Testing Treatment Effects in Repeated Measures Designs: An Update for Psychophysiological Researchers

by

H.J. Keselman
Department of Psychology
University of Manitoba

Work on this paper was supported by a grant from the Social Sciences and Humanities Research Council of Canada (# 410-95-0006). Requests for reprints should be sent to H.J. Keselman, Department of Psychology, The University of Manitoba, Winnipeg, Manitoba CANADA R3T 2N2

Abstract

In 1987, Jennings enumerated data analysis procedures that authors must follow for analyzing effects in repeated measures designs when submitting papers to *Psychophysiology*. These prescriptions were intended to counteract the effects of nonspherical data, a condition know to produce biased tests of significance. Since this editorial policy was established, additional refinements to the analysis of these designs have appeared in print in a number of sources that are not likely to be routinely read by psychophysiological researchers. Accordingly, this paper presents additional procedures that can be used to analyze repeated measurements not previously enumerated in the editorial policy. Furthermore, the paper indicates how numerical solutions can easily be obtained.

Descriptors: Repeated Measurements, main, interaction, and contrast tests, new analyses

Testing Treatment Effects in Repeated Measures Designs: An Update for Psychophysiological Researchers

In 1987, *Psychophysiology* stipulated a policy for analyzing data from repeated measures designs (Jennings, 1987). In particular, the assumptions that are required to obtain valid tests of omnibus and sub-effect hypotheses were enumerated and prescriptions for analyzing effects in such designs were stipulated. Recently, there have been additional refinements to the analysis of these designs which have appeared in print in a number of sources that are not likely to be routinely read by psychophysiological researchers. Accordingly, the purpose of this paper is to update prior recommendations.

It is important for the reader to note that the recommendations presented in this paper are based on the findings of empirical investigations which compared various alternative strategies of data analysis. The studies, individually and collectively, did not, nor could they, exhaust all conceivable parametric conditions that psychophysiological researchers may encounter in their research endeavors. Consequently, my recommendations will not always result in the optimal method of analysis; however, they should, more often than not, result in the 'best' method of analysis. Analyses based on a thorough familiarity with the phenomenon under investigation, including the mechanism(s) that is(are) responsible for generating the data will always prove superior to those based on general recommendations. The caveat of “know thy data" certainly applies to the analysis of repeated measures designs and should always be paramount when considering the appropriateness of the recommendations.

The one Between- by one Within-Subjects design

Assessing Main Effects

Researchers working in the psychophysiological area frequently adopt a repeated measures design that contains both Between-Subjects grouping variables and Within-Subjects repeated measures variables (see, for example, articles published in the January 1997, Volume 34, Number 1 issue of *Psychophysiology*). The simplest of these designs involves a single Between-Subjects grouping factor and a single Within-Subjects repeated measures factor, in which subjects ($i = 1, \ldots, n_g, \Sigma n_g = N$) are selected randomly for each level of the Between-Subjects factor ($g = 1, \ldots, G$) and observed and measured under all levels of the Within-Subjects factor ($m = 1, \ldots, M$).

To set the stage for the procedures that I will present for analyzing such designs and to help clarify notation, consider the following hypothetical research problem. Specifically, I will use the data presented in Table 1 which could represent the outcome of an experiment in which the Between-Subjects variable is susceptibility to stressors ($g = 1, \ldots, 3$) and the Within-Subjects variable is a task to be performed at four levels of challenge ($m = 1, \ldots, 4$). Readers should note that these data were obtained from a random number generator and, therefore, are not intended to reflect actual characteristics of the previously posed hypothetical problem. However, they were generated to reflect characteristics (i.e., covariance structure, relationship of covariance structure to group sample sizes, the distributional shape of the data, etc.) of repeated measures data that are likely to be obtained in psychophysiological investigations.[1] That is, these data are based on the assumption that I as well as others working in the field make (see, for example, Keselman & Keselman, 1988; 1993; Jennings, 1987; Overall & Doyle, 1994; Vassey & Thayer, 1987), namely, that psychophysiological data will not, in all likelihood, conform to the validity assumptions of the traditional tests of repeated measures effects.

Insert Table 1 About Here

In each of the groups, there are 13 observations (i.e., $n_1 = n_2 = n_3 = 13$; $\Sigma n_g = 39$). The computational procedures that will be illustrated when group sizes are unequal will be based on the data associated with subject numbers that are not enclosed in parentheses; thus, for these analyses $n_1 = 7$, $n_2 = 10$, and $n_3 = 13$ ($\Sigma n_g = 30$). Cell and marginal (unweighted) means for each data set (balanced and unbalanced) are contained in Table 2. In the illustrations that follow, the results for all analysis procedures will be presented and discussed; naturally, researchers would only compute those procedures that actually relate to their research hypotheses.

Insert Table 2 About Here

The Univariate Approach

Tests of the Within-Subjects main and interaction effects traditionally have been accomplished by the respective use of the univariate analysis of variance (ANOVA) F statistics

$$F_M = MS_M / MS_{M\times S/G} \sim F[\alpha; (M-1), (N-G)(M-1)] \text{ and} \quad (1)$$

$$F_{G\times M} = MS_{G\times M} / MS_{M\times S/G} \sim F[\alpha; (G-1)(M-1), (N-G)(M-1)], \quad (2)$$

where $\sim$ should 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, Keselman & Mendoza, 1979; Rouanet & Lepine, 1970). The sphericity assumption is satisfied if and only if the $M - 1$ contrasts (orthonormalized) among the repeated measures variable 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 contrasts are the same for all levels of this grouping factor. Jointly, these two assumptions have been referred to as multisample sphericity (Mendoza, 1980).

When the assumptions to the traditional tests have been satisfied, they will provide a valid test of their respective null hypotheses and will be uniformly most powerful for detecting treatment effects when they are present. These traditional tests are easily obtained with the major statistical packages, that is with BMDP (1994), SAS (1990), and SPSS (Norusis, 1993). Thus, when assumptions are known to be satisfied psychophysiological researchers can adopt the traditional 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. For the balanced (i.e., $N = 39$) data set given in Table 1, PROC GLM (SAS, 1990) results are $F_M = 3.95$ (3, 108; $p = .0103$) and $F_{G\times M} = 5.17$ (6, 108; $p = .0001$).

