type II error probabilities of the *t* test to

departures from population normality," *Psychological Bulletin*, 111, 352-360.

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

Welch, B. L. (1938). "The significance of the difference between two means when the population variances are unequal," *Biometrika*, 29, 350-362.

Welch, B. L. (1947). "The generalization of Students' problems 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.

Wilcox, R. R. (1994). "A one-way random effects model for trimmed means," *Psychometrika*, 59, 289-306.

Wolfinger, R. (1993). "Covariance structure selection in general mixed models," *Communication in Statistics-Simulation*, 22, 1079-1106.

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

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