British Journal of Mathematical and Statistical Psychology (2001), **54**, 1–20 *Printed in Great Britain*

The analysis of repeated measures designs: A review

H. J. Keselman*

University of Manitoba, Canada

James Algina

University of Florida, USA

Rhonda K. Kowalchuk

University of Manitoba, Canada

Repeated measures ANOVA can refer to many different types of analysis. Specifically, this vague term can refer to conventional tests of significance, one of three univariate solutions with adjusted degrees of freedom, two different types of multivariate statistic, or approaches that combine univariate and multivariate tests. Accordingly, it is argued that, by only reporting probability values and referring to statistical analyses as repeated measures ANOVA, authors convey neither the type of analysis that was used nor the validity of the reported probability value, since each of these approaches has its own strengths and weaknesses. The various approaches are presented with a discussion of their strengths and weaknesses, and recommendations are made regarding the 'best' choice of analysis. Additional topics discussed include analyses for missing data and tests of linear contrasts.

1. Introduction

Papers intended to bring to the attention of applied researchers the latest developments in data analysis strategies, which are generally introduced in statistical journals, are not uncommon in the psychological literature (see, for example, Algina & Coombs, 1996; Keselman & Keselman, 1993; Keselman, Rogan & Games, 1981; McCall & Appelbaum, 1973). Since, as McCall and Appelbaum note, repeated measures (RM) designs are one of the most common research paradigms in psychology, it is not surprising that articles of this nature pertaining to the analysis of repeated measurements have appeared periodically in our literature; for example, McCall & Appelbaum, Hertzog & Rovine (1985), Keselman & Keselman (1988), and Keselman & Algina (1996) have provided updates on analysis strategies for RM designs. Because new analysis strategies for the analysis of repeated measurements have recently

* Requests for reprints should be addressed to Professor H. J. Keselman, Department of Psychology, University of Manitoba, 190 Dysart Road, Winnipeg, Manitoba, Canada R3T 2N2 (e-mail: kesel@ms.umanitoba.ca).

appeared in the quantitatively oriented literature, we thought it timely to once again provide an update for psychological researchers.

In addition to introducing procedures that have appeared in the last five to ten years, we present a brief review of procedures that are not so new, since recent evidence suggests that even these are not commonly adopted by behavioural science researchers (see Keselman *et al.*, 1998). It is important to review these procedures since they are better (i.e., generally better control the probability of a Type I error) than the conventional univariate method of analysis and, moreover, because they provide an important theoretical link to the most recent approaches to the analysis of repeated measurements.

RM designs and analysis of variance (ANOVA) statistics are often used by behavioural science researchers to assess treatment effects (Keselman *et al.*, 1998). However, ANOVA statistics are, according to results reported in the literature, sensitive to violations of the derivational assumptions on which they are based, particularly when the design is unbalanced (i.e., group sizes are unequal) (Collier, Baker, Mandeville & Hays, 1967; Keselman & Keselman, 1993; Keselman, Keselman & Lix, 1995; Rogan, Keselman & Mendoza, 1979). Specifically, the conventional univariate method of analysis assumes that the data have been obtained from populations that have the well-known normal (multivariate) form, that the degree of variability (covariance) among the levels of the RM variable conforms to a spherical pattern, and that the data conform to independence assumptions. Since the data obtained in many areas of psychological inquiry are not likely to conform to these requirements and are frequently unbalanced (see Keselman *et al.*, 1998), researchers using the conventional procedure will erroneously claim treatment effects when none are present, thus filling their literatures with false positive claims.

However, many other ANOVA-type statistics are available for the analysis of RM designs which under many conditions will be insensitive (i.e., robust) to violations of the assumptions associated with the conventional tests or do not depend on the conventional covariance assumption (i.e., multisample sphericity). These ANOVA-type procedures include univariate tests with adjusted degrees of freedom (df), multivariate test statistics, statistics that do not depend on the conventional assumptions of multisample sphericity, and hybrid types of analyses that involve a combining of the univariate and multivariate approaches.

Another fly in this ointment relates to the vagueness associated with the descriptors typically used by behavioural science researchers to describe the statistical tests employed in the analysis of treatment effects in RM designs and the use of the associated probability value (p) to convey success or failure of the treatment. That is, describing the analysis as 'repeated measures ANOVA' does not tell the reader which repeated measures ANOVA technique was used to test for treatment effects. In addition, just reporting a p-value does not give enough information (e.g., df of the statistic) for the reader to determine what type of RM analysis was used, and thus calls into question the legitimacy of the authors' claims regarding the likelihood that the result was due to the manipulated variable and not because the test was used improperly, (i.e., when the assumptions to the test have not been met). Thus, the aim of this paper is to describe briefly how RM designs are typically analysed by researchers, and to survey the strengths and weaknesses of other ANOVA-type tests for assessing treatment effects in RM designs and thus comment on the validity of the associated p-values.

The reader should note that although we typically present the test statistic for the various approaches, they need not be examined with an eye for obtaining a numerical solution; numerical results can be obtained with specified software.

2. Older data analysis approaches

2.1. Conventional univariate tests of significance

The simplest of the between-by within-subjects RM designs involves a single between-subjects grouping factor and a single within-subjects RM factor, in which subjects ($i = 1, \ldots, n_j$, $\sum_j n_j = N$) 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 ($k = 1, \ldots, K$). In this design, the RM data are modelled by assuming that the observational vectors $\mathbf{Y}_{ij} = (Y_{ij1}\, Y_{ij2} \ldots Y_{ijk})'$ are normal, independent and identically distributed within each level j, with common mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$.

