DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 3 | Issue 2 Article 17

11-1-2004

Aligned Rank Tests As Robust Alternatives For Testing Interactions In Multiple Group Repeated Measures Designs With Heterogeneous Covariances

Xiaosheng Lei
University of Alabama, Birmingham

Janet K. Holt
Northern Illinois University

T. Mark Beasley
University of Alabama, Birmingham, mbeasley@uab.edu

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm


Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
Lei, Xiaosheng; Holt, Janet K.; and Beasley, T. Mark (2004) "Aligned Rank Tests As Robust Alternatives For Testing Interactions In Multiple Group Repeated Measures Designs With Heterogeneous Covariances," *Journal of Modern Applied Statistical Methods*: Vol. 3 : Iss. 2 , Article 17.
DOI: 10.22237/jmasm/1099268220
Available at: http://digitalcommons.wayne.edu/jmasm/vol3/iss2/17

This Emerging Scholar is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

Journal of Modern Applied Statistical Methods
November, 2004, Vol. 3, No. 2, 462-475
Copyright © 2004 JMASM, Inc.
1538 – 9472/04/$95.00

Aligned Rank Tests As Robust Alternatives For Testing Interactions In Multiple Group Repeated Measures Designs With Heterogeneous Covariances

Xiaosheng Lei
University of Alabama, Birmingham

Janet K. Holt
Northern Illinois University

T. Mark Beasley
University of Alabama, Birmingham

Data simulation was used to investigate whether tests performed on aligned ranks (Beasley, 2002) could be used as robust alternatives to parametric methods for testing a split-plot interaction with non-normal data and heterogeneous covariance matrices. Results indicated the aligned rank method do not have any distinct advantage over parametric methods in this situation.

Key words: Nonparametrics, repeated measures, covariance heterogeneity, split-plot, interaction

Introduction

Repeated measures designs involving two or more independent groups are among the most common experimental designs (see Keselman & Algina, 1996). The parametric technique used to analyze a design in which a repeated measures (i.e., within-subjects) factor is crossed with a between-subjects (i.e., independent grouping or treatment variable) factor is the split-plot analysis of variance (ANOVA). It can be expressed with the following linear model:

$$Y_{ijk} = \mu_{**} + \beta_j + \pi_{i(j)} + \tau_k + \beta\tau_{jk} + \tau\pi_{ik(j)} + \zeta_{ijk}, \quad (1)$$

where j is referenced to the J groups of the between-subjects factor, i is referenced to the n_j subjects nested within the j^{th} group, k is referenced to the K levels of the within-subjects factor, ζ_{ijk} is a random error vector, and $N = \Sigma n_j$ is the total number of subjects.

The interaction of the between-subjects and the repeated measures factors is often of most interest in many applications of the split-plot design (Boik, 1993). It is tested with an F-ratio, $F_{(Y)}$, that is distributed approximately as $F_{[(J-1)(K-1),(N-J)(K-1)]}$ under the null hypothesis:

$$H_{0(JxK)}: \beta\tau_{jk} = 0, \text{ for all } j \text{ and } k. \quad (2)$$

When the ANOVA model in (1) involves a within-subjects factor with $K > 2$, it requires the pooled within-group covariance matrix to be *spherical* (Huynh & Feldt, 1970). For the univariate $F_{(Y)}$ from model (1), the sphericity assumption implies that the random error components, ζ_{ijk}, are NID(0, σ_ζ^2) for each of the JK cells. Several procedures that correct $F_{(Y)}$ by an ε factor have been developed to adjust the degrees of freedom so that $F_{\varepsilon(Y)}$ will be a valid test of the interaction when there are departures from sphericity (e.g, Huynh, 1978).

Another suggested approach for dealing with non-spherical data is the use of multivariate tests because they do not require sphericity of the covariance matrix. Multivariate test statistics assume multivariate normality for the K repeated measures. Because repeated measures designs can be analyzed with multivariate tests applied to (K-1) transformed variables (Marascuilo & Levin, 1983), the multivariate normality assumption applied to split-plot designs implies

Xiaosheng Lei is a teacher at International Industrial and Commercial College, South-Central University for Nationalities, China. This article is based on her Master's project in Biostatistics at the UAB. Janet K. Holt is Associate Professor in the Department of Educational Technology, Research, and Assessment. This article was conducted while she was a Visiting Professor in Section Statistical Genetics, Department of Biostatistics at UAB. T. Mark Beasley is Associate Professor in the Section Statistical Genetics, Department of Biostatistics. Email: mbeasley@uab.edu.

that the random error components are independent and multivariate normal with means of zero and a common covariance matrix (i.e., NID[$\mathbf{0}_{(K\text{-}1)}$, $\mathbf{C}_K\mathbf{\Sigma}\mathbf{C}_K'$], where $\mathbf{0}_{(K\text{-}1)}$ is a (K-1) vector of zeros, $\mathbf{C}_K$ is a (K-1)xK normalized matrix of contrasts among the K repeated measures, and $\mathbf{\Sigma}$ is the KxK pooled within-group population covariance matrix. In order to pool these covariance matrices across the J groups, however, there is the implicit assumption that they are equal:

$$\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_j \ldots = \mathbf{\Sigma}_J . \quad (3)$$

If these covariance matrices are not equal, multivariate statistics are known to be invalid in terms of inflated Type I error rates, especially with unequal sample sizes (Olson, 1974).

In practice, it is likely that both the sphericity and normality assumptions are violated. However, multivariate tests are prone to inflate Type I error rates with violations of the multivariate normality assumption, especially with a small sample size to number of repeated measures (N/K) ratio (e.g., Blair, Higgins, Karniski, & Kromrey, 1994). By contrast, univariate tests are generally conservative with data sampled from heavy-tailed distributions (Wilcox, 1993). Thus, as compared to their multivariate extensions, univariate tests are noted to be more robust to non-normality. For example, simulation studies have indicated that $F_{\varepsilon(Y)}$ adequately corrects for non-sphericity (Huynh, 1978) and is reasonably robust to non-normality (Keselman, Algina, Kowalchuck, & Wolfinger, 1999). However, there are many skewed, heavy-tailed distributions that can affect the performance of both univariate (e.g., Wilcox, 1993; Zimmerman & Zumbo, 1993) and multivariate parametric tests (e.g., Blair et al., 1994; Keselman, Carriere, & Lix, 1993).

Beasley (2002) suggested an aligned rank procedure as a robust alternative to testing the interaction in split-plot designs when the normality assumption is violated. A univariate approach was detailed for situations in which the sphericity assumption holds and multivariate approach was also suggested for the more common case of non-spherical covariance structures. These procedures demonstrated more statistical power than parametric procedures when error distributions were highly skewed; however, the issue of heterogeneous covariance matrices was not addressed.