Unfortunately, as Jennings (1987) and others have indicated, the data from most applied work are unlikely to conform to the strict requirements of multisample sphericity (Keselman & Keselman, 1988; 1993; Overall & Doyle, 1994; Vassey & Thayer, 1987). The result of applying the traditional tests of significance with data that do not conform to the assumptions of multisample sphericity is that too many null hypotheses are falsely rejected (Box, 1954; Collier, Baker, Mandeville & Hayes, 1967; Kogan, 1948). Furthermore, as the degree of nonsphericity increases, the traditional repeated measures F tests becomes increasingly liberal (Collier et al.).

However, when the design is balanced (group sizes are equal) the Greenhouse and Geisser (1959) and Huynh and Feldt (1976) tests are robust alternatives to the traditional tests. As Keselman and Rogan (1980) have indicated, the Greenhouse and Geisser or Huynh and Feldt methods adjust the degrees of freedom of the usual F statistics; the adjustment for each approach is based on a sample estimate, $\hat{\epsilon}$ and $\tilde{\epsilon}$, respectively, of the unknown sphericity parameter (ϵ).

The empirical literature indicates, however, that these adjusted degrees of freedom tests are not robust when the design is unbalanced (Keselman & Keselman,

1990; Keselman, Keselman & Lix, 1995). Specifically, the tests will be liberal (conservative) when group sizes and covariance matrices are negatively (positively) paired with one another. A positive (negative) pairing refers to the case in which the smallest (largest) n_g is associated with the covariance matrix with the smallest (largest) element values.[2]

The major statistical packages [BMDP, SAS, SPSS] provide Greenhouse and Geisser (1959) and Huynh and Feldt (1976) adjusted p-values. For the balanced (i.e., $N = 39$) data set given in Table 1, PROC GLM (SAS, 1990) results for the Greenhouse and Geisser tests are $F_M = 3.95$ [3(.4497) = 1.3491, 108(.4497) = 48.5676; p = .0409] and $F_{G\times M} = 5.17$ [6(.4497) = 2.6982, 108(.4497) = 48.5676; p = .0046], where $\hat{\epsilon}$ = .4497. The corresponding Huynh and Feldt results are $F_M = 3.95$ [3(.4869) = 1.4607, 108(.4869) = 52.5852; p = .0372] and $F_{G\times M} = 5.17$ [6(.4869) = 2.9214, 108(.4869) = 52.5852; p = .0036], where $\tilde{\epsilon}$ = .4869.

The Multivariate Approach

The multivariate test of the repeated measures main effect is performed by first creating $M - 1$ difference (D) variables.[3] The null hypothesis that is tested, using Hotelling's (1931) T^2 statistic, is that the vector of population means of these $M - 1$ D variables equals the null vector. The upper $100(1 - \alpha)$ percentage points of the T^2 distribution can be obtained from the relationship

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

The multivariate test of the Within-Subjects interaction effect, on the other hand, is a test of whether the vectors of population means of the $M - 1$ D 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 $M - 1$ D variables are the

dependent variables and the grouping variable (G) is the Between-Subjects independent variable. When $G > 2$, four popular multivariate criteria are: (1) Wilk's (1932) likelihood ratio, (2) the Pillai (1955)-Bartlett (1939) trace statistic, (3) Roy's (1953) largest root criterion, and (4) the Hotelling (1951)-Lawley (1938) trace criterion. Based on Olson's work (1974), I recommend the Pillai-Bartlett criterion since it seems most robust to assumption violations. When $G = 2$, all criteria are equivalent to Hotelling's T^2 statistic.

Valid multivariate tests of the repeated measures hypotheses, 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. The empirical results indicate that the multivariate tests of the repeated measures main and interaction effects are generally robust to their assumption violations when the design is balanced and not robust when the design is unbalanced (Keselman et al., 1995). Furthermore, under most conditions that researchers are likely to encounter with real data (sample sizes, magnitude of treatment effects), a multivariate test will be more sensitive to the presence of treatment effects than the univariate tests (Algina & Keselman, 1997; Davidson, 1972). Also, as indicated, multivariate tests require fewer assumptions. Consequently, when the design is balanced I recommend that researchers adopt multivariate procedures to assess the effects of their treatments rather than the adjusted degrees of freedom tests recommended by Keselman and Rogan (1980). Multivariate tests of repeated measures designs hypotheses are easily obtained from the multivariate or repeated measures program associated with any of the three major statistical packages. PROC GLM (SAS, 1990) results for the balanced data set are $F_M = 6.00$ (3, 34; $p = .0021$) and F (Pillai's Trace)$_{G\times M} = 3.49$ (6, 70; $p = .0044$).

Additionally, researchers should note that they can estimate how many observations they need in their studies to detect repeated measures effects with either the multivariate or adjusted degrees of freedom approach (see Algina & Keselman, 1997;

Muller & Barton, 1989, 1991; O'Brien & Muller, 1993, pp. 325-333). Power analysis modules can be obtained from the Internet network through file transfer protocols (see O'Brien & Muller, 1993, p.340).

Nonpooled Corrected Degrees of Freedom Statistics

Since the effects of heterogeneous covariances on tests of mean equality in unbalanced repeated measures designs are similar to the effects of variance heterogeneity on such tests in independent groups designs, one solution to analysis problems in heterogeneous unbalanced designs parallels that found in the context of completely randomized designs.

The Keselman, Carriere and Lix (1995) Statistic. The Keselman, Carriere and Lix (1995) statistic (WJ), based on the work of Johansen (1980), is a multivariate extension of the (Welch, 1947, 1951)-James (James, 1951, 1954) procedures for completely randomized designs. The statistic does not pool across heterogeneous sources of variation (covariance matrices) and estimates error degrees of freedom from the data. Though the test statistic cannot always be obtained from the major statistical packages, Lix and Keselman (1995) present a SAS (1989) IML program that can be used to compute the Welch-James test for *any* repeated measures design. The program requires only that the user 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.