Tests of the within-subjects main and interaction effects traditionally have been accomplished by the respective use of the conventional univariate F statistics,

$$F_K = MS_K/MS_{K\times S/J} \sim F[\alpha; (K-1), (N-J)(K-1)] \tag{1}$$

and

$$F_{J\times K} = MS_{J\times K}/MS_{K\times S/J} \sim F[\alpha; (J-1)(K-1), (N-J)(K-1)], \tag{2}$$

where $\sim$ is to be read as 'is distributed as'. The validity of these tests rests on the assumptions of normality, independence of errors, and homogeneity of the treatment-difference variances—(i.e., sphericity (Huynh & Feldt, 1970; Rogan *et al.*, 1979; Rouanet & Lepine, 1970). Specifically, sphericity is satisfied if and only if $\mathbf{C}'\boldsymbol{\Sigma}\mathbf{C} = \lambda\mathbf{I}_{(K-1)}$, where $\mathbf{C}$ is a normalized (i.e., unit-length) matrix of $K-1$ orthogonal contrasts among the K repeated measurements, $\boldsymbol{\Sigma}$ is the population covariance matrix, λ is a positive scalar, and $\mathbf{I}$ is an identity matrix of order $K-1$.[1] As the diagonal and off-diagonal elements of $\mathbf{C}'\boldsymbol{\Sigma}\mathbf{C}$ equal the variances and covariances of the $K-1$ orthogonal contrasts, the sphericity assumption is satisfied if and only if the $K-1$ contrasts are independent and equally variable. Further, the presence of a between-subjects grouping factor requires that the data meet an additional assumption, namely, that the covariance matrices of these treatment differences are the same for all levels of this grouping factor. Jointly, these two assumptions have been referred to as multisample sphericity (Huynh, 1978; Mendoza, 1980; for another description of multisample sphericity, see Hertzog & Rovine, 1985, pp. 792–793). The F tests of simple RM designs containing only within-subjects variables, that is, with no between-subjects grouping variables, also depend on the sphericity assumption; however, they do not require multisample sphericity.

When the assumptions for the conventional tests have been satisfied they provide a valid test of their respective null hypotheses and are uniformly most powerful for detecting any treatment effects that are present. These traditional tests are easily obtained using the major statistical packages, such as SAS (SAS Institute, 1999) and SPSS (Norušis, 1993). Thus, when assumptions are known to be satisfied, psychological researchers can adopt the conventional procedures and report the associated p-values since under these conditions these values are an accurate reflection of the probability of rejecting the null hypothesis by chance when the null hypothesis is true.

However, McCall & Appelbaum (1973) provide a very good illustration as to why in many areas of psychology (e.g., developmental, learning), the covariances between the

[1] See Rogan *et al.* (1979) for an example that shows the form of a contrast matrix ($\mathbf{C}$) and the computation of $\mathbf{C}'\boldsymbol{\Sigma}\mathbf{C} = \lambda\mathbf{I}$.

levels of the RM variable will not conform to the required covariance pattern for a valid univariate F test. They use an example from developmental psychology to illustrate this point. Specifically, adjacent-age assessments typically correlate (i.e., covary) more highly than developmentally distant assessments (e.g., 'IQ at age 3 correlates .83 with IQ at age 4 but .46 with IQ at age 12'); this type of correlational (i.e., covariance) structure does not correspond to a spherical covariance structure. That is, for many psychological paradigms successive or adjacent measurement occasions are more highly correlated than non-adjacent measurement occasions, with the correlation between these measurements decreasing the farther apart the measurements are in the series (Danford, Hughes & McNee, 1960; Winer, 1971). Indeed, as McCall and Appelbaum (1973) note: 'Most longitudinal studies using age or time as a factor cannot meet these assumptions.' McCall and Appelbaum also indicate that the covariance pattern found in learning experiments is not likely to conform to a spherical pattern. As they note, 'experiments in which change in some behaviour over short periods of time is compared under several different treatments often cannot meet covariance requirements' (p. 403).

The result of applying the conventional tests of significance to data that do not conform to the assumptions of multisample sphericity will be that too many null hypotheses will be falsely rejected (Box, 1954; Collier *et al.*, 1967; Imhof, 1962; Kogan, 1948; Stoloff, 1970). Furthermore, as the degree of non-sphericity increases, the conventional repeated measures F tests becomes increasingly inflated (Noe, 1976; Rogan *et al.*, 1979). For example, the results reported by Collier *et al.* and Rogan *et al.* indicate that Type I error rates can approach 10% for both the test of the RM main and interaction effects when sphericity does not hold. Thus, p-values are not accurate reflections of the observed statistics occurring by chance under their null hypotheses. Rather, they indicate the probability of the statistics arising under some other distribution, a distribution characterized by a sphericity parameter that is not presumed by the conventional tests of significance. Hence, using these p-values to ascertain whether the treatment has been successful or not will give a biased picture of the nature of the treatment.

2.2. The multivariate approach

The multivariate test of the RM main effect in a simple (no between-subjects factors) or between- by within-subjects design is performed by creating $K-1$ difference variables. The null hypothesis that is tested, using Hotelling's (1931) T^2 statistic, is that the vector of population means of these $K-1$ difference variables equals the null vector (see McCall & Appelbaum, 1973, for a fuller discussion and numerical example). The upper $100(1-\alpha)$ percentage points of the T^2 distribution can be obtained from the relationship

$$F = \frac{N-J-K+2}{(N-J)(K-1)}T^2 \sim F[\alpha; K-1, N-J-K+2]. \tag{3}$$

The multivariate test of the within-subjects interaction effect, on the other hand, is a test of whether the population means of the $K-1$ difference variables are equal across the levels of the grouping variable. A test of this hypothesis can be obtained by conducting a one-way multivariate ANOVA, where the $K-1$ difference variables are the dependent variables and the grouping variable (J) is the between-subjects independent variable. When $J > 2$ four popular multivariate criteria are: (1) Wilks's (1932) likelihood ratio; (2) the Pillai–Bartlett trace statistic (Pillai, 1955; Bartlett, 1939); (3) Roy's (1953) largest root

criterion; and (4) the Hotelling–Lawley trace criterion (Hotelling, 1951; Lawley, 1938). When $J = 2$, all criteria are equivalent to Hotelling's T^2 statistic.

Valid multivariate tests of the RM hypotheses in between- by within-subjects designs, unlike the univariate tests, depend not on the sphericity assumption but only on the equality of the covariance matrices at all levels of the grouping factor as well as normality and independence of observations across subjects. Simple designs, however, in addition to normality and independence assumptions, only require that the covariance matrix be positive definite. Multivariate tests of RM designs hypotheses are easily obtained from the general linear model program associated with each of the two major statistical packages mentioned earlier.