Heterogeneity of variance is known to affect the Type I error rate of both univariate (Scheffé, 1957) and multivariate tests (Olson, 1974). Two approaches for testing interaction effects in repeated measures designs when the homogeneity of covariance assumption does not hold are the approximate degrees of freedom (*df*) multivariate Welch-James (WJ) statistic (Johansen, 1980; Keselman, Algina, Wilcox, & Kowalchuk, 2000) and the Huynh (1978) Improved General Approximation (IGA) tests. Simulation studies have shown these two approaches to be generally robust. However, under some conditions of departures form normality, sphericity and variance homogeneity, the WJ and IGA procedures have been found to yield inflated Type I error rates (Algina & Keselman, 1998; Keselman, Kowalchuk, & Boik, 2000). The purpose of this study was to investigate whether Beasley's (2002) aligned rank procedure could be used as a robust alternative to parametric procedures, when the normality and homogeneity of covariance assumptions were violated. Specifically, we investigated whether applying the WJ or IGA test to aligned ranks controlled Type I error rates when covariance matrices and sample sizes were unequal.

Rank-based competitors relax the normality assumptions by assuming that the random error components are *independent identically distributed* (IID) random variables from some continuous distribution, not necessarily the normal (i.e., NID). The rank transform concept is appealing because from a univariate perspective all data points (Y_{ijk}) are observations of one dependent variable measured under K different conditions or time points. Because the rank transform is monotonic, it is commonly believed that the null hypothesis for the parametric test of interaction (i.e., $F_{(Y)}$) from model (1) is similar to the null hypothesis for similar tests performed on ranks (e.g., $F_{(R)}$), except statistical inferences concern mean ranks. However, when test statistics for interactions used in parametric analyses of factorial designs

are applied to monotone transformations (e.g., rank transformation), the resulting tests lack an invariance property (Headrick & Sawilowsky, 2000). Specifically, the expected value of ranks for an observation in one cell will have a non-linear dependence on the original means of the other cells. Thus, interaction and main effect relationships are not expected to be maintained after rank transformations are performed (e.g., Blair, Sawilowsky, & Higgins, 1987).

Given these problems encountered by interaction tests based on the rank transform when other non-null effects are present (e.g., Blair et al., 1987; Toothaker & Newman, 1994), one solution is to treat other effects as nuisance parameters and remove them from the scores before ranking and analysis. McSweeney (1967) developed a chi-square approximate statistic for testing the interaction using aligned ranks in the two-way layout. Hettmansperger (1984) developed a linear model approach in which the nuisance effects are removed by obtaining the residuals from a regression model. However, both of these alignment procedures were developed for the two-way between-subjects factorial design and thus are not desirable because they do not remove the subjects' individual differences effect that is nested in the between-subjects factor, $\pi_{i(j)}$ from model (1). Higgins and Tashtoush (1994) proposed subtracting the subject effect and the repeated measures main effect and then ranking the aligned data from 1 to *NK* as follows:

$$R_{ijk} = \text{Rank}(Y_{ijk} - \overline{Y}_{ij*} - \overline{Y}_{*k} + \overline{Y}_{**}), \quad (4)$$

where $\overline{Y}_{*k}$ is the marginal mean of the k^{th} measure averaged over all *N* subjects, $\overline{Y}_{ij*}$ is the mean for the i^{th} subject averaged across the *K* measures, and $\overline{Y}_{**}$ is the grand mean of all *NK* observations. Following Hettmansperger (1984), this alignment could also be accomplished by obtaining the residuals from a linear model in which Y_{ijk} is regressed on a set of (*N*–1) dummy codes that represent the subjects effect ($\pi_{i(j)}$) and a set of (*K*-1) contrast codes that represent the repeated-measures main effect (τ_k) from model (1).

Univariate Approach

Consistent with Iman, Hora, and Conover (1984), Higgins and Tashtoush (1994) recommended applying the split-plot ANOVA from model (1) to the aligned ranks ($F_{(R)}$), thus replacing Y_{ijk} with R_{ijk}. It should be noted, however, that many of the properties of the original data transmit to ranks, including heterogeneity of variance (Zimmerman & Zumbo, 1993) and non-sphericity (Harwell & Serlin, 1994). Thus, when performing the split-plot ANOVA *F* on aligned ranks, *df*-correction methods may be employed if the pooled covariance matrix is non-spherical (e.g., $F_{\varepsilon(R)}$) or if the between-subjects covariance matrices are heterogeneous (e.g., $\text{IGA}_{(R)}$).

ε-adjusted F-test

With increasing departures from sphericity, the ANOVA *F*-ratio demonstrates a general lack of robustness, resulting in increasingly liberal tests. Huynh and Feldt (1976) developed an ε-adjusted test for split-plot models. Lecoutre (1991) corrected this formula so that in split-plot designs $\hat{\varepsilon}$ is replaced with $\tilde{\varepsilon}$:

$$\tilde{\varepsilon} = \frac{(N - J + 1)(K - 1)\hat{\varepsilon} - 2}{(K - 1)(N - J - (K - 1)\hat{\varepsilon})}, \quad (5)$$

where $\hat{\varepsilon}$ is a sphericity parameter estimated from the sample pooled within-group covariance matrix (**S**), $\mathbf{S} = \Sigma[(n_j\text{-}1)(N\text{-}J)]\mathbf{S}_j$. $\mathbf{S}_j$ is the sample covariance matrix for the j^{th} group with elements:

$$s_{kk'} = \Sigma\Sigma(R_{ijk} - \overline{R}_{jk})(R_{ijk} - \overline{R}_{jk'})/(n_j\text{-}1),$$

and

$$\hat{\varepsilon} = \frac{[\text{tr}(\mathbf{C}_K\mathbf{S}\mathbf{C}'_K)]^2}{(K-1)[\text{tr}(\mathbf{C}_K\mathbf{S}\mathbf{C}'_K)^2]}. \quad (6)$$

The Lecoutre adjusted test for the interaction, $F_{\varepsilon(Y)}$, is distributed approximately as

$$F_{[\tilde{\varepsilon}\,(J\text{-}1)(K\text{-}1),\ \tilde{\varepsilon}\,(N\text{-}J)(K\text{-}1)]}.$$

Keselman et al. (1999) reported that $F_{\varepsilon(Y)}$ provided effective Type I error control for non-normal data with non-spherical covariance structures; however, it demonstrated low power under several conditions. We will examine the statistical properties of calculating the $\hat{\varepsilon}$ estimate and the ε-adjusted *F*-test from the aligned ranks ($F_{\varepsilon(R)}$).