The empirical literature indicates that the Welch-James test is generally insensitive to heterogeneity of the covariance matrices and, accordingly, will provide a valid test of repeated measures hypotheses (Keselman, Algina, Kowalchuk, & Wolfinger, 1997a; Keselman, Carriere & Lix, 1993). Specifically, researchers should consider using

this statistic when they suspect that group covariance matrices are unequal and they have groups of unequal size. It should be noted, however, that to obtain a robust statistic sample sizes must be relatively large. That is, according to Keselman et al. (1993) in order to obtain a robust test of the repeated measures main effect hypothesis using this statistic, the number of observations in the smallest of groups must be three to four times the number of repeated measurements minus one (i.e., $M - 1$); to obtain a robust test of the interaction, this number must be five or six to $(M - 1)$. Algina and Keselman (in press) determined that the sample size requirements enumerated by Keselman et al. (1993) generalize to larger repeated measures designs (i.e., 6×4 and 6×8 as opposed to the 3×4 and 3×8 designs investigated by Keselman et al., 1993) for the test of the main effect but that sample size requirements had to be larger in order to obtain a robust interaction test. Nonetheless, for most situations likely to be encountered with applied data (i.e., moderate degrees of nonnormality, covariance heterogeneity, and unbalancedness), these authors recommended that researchers continue to use the Welch-James test for examining repeated measures effects.

For the data set in which group sizes are unequal (i.e., $n_1 = 7$, $n_2 = 10$, and $n_3 = 13$), Welch-James results are $WJ_M = 9.53$ (3, 11.25; $p = .002$) and $WJ_{G\times M} = 8.26$ (6, 13.89; $p = .0006$).

The Huynh (1978)-Algina (1994)-Lecoutre (1991) Statistic. Huynh (1978) developed a test of the Within-Subjects main and interaction hypotheses, the Improved General Approximation test, that is designed to be used when multisample sphericity is violated. The Improved General Approximation tests of the Within-Subjects main and interaction hypotheses are the usual statistics, F_M and $F_{G\times M}$, 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 and Algina (1996). These parameters adjust the critical value to take into account the effect that violation of multisample sphericity has on F_M and $F_{G\times M}$. If multisample sphericity holds,

$$bF[\alpha; h', h] = F[\alpha; (M-1), (N-G)(M-1)] \text{ and}$$

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

A SAS/IML (SAS Institute, 1989) program is also available for computing this test in any repeated measures design (see Algina, in press). IGA results are $F_M = 7.44$ (1.2085, 14.8387; $p = .0470$), where $b = 1.6753$ and $F_{G\times M} = 3.04$ (1.3992, 14.4387; $p = .1723$), where $c = 1.4884$.

Keselman et al. (1997a) compared the Welch-James and Improved General Approximation tests and found that both were generally able to control their rates of Type I error even when assumptions (normality and covariance homogeneity) were jointly violated. The Welch-James test, however, required a larger sample size to achieve robustness. Based on their results and recommendations and results reported by Keselman et al. (1993) and Algina and Keselman (1997), I recommend the Welch-James test for analyzing effects in repeated measures designs. Typically it will not only provide a robust test of repeated measures effects, but, as well, will generally provide a more powerful test of nonnull effects, compared to the Improved General Approximation test. Indeed, Algina and Keselman (1997) found, when Type I error rates were controlled, power differences in favor of Welch-James as large as 60 percentage points! However, if one cannot meet the recommended sample size requirements for the valid use of the Welch-James test, then the Improved General Approximation test is recommended.

A General Method

Another procedure that psychophysiological researchers can adopt to test repeated measures effects can be derived from a general formulation for analyzing effects in repeated measures models. This newest approach to the analysis of repeated

measurements is a mixed model analysis. 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. Thus, one can use this procedure to select the most appropriate covariance structure prior to testing the usual repeated measures hypotheses, e.g., F_M and $F_{G\times M}$.

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, 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. The PROC MIXED program in SAS (1996) allows researchers to examine a number of different covariance structures that could possibly describe their particular data [e.g., compound symmetric (the structure assumed by many programs for valid univariate tests), unstructured (the structure assumed by many programs for valid multivariate tests), first-order autoregressive, etc.]. 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.

In order to select an appropriate structure for one's data, PROC MIXED users can use either an Akaike (1974) or Schwarz (1978) information criteria. Keselman, Algina, Kowalchuk, and Wolfinger (1997b) compared these criteria 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 and group size and covariance balance/imbalance. Their data indicated that neither criteria 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. Not surprisingly, Keselman et al. (1997a) found that the default F-tests that PROC MIXED computes based on either of these two criteria were prone to inflated rates of Type I error. Accordingly, any presumed power benefits must be discounted when the procedure is prone to excessive rates of Type I error.

Multiple comparison procedures

Contrast Tests On Between-Subjects Marginal Means

The choice of a test statistic for contrasts on the Between-Subjects marginal means rests on the tenability of the homogeneity of variance assumption. If this assumption is tenable, then a test statistic which uses a pooled estimate of error variance (i.e., $MS_{S/G}$) in estimating the standard error of a contrast is appropriate and will provide the most powerful test. On the other hand, if the homogeneity of variance assumption is untenable, then the more appropriate test statistic is one which allows for an individual estimate of the contrast variance (Welch, 1938). Given that one seldom knows whether the variance homogeneity assumption is tenable, the safest course of action is to uniformly adopt a test statistic that is based on the separate variance approach. Indeed, research has indicated that this strategy results in only slight losses in power when the homogeneity of variance assumption is satisfied (Best & Rayner, 1987; Games & Howell, 1976). Test statistics based on the separate variance approach, often referred to as nonpooled statistics, can be obtained from programs in the popular statistical packages (e.g., the BMDP 3D program, the SAS PROC TTEST, and the SPSS T-TEST and ONEWAY procedures).

Stepwise Multiple Comparison Procedures. A class of multiple comparison procedures that can be used to test pairwise contrast hypotheses are stepwise procedures. Unlike simultaneous multiple comparison procedures [e.g., Tukey's HSD; see Kirk, 1995, p. 144] which use a constant critical value to assess statistical significance, stepwise procedures involve a succession of testing stages in which the significance criterion, and

hence the critical value, is adjusted throughout the stages. Stepwise procedures are recommended since they usually (i.e., for many nonnull mean configurations) provide a more powerful test of the multiple comparison null hypothesis. For pairwise contrasts among means, researchers can adopt any one of a number of stepwise multiple comparison procedures.