The empirical results indicate that the multivariate test of the RM main effect is generally robust to assumption violations when the design is balanced (or contains no grouping factors) and not robust when the design is unbalanced (Algina & Oshima, 1994; Keselman *et al.*, 1995; Keselman, Algina, Kowalchuk & Wolfinger, 1999*a*, 1999*b*). The interaction test is not necessarily robust even when the group sizes are equal (Olson, 1974). In particular, as was the case with the univariate tests, the multivariate tests are conservative or liberal depending on whether the covariance matrices and group sizes are positively or negatively paired. When positively paired, main as well as interaction effect rates of Type I error, can be less than 1%, while for negative pairings rates in excess of 20% have been reported (see Keselman *et al.*, 1995).

2.3. Univariate tests with adjusted degrees of freedom

When the covariance matrices for the orthonormal variables are equal but the common covariance matrix is not spherical, or when the design is balanced (group sizes are equal) the Greenhouse & Geisser (1959) and Huynh & Feldt (1976) adjusted-df univariate tests are robust alternatives to the conventional tests (see also Quintana & Maxwell, 1994, for other adjusted-df tests).

The Greenhouse and Geisser (GG) $\hat{\varepsilon}$-approximate F test is an approximate-df procedure which refers values of F to an adjusted critical value by modifying the usual numerator and denominator df according to a sample estimate ($\hat{\varepsilon}$) of the unknown sphericity parameter ε. That is,

$$F_K \dot{\sim} F[\alpha; (K-1)\hat{\varepsilon}, (N-J)(K-1)\hat{\varepsilon}] \tag{4}$$

and

$$F_{J\times K} \dot{\sim} F[\alpha; (J-1)(K-1)\hat{\varepsilon}; (N-J)(K-1)\hat{\varepsilon}], \tag{5}$$

where $\dot{\sim}$ is to be read as 'is approximately distributed as' and

$$\hat{\varepsilon} = \frac{[\mathrm{tr}(\mathbf{C}'\mathbf{S}\,\mathbf{C})]^2}{(K-1)\mathrm{tr}[(\mathbf{C}'\mathbf{S}\,\mathbf{C})]^2}, \tag{6}$$

in which $\mathbf{S}$ is the pooled sample covariance matrix which estimates $\boldsymbol{\Sigma}$ and 'tr' is the trace operator.

The Huynh and Feldt (HF) $\tilde{\varepsilon}$-approximate F test (see also Lecoutre, 1991), like the Greenhouse & Geisser (1959) adjustment, refers values of F to the sampling distribution of F based on another sample estimate of ε, one which is intended to be more accurate when

$\varepsilon \geq 0.75$. According to the HF approximation values of F are referred to

$$F_K \sim F[\alpha; (K-1)\tilde{\varepsilon}, (N-J)(K-1)\tilde{\varepsilon}] \tag{7}$$

and

$$F_{J\times K} \sim F[\alpha; (J-1)(K-1)\tilde{\varepsilon}, (N-J)(K-1)\tilde{\varepsilon}], \tag{8}$$

where

$$\tilde{\varepsilon} = \frac{(N-J+1)(K-1)\hat{\varepsilon} - 2}{(K-1)[N-J-(K-1)\hat{\varepsilon}]}. \tag{9}$$

The empirical literature indicates that the GG and HF adjusted-df tests are robust to violations of multisample sphericity as long as group sizes are equal (see Rogan *et al.*, 1979). The *p*-values associated with these adjusted statistics will provide an accurate reflection of the probability of obtaining them by chance under the null hypotheses of no treatment effects. Moreover, SAS (1999) and SPSS (Norušis, 1993) provide GG and HF adjusted *p*-values.

However, the GG and HF adjusted-df tests are not robust when the design is unbalanced (Algina & Oshima, 1994, 1995; Keselman *et al.*, 1995; Keselman & Keselman, 1990; Keselman, Lix & Keselman, 1996). Specifically, the tests are conservative (liberal) when group sizes and covariance matrices are positively (negatively) paired with one another. A positive (negative) pairing refers to the case in which the smallest n_j is associated with the covariance matrix with the smallest (largest) element values. For example, the rates when depressed can be lower than 1% and when inflated higher than 11% (see Keselman *et al.*, 1999*b*).

2.4. The combined approach

Due to the absence of a clear advantage in adopting either an adjusted univariate or multivariate approach, a number of authors have recommended that these procedures be used in combination (Barcikowski & Robey, 1984; Looney & Stanley, 1989). In order to maintain the overall rate of Type I error at α for a test of an RM effect, these authors suggested assessing each of the two tests using an $\alpha/2$ critical value. In this strategy, rejection of an RM effect null hypothesis occurs if either test is found to be statistically significant (see Barcikowski & Robey, 1984, p. 150; Looney & Stanley, 1989, p. 221). Not surprisingly, this approach to the analysis of repeated measurements results in depressed or inflated rates of Type I error when multisample sphericity is not satisfied when the design is unbalanced (see Keselman *et al.*, 1995).

3. Underused and new data analysis approaches

In addition to the Greenhouse & Geisser (1959) and Huynh & Feldt (1976) adjusted-df tests, other adjusted-df tests are available for obtaining a valid test. The test to be introduced now not only corrects for non-sphericity, but also adjusts for heterogeneity of the orthonormalized covariance matrices.

3.1. The Huynh (1978) approximate F tests

Huynh (1978) developed a test of the within-subjects main and interaction hypotheses, the improved general approximation (IGA) test, that is designed to be used when multisample

sphericity is violated. The IGA tests of the within-subjects main and interaction hypotheses are the usual statistics, F_K and $F_{J\times K}$, respectively, with corresponding critical values of $bF[\alpha; h', h]$ and $cF[\alpha; h'', h]$. The parameters of the critical values are defined in terms of the group covariance matrices and group sample sizes. Estimates of the parameters (c, b, h, h', and h'') and the correction due to Lecoutre (1991) are presented in Algina (1994) and Keselman & Algina (1996). These parameters adjust the critical value to take into account the effect that violation of multisample sphericity has on F_K and $F_{J\times K}$. If multisample sphericity holds,

$$bF[\alpha; h'h] = F[\alpha; (K-1), (N-J)(K-1)]$$

and

$$cF[\alpha; h'', h] = F[\alpha; (J-1)(K-1), (N-J)(K-1)].$$

An SAS/IML (SAS Institute, 1999) program is also available for computing this test in any RM design (see Algina, 1997).