Improved General Approximate

$F_{\varepsilon(Y)}$ was designed to correct for non-sphericity only. Jointly, the assumptions of sphericity and homoscedasticity in split-plot designs are referred to as *multi-sample sphericity* (Huynh, 1978). When covariance matrices are unequal across levels of the between-subjects factor and the design is unbalanced, the ε-adjusted *F* statistics as well as multivariate approaches are not robust for testing the interaction (Huynh, 1978; Keselman & Keselman, 1990).

In cases of arbitrary (i.e., non-spherical and/or heteroscedastic) covariance matrices, Huynh (1978) proposed the IGA procedure to estimate the *df*s for the test statistics in the split-plot design. In order to adjust the tests for violations of multi-sample sphericity, the IGA procedure uses $\tilde{c}F_{[\alpha,\tilde{h}'',\tilde{h}]}$ as the critical value for the interaction test. The statistics for these critical values are defined in terms of the separate covariance matrices for each of the *J* groups, **S***j*. Let **S*** denote a block diagonal matrix with $\mathbf{S}_j/n_j$ as the *j*th diagonal block. All off-diagonal blocks consist of a (*K* x *K*) matrix of zeros. Also let **D** = {**I**-(**1**)(**1**´)/*K*] where **I** is a *K* dimensional identity matrix and **1** is a (*K*x1) vector of ones. Define **G** as a matrix constructed of J^2(*K*x*K*) blocks. The *j*th diagonal block of **G** is n_j(1-n_j/*N*)**D** and the off-diagonal blocks are (-n_j´ n_j **D**/*N*). For testing the split-plot interaction:

$$\tilde{c} = \frac{(N-J)\mathrm{tr}(\mathbf{GS}^*)}{(J-1)\sum_{j=1}^{J}(n_j-1)\mathrm{tr}(\mathbf{DS}_j)} \qquad (7)$$

and

$$\hat{h}'' = \frac{[\mathrm{tr}(\mathbf{GS}^*)]^2}{\mathrm{tr}(\mathbf{GS}^*)^2} \quad . \qquad (8)$$

Algina and Oshima (1994) applied the Lecoutre correction to the IGA so that

$$\tilde{h}'' = \frac{(J-1)[(N-J+1)\hat{h}'' - 2(J-1)]}{(N-J)(J-1) - \hat{h}''} \qquad (9)$$

Let $\mathbf{A}_j$ = tr($\mathbf{C}_K\ \mathbf{S}_j\ \mathbf{C}_K$ `), $\mathbf{B}_j$ = tr($\mathbf{DS}_j$)^2, and $\tilde{h} = \hat{\eta}/\hat{\delta}$, where

$$\hat{\eta} = \sum_{j=1}^{J}\frac{(n_j-1)}{(n_j+1)(n_j-2)}(n_jA_j^2 - 2B_j) + \sum_{j=1}^{J}\sum_{j'\neq j}^{J}(n_j-1)(n_{j'}-1)A_jA_{j'} \qquad (10)$$

and

$$\hat{\delta} = \sum_{j=1}^{J}\frac{(n_j-1)}{(n_j+1)(n_j-2)}[(n_j-1)B_j - A_j^2]. \qquad (11)$$

We will examine the statistical properties of performing the Huyhn's (1978) IGA test on the aligned ranks (IGA$_{(R)}$).

Multivariate Approach

Another suggested approach for dealing with non-spherical data is the use of multivariate tests because they do not require sphericity of the covariance matrix. However, multivariate tests have strict sample size requirements based on the number of repeated measures. Furthermore, the degrees-of-freedom (*df*s) for the error term of the univariate $F_{(Y)}$ can be much larger than the error *df*s (df_e) for the *F* approximate tests for the multivariate approach. Thus, the multivariate approach may have less statistical power in small sample situations (Keselman & Algina, 1996).

Agresti and Pendergast (1986) recommended a multivariate *F*-test based on Hotelling's (1931) T^2 for testing repeated measures effects in a single sample design. Their results showed that this multivariate test held the Type I error rate near the nominal alpha with departures from normality and sphericity.

Harwell and Serlin (1997) confirmed these results and also demonstrated that the Akritas and Arnold (1994) chi-square approximate test, which is functionally related to the Agresti-Pendergast test, inflated Type I error rates with total sample sizes of $N = 30$ or less. However, these findings are limited to the single sample repeated measures design.

To extend the Agresti and Pendergast (1986) approach for testing the interaction in a split-plot design, define **E** as a K x K pooled-sample cross-product error matrix for the aligned ranks (4) with elements:

$$e_{kk'} = \Sigma\Sigma(R_{ijk} - \overline{R}_{jk})(R_{ijk} - \overline{R}_{jk'}) . \quad (12)$$

Let $\mathbf{E}_p$ be a JK x JK block diagonal matrix where the j^{th} block of the main "diagonal" for $\mathbf{E}_p$ is defined as $\mathbf{E}/n_j$, and all other off-diagonal blocks are zero. That is, $\mathbf{E}_p$ is the Kronecker product of a diagonal matrix $\mathbf{n}^* = \text{diag}\{1/n_1, 1/n_2, \ldots, 1/n_J\}$ and **E**, $\mathbf{E}_p = \mathbf{n}^* \otimes \mathbf{E}$. Also, define $\mathbf{R}_{JK} = [\overline{R}_{11}, \overline{R}_{12}, \ldots \overline{R}_{1K}, \overline{R}_{21}, \ldots \overline{R}_{2K}, \ldots \overline{R}_{J1}, \ldots \overline{R}_{JK}]'$ as a JK-dimensional vector of mean ranks and $\mathbf{C}_{JK}$ as a $(J\text{-}1)(K\text{-}1)$ x JK contrast matrix that represents the interaction. In general, $\mathbf{C}_{JK}$ can be defined as $\mathbf{C}_{JK} = \mathbf{C}_J \otimes \mathbf{C}_K$, where $\mathbf{C}_J$ is a $(J\text{-}1)$xJ contrast matrix for the between-subjects effect and $\mathbf{C}_K$ is a $(K\text{-}1)$xK contrast matrix for the repeated measures effect.