Fisher's (1935) Two-Stage Least Significant Difference (LSD) approach. A popular approach for testing pairwise contrasts is due to Fisher (1935). In this approach, an omnibus test is conducted at stage one and, if declared nonsignificant, all pairwise contrast hypotheses are regarded as null. If, on the other hand, the omnibus test is declared significant at stage one, then all pairwise contrasts hypotheses are tested using t statistics, each assessed at an α level of significance. Hayter (1986) showed, however, that when $G > 3$ this two-stage procedure does not limit the overall, that is, familywise rate of Type I error to α. By modifying the critical value of the stage two tests of the pairwise differences (that is, by using $q_{\alpha;\, G\text{-}1,\, \nu}/\sqrt{2}$, where q_α is a value from the Studentized range distribution), however, Hayter (1986) showed that Fisher's (1935) two-stage approach provides exact Type I error control.

Shaffer's (1979, 1986) Sequentially Rejective Bonferroni Approaches. Another stepwise multiple comparison procedure that researchers can adopt is one due to Shaffer (1986). In this procedure, the p-values associated with the test statistics are rank-ordered from smallest to largest. That is, $p_1 \leq p_2 \leq , \ldots , \leq p_w$, where $w = G(G - 1)/2$ for pairwise contrasts. At step one, the smallest p-value, p_1, is compared to α/w. If $p_1 > \alpha/w$, statistical testing stops and all pairwise contrast hypotheses (H_i, $1 \leq i \leq w$) are retained; if $p_1 \leq \alpha/w$, however, H_1 is rejected and one proceeds to test the remaining hypotheses in a similar step-down fashion by comparing the associated p-values to α/w^*, where w^* equals the maximum number of true null hypotheses, given the number of hypotheses rejected at previous steps. For example, if $w = 6$ (i.e., $G = 4$) and H_1 (say $\mu_1 - \mu_2 = 0$) was rejected at step one, Shaffer's procedure would compare p_2 to $\alpha/3$ at

step two, since only three pairwise null hypotheses could be true (i.e., $\mu_1 = \mu_3$, $\mu_1 = \mu_4$, and $\mu_3 = \mu_4$, or $\mu_2 = \mu_3$, $\mu_2 = \mu_4$, and $\mu_3 = \mu_4$). Note that this procedure is more powerful than the usual Bonferroni procedure which would compare each p-value to $\alpha/6$. Appropriate denominators for each α-stage test can be found in Shaffer's (1986) Table 2, for designs containing up to ten treatment levels (see also Seaman, Levin & Serlin, 1991).

Shaffer (1979, 1986) proposed a modification to her sequentially rejective Bonferroni procedure which involves beginning this procedure with an omnibus test. If the omnibus test is declared nonsignificant, statistical testing stops and all pairwise differences are declared nonsignificant. On the other hand, if one rejects the omnibus null hypothesis one proceeds to test pairwise contrasts using the sequentially rejective Bonferroni procedure previously described with the exception that p_1, the smallest p-value, is compared to a significance level which reflects the information conveyed by the rejection of the omnibus null hypothesis. For example, for $w = 6$, rejection of the omnibus null hypothesis implies at least one inequality of means and therefore p_1 is compared to $\alpha/3$, rather than $\alpha/6$; the remaining stages (2, 3, etc., etc.) use the w^* values given by Shaffer (1986) or Seaman et al., (1991).

Hochberg's (1988) Step-up Bonferroni Approach. Like Shaffer's (1986) procedure, the p-values associated with the test statistics are rank ordered. In Hochberg's procedure, however, one begins by comparing the largest p-value, p_w, to α/w. If $p_w \leq \alpha$, *all* hypotheses are rejected. If $p_w > \alpha$, then H_w is retained and one proceeds to compare $p_{(w-1)}$ to $\alpha/2$. If $p_{(w-1)} \leq \alpha/2$, then all remaining Hs are rejected. If not, then $H_{(w-1)}$ is retained and one proceeds to compare $p_{(w-2)}$ with $\alpha/3$, and so on. Clearly, this stepwise procedure is likely to be more sensitive in detecting pairwise differences than the usual Bonferroni procedure since on every comparison, except the last, the level of significance is larger. It is important to note that researchers can adopt other stepwise multiple comparison critical values (see Keselman, 1993, 1994; Seaman, Levin & Serlin, 1991).

The illustrations in the remainder of the paper are based on the unbalanced data set ($n_1 = 7$, $n_2 = 10$, and $n_3 = 13$). Prior to illustrating tests on the Between-Subjects marginal means, I report the omnibus test results for the Between-Subjects effect.

The test of the Between-Subjects grouping variable is statistically significant, assuming the level of significance to be .05, with either the traditional test of significance ($F = 11.74$: 2, 27; $p = .0002$) or the robust Welch (1951) test ($WJ = 4.80$: 2, 11.42; $p = .0307$). I recommend that psychophysiological researchers adopt the Welch procedure since, as indicated, it is generally robust to variance heterogeneity and is relatively sensitive in detecting treatment effects, as compared to the ANOVA F test, even when homogeneity is satisfied (Best & Rayner, 1987).

Pairwise comparisons can be computed on the Between-Subjects marginal means either following a significant omnibus test or instead of the omnibus test if postulated *a priori.* For illustration purposes, I will adopt Fisher's (1935) two-stage LSD procedure. Since the omnibus Welch (1951) test was statistically significant, one would proceed to the pairwise tests. The pairwise tests are based on the two-sample Welch (1938) test. Results for the three tests are: (a) $WJ = 8.91$ (8.18; $p = .0170$) for G_1 vs. G_2, (b) $WJ = 10.08$ (6.52; $p = .0172$) for G_1 vs. G_3, and (c) $WJ = 0.00$ (13.20; $p = .9966$) for G_2 vs. G_3. Since $G = 3$, Fisher's (1935) LSD controls the familywise Type I error rate and therefore each of the t-tests can be assessed for significance with $\alpha = .05$. Accordingly, we conclude that G_1 vs. G_2 and G_1 vs. G_3 are statistically significant.

Contrast Tests On Within-Subjects Marginal Means

Multiple comparison procedures that use a constant, that is, a pooled estimate of error variance in obtaining the standard error of a contrast do not limit the familywise rate of Type I error to α when the data do not satisfy the multisample sphericity assumption (Keselman & Keselman, 1988; Keselman, Keselman & Shaffer, 1991; Maxwell, 1980). That is, as Keselman and Keselman (1988) noted, when the assumption of multisample sphericity is not satisfied, the use of various types of pooled estimates of

error variance in estimating the standard error of Within-Subjects contrasts results in biased tests of significance, particularly when the design is unbalanced. Fortunately, a test procedure is available which provides a robust test of pairwise contrasts of repeated measures means for unbalanced (as well as balanced) nonspherical data (Keselman et al., 1991). This procedure involves the use of a nonpooled statistic (i.e., Welch) and Satterthwaite's solution for degrees of freedom (Keselman et al., 1991; Satterthwaite, 1941, 1946). The program provided by Lix and Keselman (1995) can be used to obtain numerical results.