The IGA tests have been found to be robust to violations of multisample sphericity, even for unbalanced designs where the data are not multivariate normal in form (see Keselman *et al.*, 1999*b*). This result is not surprising since these tests were specifically designed to adjust for non-sphericity and heterogeneity of the between-subjects covariance matrices. Thus, the p-values associated with the IGA tests of the repeated measures effects are accurate.

3.2. Mixed model analyses

Another procedure that researchers can adopt to test RM effects can be derived from a general formulation for analysing effects in RM models. This approach to the analysis of repeated measurements is a mixed model analysis. Advocates 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. Thus, one can use this procedure to select the most appropriate covariance structure before testing the usual RM hypotheses (e.g., F_K and $F_{J\times K}$). The first of these advantages is typically not a pertinent issue to those involved in controlled experiments, since data in these contexts are rarely missing. The second consideration, however, could be most relevant to experimenters since modelling the correct covariance structure of the data should result in more powerful tests of the fixed-effects parameters.

The linear model underlying the mixed model approach can be written as follows:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{ZU} + \mathbf{E}, \tag{10}$$

where $\mathbf{Y}$ is a vector of response scores, $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices, $\mathbf{B}$ is a vector of unknown fixed-effects parameters, $\mathbf{U}$ is a vector of unknown random effects, and $\mathbf{E}$ is the vector of random errors. 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. The model requires that $\mathbf{U}$ and $\mathbf{E}$ are normally distributed with

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

and

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

Thus, the variance of the response measure is given by

$$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}. \tag{11}$$

Accordingly, one can model $\mathbf{V}$ by specifying $\mathbf{Z}$ and covariance structures for $\mathbf{G}$ and $\mathbf{R}$. Note that the usual general linear model is arrived at by letting $\mathbf{Z} = \mathbf{0}$ and $\mathbf{R} = \sigma^2\mathbf{I}$. The choice of estimation procedure for mixed model analysis and the formation of test statistics is described in Littell, Milliken, Stroup and Wolfinger (1996, pp. 498–502).

The mixed approach, and specifically the PROC MIXED procedure in SAS (SAS Institute, 1995, 1996), allows users to fit various covariance structures for $\mathbf{G}$ and $\mathbf{R}$. For example, some of the covariance structures that can be fitted with PROC MIXED are: (a) compound symmetric (CS), (b) unstructured (UN), (c) spherical (HF), (d) first-order autoregressive (AR1), and (e) random coefficients (RC) (see Wolfinger, 1996, for specifications of these and other covariance structures). The HF structure, as indicated, is assumed by the conventional univariate F tests in the GLM program (SAS Institute, 1999), while the UN structure is assumed by GLM's multivariate tests of the RM effects. The AR1 and RC structures indicate that measurements that are closer in time could be more highly correlated than those farther apart in time. The program allows users even greater flexibility by allowing covariance structures with within-subjects and/or between-subjects heterogeneity to be modelled. In order to select an appropriate structure for one's data, PROC MIXED users can use either an Akaike (1974) or Schwarz (1978) information criterion (see Littell *et al.*, 1996, pp. 101–102).

Keselman *et al.* (1999*a*, 1999*b*) recommend adopting the optional Satterthwaite F tests rather than the default F tests when using PROC MIXED since they are typically robust to violations of multisample sphericity in cases where the default tests are not.

3.3. A non-pooled adjusted-df multivariate test

Since the effects of testing mean equality in RM designs with heterogeneous data are similar to the results reported for independent groups designs, one solution to the problem parallels those found in the context of completely randomized designs. 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. (This is in contrast to the Huynh, 1978, approach which, by use of the conventional univariate F statistics, does pool across heterogeneous sources of variance. The Huynh approach adjusts the critical value to take account of the pooling.)

Consider the RM design described previously, but allow $\mathbf{\Sigma}_j \neq \mathbf{\Sigma}_{j'}$, $j \neq j'$. Suppose under these model assumptions that we wish to test the hypothesis

$$H_0\colon \mathbf{C}\boldsymbol{\mu} = \mathbf{0}, \tag{12}$$

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–James type statistic

(WJ), according to Johansen (1980) and Keselman, Carriere & Lix (1993), is

$$T_{\mathrm{WJ}} = (\mathbf{C}\bar{\mathbf{Y}})'(\mathbf{CSC}')^{-1}(\mathbf{C}\bar{\mathbf{Y}}), \tag{13}$$

where $\bar{\mathbf{Y}} = (\bar{\mathbf{Y}}_1', \ldots, \bar{\mathbf{Y}}_J')'$, with $E(\bar{\mathbf{Y}}) = \boldsymbol{\mu}$, and the sample covariance matrix of $\bar{\mathbf{Y}}$ is $\mathbf{S} = \mathrm{diag}(\mathbf{S}_1/n_1, \ldots, \mathbf{S}_J/n_J)$, where $\mathbf{S}_j$ is the sample variance–covariance matrix of the jth 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

$$A = \frac{1}{2}\sum_{j=1}^{J}[\mathrm{tr}\{\mathbf{SC}'(\mathbf{CSC}')^{-1}\mathbf{CQ}_j\}^2 + \{\mathrm{tr}(\mathbf{SC}'(\mathbf{CSC}')^{-1}\mathbf{CQ}_j)\}^2]/(n_j - 1). \tag{14}$$

The matrix $\mathbf{Q}_j$ is a block diagonal matrix of dimension $JK \times JK$, corresponding to the jth group. The (s,t)th block of $\mathbf{Q}_j$ is $\mathbf{I}_{K\times K}$ 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 $\mathbf{C} = \mathbf{C}_{J-1} \otimes \mathbf{C}_{K-1}$.[2]