Based on Agresti and Pendergast (1986), the distribution of the statistic,

$$H_{(R)} = (\mathbf{C}_{JK}\mathbf{R}_{JK})'(\mathbf{C}_{JK}\mathbf{E}_p\mathbf{C}'_{JK})^{-1}(\mathbf{C}_{JK}\mathbf{R}_{JK}) \quad (13)$$

multiplied by $(N\text{-}1)$, should approximate a χ^2 distribution with $df = (J\text{-}1)(K\text{-}1)$ asymptotically. It should be noted that $H_{(R)}$ is the Hotelling-Lawley trace for the interaction effect from a multivariate profile analysis performed on the Rank Transformed scores. Consistent with Agresti and Pendergast (1986), transforming H to an F-test may better control Type I error rates as opposed to comparing $(N\text{-}1)H_{(R)}$ to a chi-square distribution with $df = (J\text{-}1)(K\text{-}1)$, especially with smaller sample sizes (Harwell & Serlin, 1997). Based on Hotelling (1951), $H_{(R)}$ (13) is transformed to an F approximation statistic by:

$$F_{H(R)} = [2(sn+1)/(s^2(2m+s+1))]H_{(R)} , \quad (14)$$

where $s = \min[(J\text{-}1),(K\text{-}1)]$, $m = [(|K\text{-}J|\text{-}1)/2]$, and $n = [(N\text{-}J\text{-}K)/2]$. This F approximation has numerator dfs of $df_h = [s(2m+s+1)] = [(J\text{-}1)(K\text{-}1)]$ and denominator dfs of $df_e = [2(sn+1)]$. Alternatively, a researcher could obtain a critical value for $H_{(R)}$ (13) from the sampling distribution of the Hotelling-Lawley trace using the s, m, and n parameters.

Keselman et al. (1993) suggested the use of the Welch-James test (Johansen, 1980) test for unbalanced within-subjects designs when covariance matrices were heterogeneous. The test statistic uses the same quadratic form as (13); however, separate covariance matrices are used:

$$WJ_{(R)} = (\mathbf{C}_{JK}\mathbf{R}_{JK})'(\mathbf{C}_{JK}\mathbf{S}^*\mathbf{C}'_{JK})^{-1}(\mathbf{C}_{JK}\mathbf{R}_{JK}) \quad (15)$$

where, $\mathbf{S}^*$ is a JK x JK block diagonal matrix where the j^{th} block of the main "diagonal" is defined as $\mathbf{S}_j/n_j$, and all other off-diagonal blocks are zero, $\mathbf{S}^* = \mathbf{n}^* \otimes \mathbf{S}$. The $WJ_{(R)}/c$ is distributed approximately as $F[f_1, f_2]$ with $f_1 = (J\text{-}1)(K\text{-}1)$, $f_2 = f_1\,(f_1+2)/3A$, $c = f_1 + 2A - 6A/(f_1+2)$ and

$$A = \tfrac{1}{2}\sum_{j=1}^{J}[\text{tr}\{\mathbf{S}\mathbf{C}'_K(\mathbf{C}_K\mathbf{S}\mathbf{C}')\mathbf{C}_K\mathbf{Q}_j\}^2 + \{\text{tr}(\mathbf{S}\mathbf{C}'_K(\mathbf{C}_K\mathbf{S}\mathbf{C}')\mathbf{C}_K\mathbf{Q}_j)^2\}]/(n_j - 1) .$$

The $\mathbf{Q}_j$ matrix is a JK x JK block diagonal matrix corresponding to the j^{th} group. The $(s,t)^{th}$ block of $\mathbf{Q}_j$ is $\mathbf{I}_K$ if $s=t=j$ and $\mathbf{0}$ otherwise.

Olson (1974) showed that the Pillai-Bartlett trace (V) was more robust to violations to the normality and homogeneity of covariance assumptions. Applied to the aligned ranks it is computed as:

$$V_{(R)} = (\mathbf{C}_{JK}\mathbf{R}_{JK})'(\mathbf{C}_{JK}\mathbf{T}\mathbf{C}'_{JK})^{-1}(\mathbf{C}_{JK}\mathbf{R}_{JK}) \quad (16)$$

where, **T** is the Total sum of Squares matrix with elements defined as:

$$t_{kk'} = \Sigma\Sigma(R_{ijk} - \bar{R}_{*k})(R_{ijk} - \bar{R}_{*k'}) ,$$

and $\bar{R}_{*k}$ is the aligned rank mean for the k^{th} measure for all J groups combined. $V_{(R)}$ (16) is transformed to an F approximation statistic by:

$$F_{V(R)} = [(2n+s+1)/(2m+s+1)][V/(s-V)] . \quad (17)$$

This F approximation has numerator dfs of $df_h = [s(2m+s+1)] = [(J-1)(K-1)]$ and denominator dfs of $df_e = [s(2n+s+1)]$. Again, a researcher could obtain a critical value for V (16) from the sampling distribution of the Pillai-Bartlett trace using the s, m, and n parameters.

For aligned ranks, the major purpose of the alignment process (4) is to remove the nuisance effects (i.e., main effects) so that test statistics will be sensitive to the effect of interest (i.e., interaction). The alignment process simply removes the mean values for the nuisance main effects, thus involving linear transformations of the data; however, the aligned ranks are a monotone transformation of the aligned data. Therefore, the aligned ranks (R_{ijk}) are placeholders for the percentiles of the original data (Y_{ijk}) with the nuisance location parameters removed. In either case, there is no guarantee that test statistics performed on R_{ijk} will reflect differences in location parameters without additional assumptions.

For the univariate test to be valid, under the null hypothesis in (2), not only are all J groups expected to have identical error distributions, but the error distributions for the K repeated measures are also expected to be identically distributed: NID(0, σ_ζ^2) for all j and k. Similar to this sphericity assumption for univariate parametric tests, a rank-based version simply does not require *normal* error distributions. Thus, for rank-based tests, if the univariate assumption that all JK cells have *identically* shaped error distributions with a common variance (i.e., IID[0,σ_ζ^2] for all j and k) is tenable, then statistically significant values for test statistics performed on the aligned ranks (4) implies that the interaction is due to shifts in the location parameters (Lehmann, 1998). To illustrate the shift model for the univariate approach to the split-plot design, define the null hypothesis as:

$$H_0(JxK): \mathbf{G}_1(\mathbf{Y}_1 - \mathbf{1}\Delta_1) = \mathbf{G}_2(\mathbf{Y}_2 - \mathbf{1}\Delta_2) = \ldots = \mathbf{G}_j(\mathbf{Y}_j - \mathbf{1}\Delta_j) = \ldots = \mathbf{G}_J(\mathbf{Y}_J - \mathbf{1}\Delta_J) \quad (18)$$

where $\mathbf{G}_j(\mathbf{Y}_j)$ is the K-dimensional distribution function of the original scores for the j^{th} group, $\mathbf{Y}_j$ is the NxK data matrix for the j^{th} group, $\Delta_j = [\delta_{j1}\ \delta_{j2} \ldots \delta_{jk} \ \ldots \delta_{jK}]$ is a 1xK vector of location parameters for the j^{th} group, and **1** is an Nx1 vector of ones (Agresti & Pendergast, 1986, p. 1418). By requiring the univariate IID[0,σ_ζ^2] assumption, if (18) is true then a statistically significant test statistic (i.e., $F_{(R)}$) implies that the interaction is due to shifts in location parameters, a result conceptually similar to a rejection of the parametric null hypothesis in (2).