Simultaneous Multiple Comparison Procedures. Using the nonpooled statistic, Keselman et al. (1991) investigated the robustness of various simultaneous multiple comparison procedures in nonspherical unbalanced repeated measures designs. The results of their simulations indicated that the nonpooled statistic generally limited the familywise rate of Type I error to α when used in conjunction with a Bonferroni $\{t[\alpha/(2c); \nu_s]\}$, Studentized range $\{q[\alpha; M, \nu_s]/\sqrt{2}\}$ or a Studentized maximum modulus $\{M[\alpha; c, \nu_s]\}$ critical value, where $c = M(M-1)/2$, the total number of pairwise contrasts and ν_s is error degrees of freedom based on the Satterthwaite (1941, 1946) solution (Hochberg & Tamhane, 1987; Maxwell & Delaney, 1990). In general, Keselman et al. found that a Bonferroni critical value provided the best Type I error control, followed by a Studentized maximum modulus critical value and, finally, a Studentized range critical value.

Stepwise Multiple Comparison Procedures. With respect to tests of pairwise contrasts of repeated measures means, Keselman (1993, 1994) reported that several stepwise strategies can be used to limit the familywise rate of Type I error to α for repeated measures data that do not meet the multisample sphericity assumption. For a detailed discussion of stepwise multiple comparison procedures for repeated measures designs, the reader is referred to Keselman (1993, 1994). The stepwise procedures enumerated for tests of Between-Subjects marginal means can be used here as well.

For illustration purposes, I will assume the pairwise tests were postulated *a priori* and therefore will adopt Hochberg's step-up Bonferroni procedure to assess statistical significance. The six pairwise WJ values are (a) M_1 vs. M_2: WJ = .25 (1, 7.76; p = .6291), (b) M_1 vs. M_3: WJ = 1.83 (1, 7.65; p = .2151), (c) M_1 vs. M_4: WJ = 4.86 (1, 7.14; p = .0626), (d) M_2 vs. M_3: WJ = 14.66 (1, 25.34; p = .0008), (e) M_2 vs. M_4: WJ = 25.54 (1, 12.31; p = .0003), and (f) M_3 vs. M_4: WJ = 14.01 (1, 8.86; p = .0047). The largest p-value, .6291, is compared to .05 and thus this pairwise hypothesis (M_1 vs. M_2) cannot be rejected. Next, the second largest p-value, .2151, is compared to $\alpha/2 = .05/2 = .025$. Since this p-value is greater than its criterion of significance, this pairwise difference (M_1 vs M_3) also cannot be rejected. In the third step, .0626 is compared to $\alpha/3 = .0167$ and consequently M_1 vs M_4 is retained. In the fourth step .0047 is compared to $\alpha/4 = .0125$. Since this p-value is less than its criterion of significance, M_3 vs M_4 as well as the remaining comparisons (i.e., M_2 vs M_3 and M_2 vs M_4) are declared statistically significant.

Assessing the interaction effect

Traditionally, interaction effects have been assessed using one of two methods: (1) tests of simple effects and/or (2) interaction contrasts. The choice between these two methods has depended on the hypotheses of interest.

As a preface to the discussion, it is important to note that although researchers frequently compute simple effect tests following a significant interaction such tests do not probe the interaction hypothesis.[4] Cogent discussions of this point have been presented in the literature (see, for example, Betz & Gabriel, 1978; Boik, 1993; Lix & Keselman, 1996). The presentation of simple effect tests, therefore, is intended for those researchers who compute these tests, not as a means of probing interactions, but as means for examining differences between treatments at a fixed level of one variable, when such

comparisons have interpretive meaning within a particular research context (see, for example, Toothaker, 1991, pp. 119-122).

Simple Between-Subjects Effects

In the G × M repeated measures design, the simple effect of factor G refers to the effect of the Between-Subjects factor G at a particular or fixed level of the Within-Subjects factor M. By restricting our attention to a particular level of M, we have essentially eliminated the Within-Subjects factor from the design and are left with a single-factor Between-Subjects design.

A statistic that estimates error variance on the basis of the data at a fixed level of M and, accordingly, does not require multisample sphericity is

$$F_G \text{ at } M_m = \frac{MS_{G \text{ at } M_m}}{MS_{S/G \text{ at } M_m}} \sim F[\alpha; (G - 1), (N - G)]. \quad (6)$$

Essentially, this approach is equivalent to conducting a simple Between-Subjects analysis of the grouping factor G at a particular level of M and, therefore, is dependent on the assumptions of independence of observations, normality, and homogeneity of variances. If the variance homogeneity assumption is untenable, a heterogeneous variance procedure, such as the one by Welch (1951), should be used. In order to limit the familywise Type I error rate to α, each simple effect should be assessed at a reduced significance level. For a discussion of familywise control with simple effect testing, readers are referred to Kirk (1995) and Maxwell and Delaney (1990).

For our example there are four simple effect tests: G at M_1, G at M_2, G at M_3, and G at M_4. Each of the simple effect hypotheses is tested with a Welch (1951) omnibus test statistic. Thus, we find that: (a) $WJ = 2.56$ $(2, 11.36; p = .1212)$, (b) $WJ = 3.07$ $(2, 12.95; p = .0543)$, (c) $WJ = 10.27$ $(2, 12.20; p = .0024)$, and (d)

WJ = 38.18 (2, 10.25; p = .0000), respectively, for the preceding four Between-Subjects simple effect tests. Only G at M_3 and G at M_4 are statistically significant since their p-values are less than $\alpha/4 = .05/4 = .0125$ (Bonferroni procedure).

If relevant, each of the three simple effect tests can be followed-up with contrast tests, such as pairwise comparisons between the simple effect means. Again, Welch (1938) tests would be recommended in order to circumvent the homogeneity of variance assumption. To maintain consistency with each simple effect level of significance, these tests can be assessed at the .0125/3 = .004167 level.