The empirical literature indicates that the WJ test is in many instances insensitive to heterogeneity of the covariance matrices and accordingly will provide valid p-values (see Algina & Keselman, 1997; Keselman *et al.*, 1993, 1999*a*, 1999*b*). (As a multivariate statistic, WJ does not require a spherical covariance structure.) Researchers should consider using this statistic when they suspect that group covariance matrices are unequal and they have groups of unequal size. However, to obtain a robust statistic researchers must have reasonably large sample sizes. That is, according to Keselman *et al.* (1993), when $J = 3$, in order to obtain a robust test of the RM main effect hypothesis, the number of observations in the smallest of groups ($n_{\min}$) must be three to four times the number of repeated measurements minus one ($K-1$), while the number must be five or six to one in order to obtain a robust test of the interaction effect. As J increases, smaller sample sizes will suffice for the main effect but larger sample sizes are required to control the Type I error rate for the interaction test (Algina & Keselman, 1997). Though the test statistic cannot be obtained from the major statistical packages, Lix & Keselman (1995) present a SAS/IML (SAS, 1999) program that can be used to compute the WJ test for any RM design (excluding quantitative covariates). The program requires only the user to enter the data, the number of observations per group (cell), and the coefficients of one or more contrast matrices that represent the hypothesis of interest. Lix and Keselman present illustrations of how to obtain numerical results with their SAS/IML program.

3.4. The empirical Bayes approach

Boik (1997) introduced an empirical Bayes (EB) approach to the analysis of repeated measurements. This is a hybrid approach in that it represents a melding of the adjusted-df univariate and multivariate procedures. As he notes, the varied approaches to the analysis of repeated measurements differ according to how they model the variances and covariances

[2] For a 3×4 between- by within-subjects design, a main effect contrast vector among the levels of the RM variable could look like [1 −1 0 0 1 0 −1 0 1 0 0 −1]. Though this example contains simple (pairwise) contrasts (coefficients), the vector can be any set of linearly independent contrasts.

among the levels of the RM variable. For example, as we indicated, the conventional univariate approach assumes that there is a spherical structure among the elements of the covariance matrix, whereas the multivariate approach does not require that the covariance matrix assume any particular structure, only that it be positive definite. As we have pointed out, even though users are not typically interested in the structure of the covariance matrix, the covariance model that one adopts affects how well the fixed-effect parameters of the model (e.g., the treatment effects) are estimated. An increase in the precision of the covariance estimator translates into an increase in the sensitivity that the procedure has for detecting treatment effects. As an illustration, consider the multivariate approach to the analysis of repeated measurements. Because it does not put any restrictions on the form of the covariance matrix, it can be inefficient in that many unknown parameters must be estimated (i.e., all of the variances and all of the covariances among the levels of the RM variable), and this inefficiency may mean loss of statistical power to detect treatment effects. Thus, choosing a parsimonious model should be important to applied researchers.

The EB approach is an alternative to the univariate adjusted-df approach to the analysis of repeated measurements. The adjusted-df approach presumes that a spherical model is a reasonable approximation to the unknown covariance structure, and though departures from sphericity are expected, they would not be large enough to abandon the univariate estimator of the covariance matrix. The multivariate approach allows greater flexibility in that the elements of the covariance matrix are not required to follow any particular pattern. In the EB approach the unknown covariance matrix is estimated as a linear combination of the univariate and multivariate estimators. Boik (1997) believed that a combined estimator would be better than either one individually. In effect, Boik's (1997) approach is based on a hierarchical model in which sphericity is satisfied on average, though not necessarily satisfied on any particular experimental outcome. This form of sphericity is referred to as second-stage sphericity (Boik, 1997).

Boik (1997) demonstrated, through Monte Carlo methods, that the EB approach controls its Type I error rate and can be more powerful than either the adjusted-df or multivariate procedure for many non-null mean configurations. Researchers can make inferences about the RM effects by computing hypothesis and error sums of squares and cross product matrices with Boik's formulae and obtain numerical solutions with any of the conventional multivariate statistics (see Boik, 1997, p. 162 for an illustration).

4. Discussion

The aim of this paper was to indicate that 'repeated measures ANOVA' can refer to a number of different types of analysis for RM designs. Specifically, we indicated that repeated measures ANOVA could be construed to mean the conventional tests of significance, the adjusted-df univariate test statistics, a multivariate analysis, a multivariate analysis that does not require the assumptions associated with the usual multivariate test, or a combined univariate–multivariate test. In addition, by indicating the strengths and weaknesses of each of these approaches, we intended to convey the validity or lack thereof that can be associated with the *p*-values corresponding with each of these approaches. Thus, researchers can better convey the validity of their findings by indicating the type of 'repeated measures ANOVA' that was used to assess treatment effects. We summarize the advantages/disadvantages of the various approaches, indicating as well how numerical results can be obtained, in Table 1.

In conclusion, we feel it is rarely legitimate to use the conventional tests of significance since data are not likely to conform to the very strict assumptions associated with this procedure. On the other hand, researchers should take comfort in the fact that there are many viable alternatives to the conventional tests of significance. Furthermore, we believe that we can offer simple guidelines for choosing between them, guidelines which, by and large, are based on whether group sizes are equal or not.[3] That is, for simple RM designs containing no between-subjects variables or for between- by within-subjects designs having groups of equal size, we recommend either the empirical Bayes or the mixed model approach. Boik (1997) demonstrated that his approach will typically provide more powerful tests of RM effects than uniformly adopting either an adjusted-df univariate approach or a multivariate test statistic. Furthermore, numerical results can easily be obtained with a standard multivariate program. The mixed model approach is also likely to provide more powerful tests of RM effects than the adjusted-df univariate and multivariate approaches because researchers can model the covariance structure of their data. Furthermore, for designs that contain between-subjects grouping variables, heterogeneity across the levels of the grouping variable can also be modelled. To the extent that the actual covariance structure of the data resembles the fit structure, it is likely that the mixed model approach will provide more powerful tests than the empirical Bayes approach; however, this observation has not yet been confirmed through empirical investigation. A caveat to this recommendation is that when covariance matrices are suspected to be unequal, a safer course of action, in terms of Type I error protection, is to use an adjusted-df univariate test.[4] That is, some findings suggest that the EB and mixed model approaches may result in inflated rates of Type I error when covariance matrices are unequal and sample sizes are small, even when group sizes are equal (see Keselman *et al.*, 1999*a*, 1999*b*; Keselman, Kowalchuk & Boik, 2000; Wright & Wolfinger, 1996).