To illustrate the shift model for the multivariate approach to the split-plot design, define the null hypothesis as:

$$H_0(JxK): G_1(Y_{1k} - \delta_{1k}) = G_2(Y_{2k} - \delta_{2k}) = \ldots = G_j(Y_{jk} - \delta_{jk}) = \ldots = G_J(Y_{Jk} - \delta_{Jk}) , \quad (19)$$

$$\text{for } k = 1, \ldots K .$$

$G_j(Y_{jk})$ is the one-dimensional distribution function of the k^{th} repeated measure for the j^{th} group, Y_{jk} is the Nx1 data matrix for the j^{th} group on the k^{th} measure and δ_{jk} is a scalar location parameter for the jk^{th} cell. This is similar to the NID[$\mathbf{0}_{(K-1)}$,$\mathbf{C}_K\mathbf{\Sigma}\mathbf{C}_K'$] assumption for multivariate parametric tests except normal error distributions are not required. Under the multivariate model assumption that the random error vectors are IID[$\mathbf{0}_{(K-1)}$,$\mathbf{C}_K\mathbf{\Sigma}\mathbf{C}_K'$] across the J groups, if (19) is true then a statistically significant multivariate test statistic performed on R_{ijk} implies that the interaction is due to shifts in location parameters. Again, this is a result conceptually similar to a rejection of the parametric null hypothesis in (2) and thus tests of shift parameter models (18 or 19) could be

used as robust alternatives to parametric procedures for testing interactions.

Note that the null hypotheses (18) and (19) are equivalent in terms of location parameters. If (18) is true so is (19); however, if (19) is true, it does not imply that (18) is true. Likewise, a false (18) does not imply a false (19). These distinctions are important because in order to test a null hypothesis of shifts in location parameters analogous to the null hypothesis in (2), the univariate null model for ranks (18) requires an assumption that the data for all *JK* cells are sampled from identically shaped distributions with a common variance. By contrast, the multivariate null model for ranks (19) only requires an assumption that the error distributions for each of the *K* repeated measures are identical for each of the *J* groups; however, there is no assumption that the error distributions for all *K* repeated measures are identically distributed. Thus, the relationship between the multivariate approach to analyzing aligned ranks and the *F*-ratio performed on aligned ranks is analogous to the relationship of the multivariate approach to repeated measures designs and the univariate approach that requires the sphericity assumption (Agresti & Pendergast, 1986).

Strictly speaking, if the assumption in (3) does not hold (i.e., the covariance matrices are heterogeneous), then neither the univariate (i.e., IID[0, σ_ζ^2] for all *j* and *k*) nor multivariate IID[$\mathbf{0}_{(K-1)}$,$\mathbf{C}_K\mathbf{\Sigma}\mathbf{C}_K'$] assumptions hold. The IGA test, Welch-James statistic, and the Pillai trace criterion have been shown to be generally robust to departures from homogeneous covariance asumption (3) for testing interaction among location parameters when normality holds. Thus, we investigated the use of the IGA, Welch-James, and Pillai tests applied to aligned ranks (4) as a robust alternative to testing interactions among location parameters (i.e., shift models 18 and 19) when assumptions of normality, sphericity, and homogeneous covariance matrices (3) do not hold.

Methodology

A 3 (sample size: *N* = 30, 90, and 150) x 3 (balanced, conservative unbalanced, and liberal unbalanced samples) x 2 (covariance structure: spherical and non-spherical) x 3 (shape of error distribution: normal, double exponential, and exponential) factorial design was employed for this simulation study. For each of these conditions, 10,000 replications were generated using SAS/IML 8.2 (SAS Institute, 2001). Comparisons were made among procedures for testing the interaction effect in a *J*=3 x *K*=4 split-plot design at the α=0.05 significance level. For the aligned ranks (R_{ijk}), the following nine statistics were calculated: (a) the conventional *F*-test; (b) the Lecoutre (1991) ε-adjusted *F*; (c) the $\text{IGA}_{(R)}$; (d) $H_{(R)}$ (13) using a critical value from the Hotelling-Lawley trace distribution, (e) the *F* approximate test for $H_{(R)}$ (14); (f) the $WJ_{(R)}$ test (15), (g) $V_{(R)}$ (16) using a critical value from the Pillai-Bartlett trace distribution, and (h) the *F* approximate test for $V_{(R)}$ (17).

For a *J*=3 x *K*=4 split-plot design, the parameters for both the Hotelling-Lawley trace and Pillai-Bartlett trace distribution are *s* = 2, *m* = 0, *n* = 11.5 for *N* = 30, *n* = 41.5 for *N* = 90, and *n* = 71.5 for *N* = 150. The α=.05 critical values for *H* are 0.587, 0.156, and 0.089 for *N* = 30, 90 and 150, respectively. The α=.05 critical values for *V* are 0.407, 0.139, 0.086 for *N* = 30, 90, and 150, respectively.

The *N* = 30 condition was chosen because it has been used in other simulation studies (e.g., Agresti & Pendergast, 1986; Blair et al., 1987). Also, Harwell and Serlin (1997) reported that for a single sample, repeated measures design the multivariate *F* approximate test of rank transformed scores inflated Type I error rates with a total sample size of *N* = 30. For an unbalanced sample size, we used **n** = {5, 10, 15} for the "conservative" or positive pairing and the reverse for the "liberal" or negative pairing. For an unbalanced sample size with *N*=90 and *N*=150, we used **n** = {15, 30, 45} and **n** = {25, 50, 75}, respectively, for the "conservative" or positive pairings and the reverse for the "liberal" or negative pairings.

The double exponential distribution was chosen as a condition where the errors were symmetric but heavy-tailed with skewness and kurtosis values of γ_1=0 and γ_2=3, respectively. The exponential distribution was selected as a condition where the errors were skewed (γ_1=2)

and extremely heavy-tailed (γ_2=6). Wilcox (1993) has noted that heavy-tailed distributions are common in practice and tend to inflate variances which in turn reduces power. In the case of empirical alpha rates, heavy-tailed distributions are likely to lead to Type I error rates that are below the nominal alpha. Micceri (1989) reported that 30.9% of the data from educational and psychological research had asymmetry as extreme as that of the exponential distribution. Furthermore, the exponential distribution condition is similar to the lognormal distribution (γ_1=1.75; γ_2=5.90) used in other simulation studies (e.g., Algina & Keselman, 1998; Algina & Oshima, 1994; Keselman et al., 1993). Moreover, it is representative of skewed, heavy-tailed distributions found in experimental psychology, most notably reaction time data (Zumbo & Coulombe, 1997).