Simple Within-Subjects Effects

In our G × M design, the simple effect of factor M refers to the effect of the Within-Subjects factor M at a particular level of the Between-Subjects factor G. By focusing our attention on a fixed level of G, the Between-Subjects factor is effectively eliminated and we are left with a single-factor Within-Subjects design. Accordingly, one can adopt univariate or multivariate approaches to the analysis of simple Within-Subjects effects.

Adopting a univariate approach, the simple effects of the Within-Subjects factor M can be tested with a statistic that corrects for nonsphericity, namely, an adjusted degrees of freedom approach due to Greenhouse and Geisser (1959) ($\widehat{\epsilon}$) or Huynh and Feldt (1976) ($\widetilde{\epsilon}$). Using the Greenhouse and Geisser adjustment, the test would be

$$F_{M \text{ at } G_g} = \frac{MS_{M \text{ at } Gg}}{MS_{M \times S/G_g}} \sim F[\alpha; (M-1)\widehat{\epsilon}, (n_g - 1)(M-1)\widehat{\epsilon}], \qquad (7)$$

where $\widehat{\epsilon}$ (or $\widetilde{\epsilon}$) is estimated on the basis of $\mathbf{S}_g$, the sample covariance matrix of group g.

The three Greenhouse and Geisser approximate degrees of freedom tests equal: (a) F = 0.85 (1.09, 6.58; p = .4007, $\widehat{\epsilon}$ = .3653), (b) F = 19.23 (1.74, 15.64; p = .0001,

$\widehat{\epsilon} = .5791$), and (c) F = .18 (1.92, 23.07; p = .8291, $\widehat{\epsilon} = .6407$), respectively, for the three simple effect tests of M at G_1, M at G_2, and M at G_3.

As with the univariate approach, I prefer the use of a separate error matrix (nonpooled) in arriving at a multivariate statistic. Using this approach, the M repeated measurements for the n_g subjects at a fixed level of G are transformed into M – 1 D variables and a test of the simple Within-Subjects effect is performed using Hotelling's (1931) T^2 statistic. Upper $100(1 - \alpha)$ percentage points of Hotelling's T^2 distribution can be obtained from the relationship

$$F = \frac{n_g - M + 1}{(n_g - 1)(M - 1)} T^2 \sim F[\alpha; M - 1, n_g - M + 1] . \quad (8)$$

The three multivariate tests equal: (a) F = 10.43 (3, 4; p = .0232), (b) F = 20.07 (3, 7; p = .0008), and (c) F = .37 (3, 10; p = .7748), respectively, for the three simple effect tests M at G_1, M at G_2, and M at G_3 (Note that the multivariate criteria are equal in this instance.).

Finally, with either approach and in order to limit the familywise rate of Type I error to α for the set of simple effect tests, each simple effect test should be conducted using a reduced significance level (Kirk, 1995; Maxwell & Delaney, 1990). That is, a Bonferroni procedure can be adopted to control the overall level of significance. Accordingly, with either approach to simple effect testing, only M at G_2 is statistically significant since its p-value is less than $\alpha/3 = .05/3 = .0167$.

Interaction Contrasts

A second method of assessing the interaction effect is to perform a series of interaction (tetrad) contrasts. Tetrad contrasts, are most useful for teasing out interaction effects in large factorial designs (Lix & Keselman, 1995, 1996). As previously indicated, this

method truly explores interaction effects, which is *not* the case with simple main effect tests (Boik, 1993).

Adopting a multivariate approach and letting $D_{ig} = X_{ig1} - X_{ig2}$, the interaction contrast can be conceptualized as a Between-Subjects contrast ($\widehat{\psi}$) between say G_1 and G_2 on the dependent variable, D_{ig}. In this conceptualization, the test statistic is

$$t = \frac{\widehat{\psi}}{\sqrt{\frac{\sum_g c_g^2 s_{g(D)}^2}{n_g}}}, \tag{9}$$

where $s_{g(D)}^2$ is the variance of the D variable at level g. This statistic can be approximated by Student's t distribution with estimated Welch (1938) degrees of freedom given by

$$\nu_W = \frac{[\sum_g c_g^2 s_g^2/n_g]^2}{\sum_g \frac{[c_g^2 s_g^2/n_g]^2}{n_g - 1}}. \tag{10}$$

Familywise control can be achieved using a Bonferroni critical value, $t[\alpha/(2c); \nu_W]$, where $c = [G(G-1)/2][M(M-1)/2]$ (see Lix & Keselman, 1996). For our data set, there are a total of $c = 3 \times 6 = 18$ tetrad contrasts. Adopting the nonpooled statistic the contrast tests equal:

Contrast	WJ	df	p-value
$(GM_{11} - GM_{12})$ vs. $(GM_{21} - GM_{22})$	1.08	7.38	.3258
$(GM_{11} - GM_{12})$ vs. $(GM_{31} - GM_{32})$	0.21	6.35	.6650
$(GM_{21} - GM_{22})$ vs. $(GM_{31} - GM_{32})$	2.97	13.51	.1077
$(GM_{11} - GM_{13})$ vs. $(GM_{21} - GM_{23})$	0.69	7.34	.4314
$(GM_{11} - GM_{13})$ vs. $(GM_{31} - GM_{33})$	0.02	6.28	.8842
$(GM_{21} - GM_{23})$ vs. $(GM_{31} - GM_{33})$	7.94	12.77	.0147
$(GM_{11} - GM_{14})$ vs. $(GM_{21} - GM_{24})$	0.63	6.76	.4532
$(GM_{11} - GM_{14})$ vs. $(GM_{31} - GM_{34})$	0.44	6.36	.5302
$(GM_{21} - GM_{24})$ vs. $(GM_{31} - GM_{34})$	24.16	16.81	.0001
$(GM_{12} - GM_{13})$ vs. $(GM_{22} - GM_{23})$	0.39	14.72	.5394
$(GM_{12} - GM_{13})$ vs. $(GM_{32} - GM_{33})$	8.49	15.34	.0105
$(GM_{22} - GM_{23})$ vs. $(GM_{32} - GM_{33})$	4.69	19.58	.0429
$(GM_{12} - GM_{14})$ vs. $(GM_{22} - GM_{24})$	0.00	9.81	.9725
$(GM_{12} - GM_{14})$ vs. $(GM_{32} - GM_{34})$	9.36	8.08	.0154
$(GM_{22} - GM_{24})$ vs. $(GM_{32} - GM_{34})$	21.75	17.34	.0002
$(GM_{13} - GM_{14})$ vs. $(GM_{23} - GM_{24})$	0.11	7.85	.7486
$(GM_{13} - GM_{14})$ vs. $(GM_{33} - GM_{34})$	3.85	6.90	.0913
$(GM_{23} - GM_{24})$ vs. $(GM_{33} - GM_{34})$	25.23	16.84	.0001.