In those (fairly typical) cases where the group sizes are unequal and one does not know that the group covariance matrices are equal, researchers should use either the IGA or Welch–James tests. We feel quite comfortable in recommending the WJ and IGA tests as general analytic tools for the analysis of repeated measurements. We believe they are preferable to the conventional univariate (including adjusted-df univariate tests) and multivariate methods because they will typically control rates of Type I error where the conventional methods of analysis (and newer ones as well) will not. Furthermore, results indicate that the power to detect effects will not be substantially reduced when using WJ or IGA when the assumptions on the conventional procedures are satisfied (see Algina & Keselman, 1998). Thus, there is nothing to lose (with respect to power) and everything to

[3] We caution readers that our recommendations are no substitute for carefully examining the characteristics of their data and basing their choice of a test statistic on this examination. There are a myriad of factors (scale of measurement, distributional shape, outliers, etc.), not considered for the sake of simplicity in formulating our recommendations, which could result in other data analysis choices (nonparametric analyses, analyses based on robust estimators rather than least-squares estimators, transformations of the data, etc.). Furthermore, the empirical literature that has been published regarding the efficacy of the new procedures reviewed in this paper is extremely limited, and future findings may accordingly result in better recommendations.

[4] It is unknown to what extent covariance matrices are unequal between groups in RM designs since researchers do not report their sample covariance matrices. However, we agree with other researchers who investigate the operating characteristics of statistical procedures that the data in psychological experiments is likely to be heterogeneous (see DeShon & Alexander, 1996; Wilcox, 1987). Accordingly, the safest course of action, when group sizes are unequal, is to adopt a procedure that allows for heterogeneity. The empirical literature also indicates that one will not suffer substantial power losses by using a heterogeneous test statistic when heterogeneity does not exist (see Algina & Keselman, 1998; Keselman *et al.*, 1999*b*).

Table 1. Data analysis procedures for repeated measures designs

METHOD	REQUIREMENTS/CONSIDERATIONS/ ISSUES	EMPIRICAL FINDINGS	OBTAINING NUMERICAL RESULTS
Conventional *F* tests	• Require, among other assumptions, the unlikely to be satisfied assumption of multisample sphericity	• Inflated or depressed Type I error rates when data do not conform to multisample sphericity	• The major statistical packages (e.g., SAS, SPSS) compute these tests
Multivariate *F* tests	• Require, among other assumptions, homogeneity of the between-subjects covariance matrices	• Generally robust to heterogeneity of the covariance matrices when group sizes are equal • Not robust to covariance heterogeneity when group sizes are unequal • Multivariate test of the interaction effect may be non-robust to non-normality	• The major statistical packages compute these tests
Adjusted-df univariate test statistics: Greenhouse & Geisser (1959), Huynh & Feldt (1976)	• Require, among other assumptions, homogeneity of the between-subjects covariance matrices	• Inflated or depressed Type I error rates when covariance matrices are unequal, particularly when group sizes are unequal	• The major statistical packages compute these tests
Combined approach (Barcikowski & Robey, 1984)	• Uses both the adjusted-df univariate and multivariate tests to analyse effects, dividing the level of significance between the two tests	• Inflated or depressed Type I error rates when data are heterogeneous and non-normal	• The major statistical packages can be used to compute these tests
Huynh's (1978) IGA *F* tests	• Derived to be applicable to data that do not conform to multisample sphericity	• Robust to violations of multisample sphericity • Robust even when group sizes are unequal and relatively small • Robust to non-normality when robust estimators (i.e., trimmed means and Winsorized variances and covariances) are substituted for the least-squares estimators	• A SAS/IML program can be obtained from Algina (1997) • Applicable to any RM design that does not contain covariates or continuous variables

Mixed model *F* tests	• Allow the covariance structure of data to be modelled before conducting tests of the RM effects • Allow missing data across the levels of the RM variable • Allow between-subjects and/or within-subjects heterogeneity • Multiple comparisons of RM effects can be obtained through this procedure	• The default *F* tests are prone to distorted Type I error rates when covariance matrices are heterogeneous, group sizes are unequal and data are non-normal in form • The Satterthwaite optional *F* tests provide reasonably good protection against Type I errors	• Results can be obtained from PROC MIXED (SAS Institute, 1999)
Welch–James adjusted-df multivariate *F* tests (Keselman *et al.*, 1993)	• Sample sizes must conform to prescriptions given by Keselman *et al.* (1993) and Algina & Keselman (1997)	• Generally robust to covariance heterogeneity and non-normality if sample size requirements are met • Interaction test requires larger sample sizes in order to be robust to non-normality • Robust results can be achieved with reasonably moderate sample sizes when robust estimators are substituted for the least-squares estimators • Algina & Keselman (1997) found that the WJ test can have substantially more power to detect effects than the IGA approach • Multiple comparisons of RM effects can be obtained	• Lix & Keselman (1995) provide an SAS/IML program that can be used to obtain numerical results in any RM design not containing covariates or continuous variables • Keselman *et al.* (2001) provide an SAS/IML program that computes tests of significance with least-squares and/or robust estimators with or without boot strapping
Empirical Bayes approach (Boik, 1997)	• A hybrid approach that combines the adjusted-df univariate and multivariate approaches to the analysis of RM	• EB can be more powerful than either the adjusted-df or multivariate approach (Boik, 1997) • Generally robust to covariance heterogeneity when group sizes are equal • Type I error rates can be inflated or depressed when covariance matrices are heterogeneous when group sizes are unequal	• Hypothesis and error sums of squares and cross product matrices can be computed with formulae provided by Boik (1997) and then can be input to any multivariate test statistic

gain (with respect to Type I error control) by adopting one of these two approaches to the analysis of repeated measurements. Of the two, we generally recommend the WJ approach; based upon power analyses, it appears that it can have substantial power advantages over the IGA test (Algina & Keselman, 1997).