Using the SAS/IML RANNOR function, a (n_j) by (K=4) matrix of normally distributed random variates with zero means and unit variances ($\mathbf{X}_j$) was generated for each of the J=3 groups. A covariance matrix $\mathbf{\Sigma}_j$ was subsequently imposed on the $\mathbf{X}_j$ scores by deriving a KxK matrix of principal component coefficients, $\mathbf{F}$, from the pre-specified covariance matrix ($\mathbf{\Sigma}_j$) and pre-multiplying it by the transpose of $\mathbf{X}_j$ to create a data matrix $\mathbf{Y}_j$ that simulates $\mathbf{\Sigma}_j$:

$$\mathbf{Y}_j' = \mathbf{F}\,\mathbf{X}_j' \qquad (20)$$

(Beasley, 1994; Kaiser & Dickman, 1962).

In the first condition, all population correlations between measures (i.e., off-diagonal elements of $\mathbf{\Sigma}_j$) were ρ = 0.60. This condition yielded results for a spherical covariance structure (ε = 1) in which case the univariate F-tests should not inflate Type I error rates with homogeneous covariance matrices. In the second condition, covariance structures with ε = 0.64 were imposed. The pairwise intercorrelations were ρ_{12} and ρ_{34} = 0.70 with all other population correlations equal to 0.30. These values were taken from Headrick and Sawilowsky (1999) and represent a realistic situation in which the sphericity assumption is violated because a measure taken at time point k=1 is more correlated with a measure taken at time k=2 than it is with measures taken later in the experiment (i.e., time points k=3 and 4). Likewise, measures taken at time points k=3 and 4 were more correlated with each other than with previous measurements.

Two conditions of error non-normality were simulated: exponential and double exponential. To simulate the error distributions for both non-normal conditions, intermediate population correlation values were derived (see Headrick & Sawilowsky, 1999) for each of the three covariance structure conditions described above. First, the random normal variates ($\mathbf{X}_j$) were generated. Then, a matrix of principal component coefficients, $\mathbf{F}$, was derived from the intermediate values for the pre-specified correlation matrix. Subsequently, covariance structures with the intermediate values were imposed using (20). Then, data transformations using an extended Fleishman (1978) power method were performed (Headrick & Sawilowsky, 1999).

This process yielded data with zero means, unit variances, and the expected covariance structure ($\mathbf{\Sigma}_j$) after the non-linear transformations were performed to make these values non-normal. Thus, these values were transformed so that the variances and shapes of each of the K error components were the same. This transformation process was also completed for each of the J=3 groups so that there were no between-group differences in variance or shape. Thus, under conditions in which the covariance matrices were homogeneous and spherical, the random error components (ζ_{ijk}) were IID(0, σ_ζ^2) for each of the JK cells, which permitted an investigation of the test statistics as robust alternative tests of interaction in terms of a univariate shift model for location parameters (18). Under conditions in which the covariance structures were homogeneous but not spherical, however, only the less restrictive multivariate assumption (IID[$\mathbf{0}_{(K\text{-}1)}$,$\mathbf{C}_K\mathbf{\Sigma}\mathbf{C}_K'$]) was valid, thus creating a violation of the assumptions for the univariate parametric F-tests.

To impose heterogeneous variances, the second group (j=2) was multiplied by $\sqrt{3}$ and

the third group *(j=3)* was multiplied by $\sqrt{5}$, thus yielding a $\boldsymbol{\Sigma}_1 = 3\boldsymbol{\Sigma}_2 = 5\boldsymbol{\Sigma}_3$ ratio. This variance ratio has been used in several other simulation studies (e.g., Keselman, et al., 2000). A repeated measures main effect pattern resulting in no interaction was imposed (Blair et al., 1987, p. 1143) after multiplication to increase variance was completed. Specifically for group 1, a vector of constants, $\mathbf{c}_1$ = [0 0 1 0], was added to each observation for the K=4 repeated measures. For group 2, $\mathbf{c}_2$ = [-.5 -.5 .5 -.5]. For group 3, $\mathbf{c}_3$ = [-1 -1 0 -1].

When covariance matrices were not homogeneous then both univariate and multivariate IID assumptions were violated, and thus, we investigated whether tests performed on aligned ranks (4) can be used as robust alternatives to testing interactions among location parameters under this extreme violation of the shift model assumptions.

Results

For all tables, $F_{(R)}$ refers to the univariate ANOVA F-test, $F_{\varepsilon(R)}$ refers to the Lecoutre (1991) ε-adjusted F, $\text{IGA}_{(R)}$ refers to the Improved General Approximate, $H_{(R)}$ refers to testing the Hotelling-Lawley trace (13) with a critical value from its referent distribution, $F_{\text{H}(R)}$ refers to the F approximation (14), $WJ_{(R)}$ refers to the Welch-James test (15), $V_{(R)}$ refers to testing the Pillai-Bartlett trace (13) with a critical value from its referent distribution, $F_{\text{H}(R)}$ refers to the F approximation (14), $WJ_{(R)}$ refers to the Welch-James test (15), $V_{(R)}$ refers to testing the Pillai-Bartlett trace (13) with a critical value from its referent distribution, and $F_{V(R)}$ refers to the F approximation (18). The subscript R indicates that the tests were performed on the aligned ranks (R_{ijk}). The results for the condition in which the K=4 repeated measures were equicorrelated and thus spherical are labeled as ε = 1.00 and ε = 0.64 refers to the non-spherical condition.

For this study, tests that demonstrated a Type I error rate lower than 0.05 were considered conservative but acceptable, while those with rates that were significantly above the nominal alpha were considered unacceptably liberal. Given α=0.05 and 10,000 replications, a simulated estimate has a standard error of 0.0022. Thus, for empirical estimates of Type I error rates, any rejection rate two standard errors above 0.05 (i.e., 0.0544) was considered significantly liberal. This is consistent with Bradley's (1978) criterion of non-robustness in which the empirical Type I error rate should never exceed 1.1α. Likewise, any rejection rate below 0.0456 was considered significantly below the nominal alpha (i.e., conservative).

Tables 1, 2, and 3 show the rejection rates for the eight tests under conditions of heterogeneous covariance matrices. It is apparent that, for the conditions simulated in this study, none of the tests adequately controlled the Type I error rate when assumption (3) did not hold. As expected, most tests, with the exception of $\text{IGA}_{(R)}$ and $WJ_{(R)}$, produced rejections rates well above the nominal alpha with a liberal sample size-covariance pairing.