With a Bonferroni critical value ($.05/18 = .0028$), only $(GM_{21} - GM_{24})$ vs. $(GM_{31} - GM_{34})$, $(GM_{22} - GM_{24})$ vs. $(GM_{32} - GM_{34})$, and $(GM_{23} - GM_{24})$ vs. $(GM_{33} - GM_{34})$ would be judged statistically significant. Interpretively, there are just three pairwise differences between levels of the repeated measures variable (M_1 vs. M_4, M_2 vs. M_4, and M_3 vs. M_4) that vary among just two levels of the Between-Subjects grouping variable (G_2 and G_3) .

Summary

The intention of this article was to present tests of hypotheses and sub-hypotheses in repeated measures designs that have not previously been discussed in this journal (Jennings, 1987; Keselman & Keselman, 1988; Keselman & Rogan, 1980). Specifically, methods for obtaining valid tests of the repeated measures main and interaction effect hypotheses as well as for probing these effects were presented. In addition to presenting

methods not previously discussed, the paper indicated how users can obtain solutions with various statistical algorithms that are readably available.

The recommendations offered differed according to whether the design was balanced or unbalanced; that is, researchers should choose an analysis strategy based on whether group sizes are equal or unequal. When equal, multivariate techniques were recommended. On the other hand, when group sizes are unequal, the multivariate tests will be invalid when the covariance homogeneity assumption is not satisfied, particularly when data are also nonnormal. Accordingly, in this case, it was recommended that researchers adopt the Welch-James test to investigate omnibus and sub-effect hypotheses. It is important to note that the Welch-James approach compares favorably to the multivariate approach with regard to sensitivity to detect nonnull effects and, hence, researchers wishing to follow a unified approach to significance testing can choose to uniformly adopt the Welch-James nonpooled statistic for both balanced and unbalanced designs.

Footnotes

1. The data were generated from a multivariate lognormal distribution with marginal distributions based on $Y_{ijk} = \exp(X_{ij})$ $(i = 1, \ldots, n_j)$ where X_{ijk} is distributed as N(0, .25); this distribution has skewness and kurtosis values of 1.75 and 5.90, respectively. Furthermore, the correlational (covariance) structure of the data was determined by setting the sphericity parameter ϵ at .57. Additionally, the between-subjects covariance matrices were made to be unequal such that the elements of the matrices were in the ratio of 1:3:5 (i.e., $\boldsymbol{\Sigma}_1 = 1/3\boldsymbol{\Sigma}_2 = 5/3\boldsymbol{\Sigma}_3$). When group sizes were unequal they were negatively related to the unequal covariance matrices. That is, the smallest n_j was associated with the covariance matrix containing the largest element values and the largest n_j was associated with the covariance matrix containing the smallest element values (See Footnote 2).

2. For our 3×4 design, *if* the covariance matrices equaled

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 8 & 2 & -1 & -2 \\ & 6 & 0 & 0 \\ & & 4 & 1 \\ & & & 2 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 16 & 9 & 5 & 3 \\ & 12 & 5 & 4 \\ & & 8 & 4 \\ & & & 4 \end{bmatrix}, \text{ and } \boldsymbol{\Sigma}_3 = \begin{bmatrix} 24 & 16 & 11 & 8 \\ & 18 & 10 & 8 \\ & & 12 & 7 \\ & & & 6 \end{bmatrix} \text{ and}$$

sample sizes were $n_1 = 13$, $n_2 = 10$, and $n_3 = 7$, then a negative pairing (relationship) of covariance matrices and sample sizes exists; however, if $n_1 = 7$, $n_2 = 10$, and $n_3 = 13$, then a positive relationship between the two exists.

3. In a 3×4 design, a contrast vector to compare means {e.g., $[\mu_1\ \mu_2\ \mu_3\ \mu_4]$ }among the levels of the repeated measures variable could be defined as [1 -1 0 0 1 0 -1 0 1 0 0 -1]. Though this example contains simple (D $\equiv$ pairwise) contrasts (coefficients), the vector can contain any set of linearly independent contrasts.

4. It is easy to show how comparisons between means at a fixed level of one variable do not merely represent a probing of interaction effects. For example, in a two-way factorial design let $Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$ be the less than full rank model

($j = 1, \ldots$, J, $k = 1, \ldots$, K), where α_j and β_k are the effects for the row and column variables, respectively, and $(\alpha\beta)_{jk}$ represents the interaction effect. In terms of the parameters of the model, a comparison between say $\mu_{11} - \mu_{12}$ (a comparison between columns one and two within the first row of j) would be equivalent to $[\mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}] - [\mu + \alpha_1 + \beta_2 + (\alpha\beta)_{12}] \quad = (\beta_1 - \beta_2) + [(\alpha\beta)_{11} - (\alpha\beta)_{12}]$. Thus, this comparison confounds effects due to the column variable with interaction effects.

References

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

Algina, J. (in press). 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.

Algina, J. & Keselman, H.J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. Psychological Methods, 2, 208-218.

Algina, J., & Keselman, H.J. (in press). A power comparison of the Welch-James and

Improved General Approximation tests in the split-plot design. Journal of Educational and Behavioral Statistics.

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

Best, D.J., & Rayner, J.C.W. (1987). Welch's approximate solution for the Behrens-Fisher problem. Technometrics, 29**,** 205-210.

Betz, M.A., & Gabriel, K.R. (1978). Type IV errors and analysis of simple effects. Journal of Educational Statistics, 3, 121-143.

Boik, R.J. (1993). The analysis of two-factor interactions in fixed effects linear models. Journal of Educational Statistics, 18, 1-40.

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.

BMDP Statistical Software Inc. (1994). BMDP New System for Windows, Version 1. Author, Los Angeles.

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.

Davidson, M.L. (1972). Univariate versus multivariate tests in repeated measures experiments. Psychological Bulletin, 77, 446-452.

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

Games, P.A., & Howell, J.F. (1976). Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study. Journal of Educational Statistics, 1, 113-125.

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

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

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

Hochberg, Y., & Tamhane, A.C. (1987). Multiple Comparison Procedures. New York: John Wiley.