The SAS/IML program (SAS Institute, 1999) presented by Lix & Keselman (1995) can be used to obtain numerical results. However, according to results provided by Keselman *et al.* (1993) and Algina & Keselman (1998), sample sizes cannot be small. When sample sizes are unequal and small, we recommend the IGA test.

When researchers feel that they are dealing with populations that are non-normal in form—Tukey (1960) suggests that most populations are skewed and/or contain outliers—and thus subscribe to the position that inferences pertaining to robust parameters are more valid than inferences pertaining to the usual least-squares parameters, then either the IGA or WJ procedure, based on robust estimators, can be adopted. Results provided by Keselman, Algina, Wilcox & Kowalchuk (2000) certainly suggest that these procedures will provide valid tests of the RM main and interaction effect hypotheses (of trimmed population means) when data are non-normal, non-spherical and heterogeneous. Numerical results can be obtained with the SAS/IML program provided by Keselman *et al.* (2001).

5. Postscripts

5.1. Missing data

We remind the reader that in some areas of psychological research data may be missing over time. Mixed model analyses can provide numerical solutions based on all of the available data, as opposed to statistical software that derives results from complete cases (e.g., PROC GLM in SAS). Alternatively, multiple imputation (Rubin, 1987; Schafer, 1997) can be used in conjunction with software that calculates results from complete cases. However, the validity of these approaches depends on the mechanism that causes the data to be missing. Rubin (1976), Little & Rubin (1987) and Wang-Clow, Lange, Laird & Ware (1995) describe three mechanisms that can cause data to be missing. Citing Diggle & Kenward (1994), Little (1995) describes a fourth. Aspects of the following presentation assume that if a subject does not contribute data on a particular occasion, he or she does not contribute data subsequently. We refer to this as dropout.

Missing completely at random (MCAR). This process assumes that missing data occur at random and that missingness does not depend on individual characteristics or treatment. Thus dropout rates do not vary across treatment levels and dropout is not predictable from any of the variables in the study. Clearly MCAR is a very strong assumption. If data are MCAR, analysis of complete cases is not biased but is inefficient because data are discarded for respondents who have been observed on at least some of the measurement occasions. The maximum likelihood analysis implemented in PROC MIXED and multiple imputation are more efficient.

Covariate-dependent dropout (CDD). Here dropping out is dependent on between-subjects and within-subjects covariates that are fixed in the study. These covariates include the treatments. An example of CDD would be if subjects dropped out of a diet intervention study because they were unwilling to adhere to the diet regimen and not because of their weight

gain or loss (Wang-Clow *et al.*, 1995). Complete case analyses are unbiased but inefficient under CDD (Little, 1995). Correct analyses can be obtained by using the maximum likelihood analysis implemented in PROC MIXED or by using multiple imputation.

Missing at random (MAR). When data are missing at random, missingness depends on the observed values of the dependent variables and on the covariates. For example, discussing dropout, Wang-Clow *et al.* (1995, p. 295) report that data are MAR if 'attrition occurs at random, but with a probability that depends on an individual's previously observed response'. An example, of the MAR process would be if, in a study designed to assess the effectiveness of a drug in reducing weight among obese patients, subjects dropped out because they attained their desired weight loss. When data are MAR, correct analyses can be obtained by using maximum likelihood analysis implemented in PROC MIXED or by using multiple imputation.

Non-ignorable missingness (NI). Non-ignorable dropout means that the missing-data mechanism is not ignorable and must be explicitly taken into account in the data analysis. Two varieties of non-ignorable dropout have been described in the literature (Little, 1995). NI outcome-based dropout means that dropping out is predictable from the unrecorded scores on the dependent variable. Wang-Clow *et al.* (1995, p. 294) give as an example of this mechanism a study designed to control blood pressure where patients with home blood-pressure kits decide not to return to the study based on their own home measurements. NI random-effect dropout means that dropping out is predictable from the random effects in equation (10). For example, if subjects' performances over time are modelled as linear functions over time, and if subjects who have large slopes (i.e., are changing more quickly) are more likely to dropout, then, the missing-data mechanism is NI random-effect-dependent. When the missing-data mechanism is NI, two modelling approaches may be used (Little, 1995). In selection models, the missing-data mechanism is explicitly modelled along with modelling the dependent variable. In pattern-mixture models, data are stratified by missing-data patterns. Little (1995) has advocated the use of patter-mixture models for taking account of NI dropout. Drawing on the extensive work of Little (1993, 1994, 1995), Hedeker and Gibbons (1997) provide a recent presentation of patter-mixture models in the context of repeated measures analysis.

5.2. Multiple comparisons

The reader should note that the mixed model, WJ and EB approaches can also be applied to tests of contrasts (see SAS Institute, 1992, Chapter 16; Lix & Keselman, 1995, 1996; Boik, 1997). Our preference is for the approach presented by Keselman, Keselman & Shaffer (1991)—in effect a WJ approach—which can now be implemented with PROC MIXED. Accordingly, we present a brief description of the Keselman *et al.* approach (a detailed presentation can be found in Kowalchuk & Keselman, 2000).

Keselman *et al.* (1991) presented a statistic, which when combined with various multiple comparison critical values—Hochberg's (1988) sequentially acceptive step-up Bonferroni approach or Shaffer's (1986) sequentially rejective step-down Bonferroni approach—is robust to the effects of covariance heterogeneity and non-normality in unbalanced non-spherical RM designs. Their statistic, like the omnibus multivariate statistic, allows for a

general UN covariance structure and uses Satterthwaite's (1946) solution for error df. Numerical results can be obtained with the SAS/IML program provided by Lix & Keselman (1995). However, numerical results can also be obtained with the SAS PROC MIXED program. In particular, PROC MIXED allows users to compute linear contrast tests after selecting an 'appropriate' covariance structure. If a user selects a heterogeneous (across groups) UN covariance structure with the optional Satterthwaite df solution, he/she is in fact computing the statistic first defined by Keselman *et al.* Thus, their robust statistic can be obtained from a widely available statistical package.