Also as expected, rejection rates for most tests were significantly below the nominal alpha with a conservative sample size-covariance pairing. The $\text{IGA}_{(R)}$ and $WJ_{(R)}$ were the best at controlling the Type I error rate. That is, these two procedures had rejection rates that were closest to the nominal alpha but were nevertheless unacceptably liberal under many conditions. Rejection rates for $\text{IGA}_{(R)}$ were similar for both sample sizes of N=30 and 90. By contrast, rejection rates for $WJ_{(R)}$ became less liberal with an increase in sample size from N=30 to 90. Therefore, $WJ_{(R)}$ was more sensitive to smaller sample sizes. A larger sample size of N=150 was used to investigate whether the $\text{IGA}_{(R)}$ and $WJ_{(R)}$ tests would eventually yield Type I error rates near the nominal alpha. Although these rejection rates reported in Table 3 are closer to α=0.05, these values were consistently around 6 to 7.5% rejection.

Conclusion

One reason to use tests based on aligned ranks is that they have demonstrated superior power for detecting interactions in split-plot designs when error distributions are identically skewed with a common variance (Beasley, 2002). However, heterogeneous covariance matrices violate both the univariate (i.e., IID[0, σ_{ζ}^{2}] for all *j* and *k*) and multivariate IID[$\mathbf{0}_{(K-1)}$,$\mathbf{C}_K\mathbf{\Sigma}\mathbf{C}_K'$] assumptions. Results indicated that although the $WJ_{(R)}$ and $IGA_{(R)}$ produced relatively stable rejection rates across sample size – covariance pairing conditions, both tests yielded rejection rates significantly above the nominal alpha. However, $WJ_{(R)}$ required a much larger sample size (N=150) to produce rejection rates consistently around 6 to 7.5%. Perhaps, additional *df* correction could be applied, but it must be considered that the conditions imposed in this simulation study are rather extreme violations of the IID assumptions. Furthermore, for sample sizes this large the Type I error rates for the Welch-James test performed on the original non-normal ($WJ_{(Y)}$) are as close to the nominal alpha (Keselman et al., 2000) as the error rates for the Welch-James test performed on the aligned ranks ($WJ_{(R)}$; see Table 3). Moreover, for larger sample sizes the expected power advantage of $WJ_{(R)}$ over $WJ_{(Y)}$ is likely to be negligible, except for extremely small interaction effects. Thus, when covariance matrices are drastically unequal, it appears that aligned rank procedures cannot be used as robust alternatives to testing interaction among location parameters (i.e., shift models 18 and 19). Therefore, issues concerning the interpretation of rank-based tests are of concern.

Multivariate procedures performed on aligned ranks test a null hypothesis of distributional equivalence across the *J* groups for each of the *K* measures (Beasley, 2002).

However, situations where distributional equivalence does not hold but location parameters are identical only occur in symmetric distributions (Vargha & Delaney, 1998). Hence the null hypothesis being tested with asymmetric distributions and heterogeneous variances with rank data becomes one of location and variance differences. In other words, imposing the situation of unequal variances violates the restrictive assumption of the shift model (Lehmann, 1998) and explains the inflated Type I error rates that occur in the $F_{(R)}$ results. The effects of distributional nonequivalence are manifested in the Type I error rates of the other rank statistics tested in this study, including the Welch-James, the IGA, and the Pillai trace.

Therefore, WJ and IGA, as well as other tests, performed on the aligned ranks cannot be used as robust alternatives to testing the interaction in a split-plot design when assumption (3) does not hold. That is, when covariance matrices are heterogeneous, tests performed on the aligned ranks will detect between-group distributional differences to some extent, and thus, a statistically significant result cannot be attributed solely to differences among location parameters.

This is important because there are situations where the interaction null hypothesis in (19) would be rejected and the researcher might assume it was due to differences in location parameters when in actuality the rejection resulted from other between-group distributional (i.e., variance, shape) differences (Agresti & Pendergast, 1986; Beasley, 2002; Serlin & Harwell, 2001; Vargha & Delaney, 1998). For this reason, we do not recommend the Welch-James, the IGA, or the Pillai trace as tests of interaction among location parameters if covariance heterogeneity is suspected.

Table 2. Empirical Type I Error Rates (α=.05) for the Interaction Tests in the Presence of a Repeated Measures Main Effect (c = .50) with a $\mathbf{\Sigma}_1 = 3\mathbf{\Sigma}_2 = 5\mathbf{\Sigma}_3$ ratio and N=90.

n_1 n_2 n_3	Normal		Double Exponential		Exponential	
30 30 30[B]	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64
$F_{(R)}$	.0748	.0981	.0734	.1015	.0797	.0982
$F_{\varepsilon(R)}$	.0743	.0744	.0732	.0746	.0786	.0737
IGA$_{(R)}$	.0692	.0694	.0676	.0688	.0730	.0690
$H_{(R)}$	.0766	.0736	.0724	.0731	.0761	.0751
$F_{H(R)}$	.0803	.0788	.0766	.0761	.0806	.0797
$WJ_{(R)}$	.0776	.0725	.0776	.0755	.0826	.0823
$V_{(R)}$	.0787	.0770	.0746	.0741	.0785	.0776
$F_{V(R)}$	.0766	.0744	.0724	.0710	.0767	.0751
15 30 45[C]	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64
$F_{(R)}$	.0263	.0473	.0300	.0521	.0335	.0541
$F_{\varepsilon(R)}$	.0260	.0323	.0299	.0366	.0333	.0375
IGA$_{(R)}$	.0601	.0602	.0644	.0649	.0647	.0623
$H_{(R)}$	.0266	.0255	.0290	.0287	.0332	.0331
$F_{H(R)}$	.0298	.0273	.0314	.0310	.0345	.0355
$WJ_{(R)}$	.0713	.0727	.0783	.0823	.0875	.0842
$V_{(R)}$	.0275	.0253	.0291	.0295	.0335	.0338
$F_{V(R)}$	.0263	.0245	.0286	.0285	.0320	.0327
45 30 15[L]	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64
$F_{(R)}$	.1441	.1518	.1374	.1460	.1259	.1483
$F_{\varepsilon(R)}$	.1433	.1204	.1373	.1138	.1245	.1154
IGA$_{(R)}$	.0731	.0712	.0691	.0667	.0667	.0711
$H_{(R)}$	.1382	.1370	.1366	.1307	.1197	.1270
$F_{H(R)}$	.1444	.1434	.1428	.1368	.1261	.1346
$WJ_{(R)}$	.0805	.0754	.0782	.0710	.0740	.0763
$V_{(R)}$	.1414	.1397	.1402	.1346	.1244	.1307
$F_{V(R)}$	.1389	.1359	.1360	.1314	.1221	.1284

Table 3. Empirical Type I Error Rates (α=.05) for the Interaction Tests in the Presence of a Repeated Measures Main Effect (c = .50) with a $\mathbf{\Sigma}_1 = 3\mathbf{\Sigma}_2 = 5\mathbf{\Sigma}_3$ ratio and N=150.