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. Berkeley: University of California Press.

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

Huynh, H.S., & 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.

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.

Jennings, J.R. (1987). Editorial policy on analyses of variance with repeated measures. Psychophysiology, 24, 474-475.

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. (1993). Stepwise multiple comparisons of repeated measures means under violations of multisample sphericity. In: Fred M. Hoppe (ed.) Multiple Comparisons, Selection, and Applications in Biometry. New York: Marcel Dekker, Inc.

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

Keselman, H.J., Algina, J., Kowalchuk, R.K., & Wolfinger, R.D. (1997a). A comparison of recent approaches to the analysis of repeated measurements. Unpublished manuscript.

Keselman, H.J., Algina, J., Kowalchuk, R.K., & Wolfinger, R.D. (1997b). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. Unpublished manuscript.

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

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

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

Keselman, H.J., Keselman, J.C., & Shaffer, J.P. (1991). Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity. Psychological Bulletin, 110**,** 162-70.

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

Kirk, R.E. (1995). Experimental design: Procedures for the behavioral sciences. 3rd Ed. New York: Brooks/Cole.

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

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.

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

Lix, L.M., & Keselman, H.J. (1995). Approximate degrees of freedom tests: A unifiedperspective 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.

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

Maxwell, S.E., & Delaney, H.D. (1990). Designing Experiments and Analyzing Data: A Model Comparison Perspective, Belmont California: Wadsworth.

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

Mendoza, J.L., Toothaker, L.E., & Crain, B.R. (1976). Necessary and sufficient conditions for F ratios in the $L \times J \times K$ factorial design with two repeated factors. Journal of the American Statistical Association,71**,** 992-993.

Muller, K.E. & Barton, C.N. (1989). Approximate power for repeated measures ANOVA lacking sphericity. Journal of the American Statistical Association, 84**,** 549-555.

Muller, K.E. & Barton, C.N. (1991). Correction to "Approximate power for repeated measures ANOVA lacking sphericity." Journal of the American Statistical Association, 86**,** 255-256.

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

Norusis, M.J. (1993). SPSS for Windows, Advanced Statistics, Release 6.0. SPSS Inc., Chicago.

O'Brien, R.G., & Muller, K.E. (1993). Unified power analysis for t-tests through multivariate hypotheses. In: L.K Edwards ed. Applied analysis of variance in behavioral science. New York: Marcel Dekker.

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

Overall, J.E., & Doyle, S.R. (1994). Estimating sample sizes for repeated measurement designs. Controlled Clinical Trials, 15, 100-123.

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

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.

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

SAS Institute. (1990). SAS/STAT User's Guide: Volume 2, GLM-VARCOMP, Version 6. Author, Cary, NC.

SAS Institute. (1996). SAS/STAT Software: Changes and Enhancements through release 6.11; Author, Cary, NC.

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

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

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

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

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

Toothaker, L.E. (1991). Multiple comparisons for researchers. Newbury Park: Sage.

Vassey, M.W., & Thayer, J.F. (1987). The continuing problem of false positives in repeated measures ANOVA in psychophysiology: A multivariate solution. Psychophysiology, 24, 479-486.

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 Student's problem when several different population variances are involved. Biometrika, 34, 28-35.

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

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

Table 1. Hypothetical data for a 3 X 3 Repeated Measures Design

Subject	M1	M2	M3	M4
G1				
1	14.26	4.22	4.24	5.83
2	-1.35	3.61	4.79	6.98
3	-6.65	-1.45	1.45	4.21
4	7.22	4.79	7.12	8.10
5	-0.44	1.03	2.33	3.03
6	0.58	3.35	4.40	9.41
7	13.91	6.47	7.40	5.30
(8)	7.73	-0.07	0.80	0.00
(9)	-4.51	-2.14	-0.92	-1.18
(10)	1.60	2.90	-0.53	-1.08
(11)	3.24	2.26	0.68	0.68
(12)	7.41	0.82	0.61	-1.18
(13)	5.58	1.22	0.16	-0.14

Table 1, continued

Subject	M1	M2	M3	M4
G2				
14	-0.99	-1.38	0.14	2.38
15	-2.46	-1.54	0.01	2.53
16	-4.28	-2.79	-0.65	2.35
17	-1.27	1.03	1.94	2.89
18	-4.11	0.10	1.18	0.95
19	-3.33	2.48	1.51	2.66
20	-3.51	-2.06	0.13	1.94
21	-4.78	-1.49	1.19	3.34
22	0.46	1.88	1.45	4.52
23	5.88	2.88	2.78	4.94
(24)	-3.51	-2.06	0.13	1.94
(25)	-4.78	-1.49	1.19	3.34
(26)	0.46	1.88	1.45	4.52

Table 1, continued

Subject	M1	M2	M3	M4
G3				
27	-1.13	0.06	0.23	0.87
28	-1.30	-0.15	0.41	0.87
29	3.20	0.92	1.03	-0.06
30	3.65	1.67	1.16	0.41
31	-1.79	-1.05	-0.80	-0.06
32	-0.24	-1.06	0.03	0.79
33	0.33	0.60	0.53	0.45
34	0.09	2.05	0.29	0.40
35	-1.15	-1.04	-0.26	0.25
36	-1.11	0.02	0.08	0.10
37	-0.59	0.18	1.25	0.60
38	2.34	1.38	2.21	0.33
39	1.87	4.03	0.79	1.02

Note: M1-M3: levels of the Within-Subjects Repeated Measures variable; G1-G3: levels of the Between-Subjects grouping variable. For the balanced data set n_1=13, n_2=13, and n_3=13. For the unbalanced data set n_1=7, n_2=10, and n_3=13.

Table 2. Cell and Marginal Unweighted Means

(a) Balanced data set:

	M1	M2	M3	M4	Row Mean
G1	3.74	2.08	2.50	3.07	2.85
G2	-2.02	-0.20	0.96	2.95	.42
G3	0.32	.59	.53	.46	.48
Column Mean	0.68	.82	1.33	2.16	1.25

(b) Unbalanced data set:

	M1	M2	M3	M4	Row Mean
G1	3.93	3.15	4.53	6.12	4.43
G2	-1.84	-0.09	0.97	2.85	.47
G3	0.32	.59	.53	.46	.48
Column Mean	0.80	1.22	2.01	3.14	1.79

Note: See the note from Table 1.