Furthermore, Kowalchuk & Keselman (2000) found that, unlike the results reported for tests of omnibus effects, tests of contrasts based on always fitting the heterogeneous unstructured covariance structure were typically robust to conditions of non-sphericity, covariance heterogeneity and non-normality in unbalanced designs; moreover, they were typically just as powerful as tests based on knowing and fitting the true covariance structure for the data. That is, unlike the omnibus tests (SAS's default and Satterthwaite *F* tests), tests of contrasts can routinely be computed with a heterogeneous UN covariance structure. Kowalchuk and Keselman hypothesized that because a restricted number of levels of the RM variable (e.g., only two levels for pairwise tests) are involved in computing the standard error of the contrast, it is not that crucial to know the correct form of the covariance structure for tests of contrasts (at least for the structures that they investigated—UN, AR1, RC). Therefore, based on the results provided by Keselman *et al.* (1991) and Kowalchuk & Keselman (2000), we recommend the procedure presented by Keselman *et al.* for computing tests of contrasts of RM effects. We also note that the literature indicates that when data have been obtained from skewed long-tailed distributions, power and Type I error rates for Welch type tests (Welch, 1938) can be affected. In such situations, researchers can substitute robust estimators (i.e., trimmed means and Winsorized variances and covariances) for the least-squares estimators in the Keselman *et al.* statistic (see Keselman, Lix & Kowalchuk, 1998; Keselman *et al.*, 2001).

Acknowledgements

Work on this paper was supported by a grant from the Social Sciences and Humanities Research Council of Canada (no. 410-95-0006).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions 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., & Coombs, W. T. (1996). A review of selected parametric solutions to the Behrens-Fisher problem. In B. Thompson (Ed.), *Advances in social science methodology*, *Vol. 4* (pp. 137–171) Greenwich, CT: JAI Press.

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, I. Effects of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, *25*, 290–302.

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.

Danford, M. B., Hughes, H. M., & McNee, R. C. (1960). On the analysis of repeated-measurements experiments. *Biometrics*, *16*, 547–565.

DeShon, R. P., & Alexander, R. A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal. *Psychological Methods*, *1*, 261–277.

Diggle, P., & Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, *43*, 49–94.

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

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*, 64–78.

Hertzog, C., & Rovine, M. (1985). Repeated-measures analysis of variance in developmental research: Selected issues. *Child Development*, *56*, 787–809.

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

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, *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*, Vol. 2 (pp. 23–41). University of California Press, Berkeley.

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

Huynh, H., & 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., & 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.

Imhof, J. P. (1962). Testing the hypothesis of no fixed main-effects in Scheffé's mixed model. *Annals of Mathematical Statistics*, *33*, 1085–1095.

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

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., & 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. (1999*a*). 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. (1999*b*). 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., Wilcox, R. R., & Kowalchuk, R. K. (2001). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch–James test again. *Educational and Psychological Measurement*, *60*, 925–938.

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

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 violations of multisample sphericity. *Psychological Bulletin*, *110*, 162–170.

Keselman, H. J., Kowalchuk, R. K., & Boik, R. J. (2000). An examination of the robustness of the empirical Bayes and other approaches for testing main and interaction effects in repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, *53*, 51–67.

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., Wilcox, R. R., & Lix, L. M. (2001). A robust approach to hypothesis testing. Paper presented at the annual meeting of the Western Psychological Association, Maui, HI.

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.

Kogan, L. S. (1948). Analysis of variance: Repeated measurements. *Psychological Bulletin*, *45*, 131–143.

Kowalchuk, R. K., & Keselman, H. J. (2000). Mixed model pairwise multiple comparison procedures of repeated measures means. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Lawley, D. N. (1938). A generalization of Fisher's *z* test. *Biometrika*, *30*, 180–187, 467–469.

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

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

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*, 125–134.

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.
Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
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.
Lix, L. M., & Keselman, H. J. (1996). Interaction contrasts in repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, *49*, 147–162.
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.
McCall, R. B., & Appelbaum, M. I. (1973). Bias in the analysis of repeated-measures designs: Some alternative approaches. *Child Development*, *44*, 401–415.
Mendoza, J. L. (1980). A significance test for multisample sphericity. *Psychometrika*, *45*, 495–498.
Noe, M. J. (1976). A Monte Carlo survey of several test procedures in the repeated measures design. Paper presented at the meeting of the American Educational Research Association, April, San Francisco.
Norušis, M. J. (1993). *SPSS for Windows, Advanced Statistics, Release 6.0*. Chicago: SPSS Inc.
Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, *69*, 894–908.
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.
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 its use in multivariate analysis. *Annals of Mathematical Statistics*, *24*, 220–238.
Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
SAS Institute (1992). *SAS Technical Report: SAS/STAT Software Change and Enhancements Release 6.07*. Author, Cary, NC.
SAS Institute (1995). *Introduction to the MIXED procedure*. Cary, NC: SAS.
SAS Institute (1996). *SAS/STAT software: Changes and enhancements through Release 6.11*. Cary, NC: SAS.
SAS Institute (1999). *SAS/STAT user's guide*, *Version 7*. Cary, NC: SAS.
SAS Institute (1999). *SAS/IML: User's guide*, *Version 8*. Author Cary, NC: SAS.
Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110–114.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.
Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, *81*, 826–831.
Stoloff, P. H. (1970). Correcting for heterogeneity of covariance for repeated measures designs of the analysis of variance. *Educational and Psychological Measurement*, *30*, 909–924.
Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In Olkin *et al.* (Eds.), *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.
Wang-Clow, F., Lange, M., Laird, N. M., & Ware, J. H. (1995). A simulation study of estimators for rates of change in longitudinal studies with attrition. *Statistics in Medicine*, *14*, 283–297.
Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, *38*, 330–336.

Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology*, *38*, 29–60.
Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, *24*, 471–494.
Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.
Wolfinger, R. D. (1996). Heterogeneous variance–covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, *1*, 205–230.
Wright, S. P., & Wolfinger, R. D. (1996). Repeated measures analysis using mixed models: Some simulation results. Paper presented at the Conference on Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions, Nantucket, MA.

Received 25 October 1999; revised version received 28 April 2000