n_1 n_2 n_3	Normal		Double Exponential		Exponential	
50 50 50[B]	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64
$F_{(R)}$	.0784	.0912	.0734	.0993	.0771	.0958
$F_{\varepsilon(R)}$	.0780	.0677	.0731	.0750	.0764	.0678
IGA$_{(R)}$	.0719	.0651	.0687	.0707	.0716	.0645
$H_{(R)}$	.0774	.0668	.0730	.0768	.0759	.0745
$F_{H(R)}$	.0794	.0684	.0743	.0789	.0775	.0767
$WJ_{(R)}$	.0798	.0706	.0740	.0778	.0791	.0802
$V_{(R)}$	.0784	.0663	.0724	.0779	.0759	.0747
$F_{V(R)}$	.0776	.0655	.0713	.0766	.0752	.0732
25 50 75[C]	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64
$F_{(R)}$	.0266	.0477	.0274	.0481	.0387	.0573
$F_{\varepsilon(R)}$	.0264	.0338	.0274	.0344	.0384	.0421
IGA$_{(R)}$	.0577	.0597	.0605	.0601	.0686	.0679
$H_{(R)}$	.0265	.0271	.0281	.0295	.0381	.0336
$F_{H(R)}$	.0275	.0282	.0289	.0305	.0391	.0344
$WJ_{(R)}$	.0666	.0732	.0670	.0729	.0833	.0802
$V_{(R)}$	.0266	.0270	.0283	.0287	.0381	.0341
$F_{V(R)}$	.0262	.0263	.0275	.0284	.0372	.0337
75 50 25[L]	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64	ε = 1.00	ε = 0.64
$F_{(R)}$	.1460	.1563	.1373	.1593	.1378	.1463
$F_{\varepsilon(R)}$	.1455	.1181	.1367	.1247	.1370	.1142
IGA$_{(R)}$	.0697	.0668	.0720	.0711	.0767	.0721
$H_{(R)}$	.1433	.1402	.1381	.1356	.1338	.1297
$F_{H(R)}$	.1464	.1431	.1404	.1387	.1366	.1324
$WJ_{(R)}$	.0765	.0755	.0792	.0744	.0787	.0751
$V_{(R)}$	.1437	.1403	.1373	.1366	.1347	.1305
$F_{V(R)}$	.1420	.1393	.1361	.1354	.1332	.1301

References

Agresti, A., & Pendergast, J. (1986). Comparing mean ranks for repeated measures data. *Communications in Statistics (A): Theory & Method, 15,* 1417-1433.

Akritas, M.G., Arnold, S. F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated-measures designs. *Journal of the American Statistical Association, 89,* 336-343.

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 & 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 & Statistical Psychology, 47,* 151-165.

Beasley, T. M. (1994). CORRMTX: Generating correlated data matrices in SAS/IML. *Applied Psychological Measurement, 18*, 95.

Beasley, T. M., (2002) Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multivariate Behavioral Research*, *37*, 197-226.

Blair, R. C., Higgins, J. J., Karniski, W., & Kromrey, J. D. (1994). A study of multivariate permutation tests which may replace Hotelling's $\underline{T}^2$ test in prescribed circumstances. *Multivariate Behavioral Research, 29*, 141-163.

Blair, R. C., Sawilowsky, S. S., Higgins, J. J. (1987). Limitations of the rank transform statistic in test for interactions. *Communications in Statistics (B): Simulation & Computation, 16*, 1133-1145.

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

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

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.

Harwell, M. R., Serlin, R. C. (1994). A Monte Carlo study of the Friedman test and some competitors in the single factor, repeated measures design with unequal covariances. *Computational Statistics & Data Analysis, 17*, 35-49.

Harwell, M. R., Serlin, R. C. (1997). An empirical study of five multivariate tests for the single-factor repeated measures model. *Communications in Statistics (B): Simulation & Computation, 26*, 605-618.

Headrick, T. C., Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika*, *64*, 25-35.

Headrick, T. C., Sawilowsky S. S. (2000). Properties of the rank transformation in factorial analysis of covariance. *Communications in Statistics (B): Simulation & Computation, 29*, 1059-1088.

Hettmansperger, T .P. (1984). *Statistical Inference Based on Ranks*. Wiley, New York.

Higgins, J. J., Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World, 1*, 201-211.

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. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics & Probability*, 2, 23-41.

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

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

Huynh, H., Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 1*, 69-82.

Iman, R. L., Hora, S. C., Conover, W. J. (1984). Comparison of asymptotically distribution-free procedures for the analysis of complete blocks. *Journal of the American Statistical Association*, *79*, 674-685.

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

Kaiser, H. F., Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika***,** *27*, 179-182.

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

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

Keselman, H. J., Algina, J., Wilcox, R. R. Kowalchuk, R. K. (2000). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James test again. *Educational & Psychology 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, J. C., Keselman, H. J. (1990). Analysing unbalanced repeated measures designs. *British Journal of Mathematical & Statistical Psychology, 43*, 265-282.

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 & Statistical Psychology, 53*, 51-67.

Lecoutre, B. (1991). A correction for the e approximate test in repeated measures designs with two or more independent groups. *Journal of Educational Statistics, 16*, 371-372.

Lehmann, E. L. (1998). *Nonparametrics: Statistical methods based on ranks* (Revised 1st Ed.). Upper Saddle River, NJ: Prentice-Hall.

Marascuilo, L. A., Levin, J. R. (1983). *Multivariate methods for the social science*: *A researcher's handbook*. Monterey, CA, Brooks/Cole.

McSweeney, M. (1967). An empirical study of two proposed nonparametric test for main effects and interaction (Doctoral dissertation, University of California-Berkeley, 1968). *Dissertation Abstracts International*, *28*(11), 4005.

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

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

SAS Institute, 2001. *SAS/IML user's guide* (Release 8.2). Cary, NC.

Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.

Serlin, R. C., Harwell, M. R. (April, 2001). *A review of nonparametric test for complex experimental designs in educational research*. Paper presented at the American Educational Research Association. Seattle, WA.

Toothaker, L. E., Newman, D. (1994). A. Nonparametric competitors to the two way ANOVA. *Journal of Educational & Behavioral Statistics, 19*, 237-273.

Vargha, A., Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational & Behavioral Statistics, 23*, 170-192.

Wilcox, R. (1993). Robustness in ANOVA. In Edwards, E. (Ed.), *Applied analysis of variance in the behavioral sciences*, pp. 345-374. Marcel Dekker, New York.

Zimmerman, D., Zumbo, B. D. (1993). Relative power of the Wilcoxon test, the Friedman test, and the repeated-measures ANOVA on ranks. *Journal of Experimental Education, 62*, 75-86.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology, 51*, 139-149.