Multivariate Behavioral Research, *37* (2), 197-226
Copyright © 2002, Lawrence Erlbaum Associates, Inc.

Multivariate Aligned Rank Test for Interactions in Multiple Group Repeated Measures Designs

T. Mark Beasley
Department of Biostatistics
University of Alabama at Birmingham

This study showed that a multivariate test of interactions for aligned ranks in a split-plot design controlled Type I error rates for non-normal data with non-spherical covariance structures. Furthermore, it performed well in the presence of a strong repeated measures main effect, whereas tests performed on rank transformed scores demonstrated severely inflated Type I error rates. This test also demonstrated more statistical power than parametric tests performed on non-normal data sampled from a skewed, heavy-tailed distribution. Methods for conducting multiple comparisons are proposed.

Aligned Rank Test for Testing Interactions in Multiple Group Repeated Measures Designs

Repeated measures designs involving two or more independent groups are among the most common experimental designs in behavioral research (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:

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

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 $N = \boldsymbol{\Sigma} n_j$ is the total number of subjects. In applications of the split-plot design in behavioral research, the interaction of the between-subjects and the repeated measures factors is usually of most interest (Boik, 1993). It is tested with an

The author would like to thank Todd C. Headrick and Bruno D. Zumbo for their support in completing this article. Correspondence should be addressed to: T. Mark Beasley, Department of Biostatistics, University of Alabama at Birmingham, RPHB 327, 1530 Third Avenue South, Birmingham, AL 35294

F-ratio, $F_{(Y)}$, that is distributed approximately as $F_{[(J-1)(K-1),(N-J)(K-1)]}$ under the null hypothesis:

$$H_{0(J \times K)}\colon \beta\tau_{jk} = 0, \quad \text{for all } j \text{ and } k. \tag{2}$$

When the ANOVA model in Equation 1 involves a within-subjects factor with $K > 2$, it requires the pooled within-group covariance matrix to have a specific form (i.e., a *sphericity* assumption) in order for the sampling distribution of the $F_{(Y)}$ test of the interaction to approximate $F_{[(J-1)(K-1),(N-J)(K-1)]}$ under the interaction null hypothesis in Equation 2 (Huynh & Feldt, 1970). With increasing departures from sphericity, the ANOVA *F*-ratio demonstrates a general lack of robustness, resulting in increasingly liberal tests. Unfortunately, the traditional univariate $F_{(Y)}$ is commonly (mis)used when the sphericity assumption is violated (Keselman et al., 1998; Robey & Barcikowski, 1995).

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 $\hat{\varepsilon}$:

$$\tilde{\varepsilon} = \frac{(N-J+1)(K-1)\hat{\varepsilon}-2}{(K-1)[N-J-(K-1)\hat{\varepsilon}]}, \tag{3}$$

where $\hat{\varepsilon}$ is a sphericity parameter estimated from the sample pooled within-group covariance matrix (see Winer, Brown, & Michels, 1991, p. 257). The Lecoutre adjusted test for the interaction, $F_{\varepsilon(Y)}$, is distributed approximately as $F_{[\tilde{\varepsilon}(J-1)(K-1),\ \tilde{\varepsilon}(N-J)(K-1)]}$. Keselman, Algina, Kowalchuk, and Wolfinger (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.

$F_{\varepsilon(Y)}$ was designed to correct for non-sphericity only. Thus, in cases of non-spherical and/or heteroscedastic between-subjects covariance matrices, Huynh (1978) proposed the general approximate (GA) procedure to estimate the *df*s for the univariate *F*-tests in the split-plot design and the improved general approximate (IGA) for situations where covariance matrices are close to possessing sphericity and/or homoscedasticity. Both GA and IGA exhibit Type I error rates substantially below the nominal alpha in many conditions (Algina & Oshima, 1994). Because the current study does not focus on the heterogeneity of covariance matrices issue and the general approximate procedures have demonstrated low power in preliminary analyses, these techniques are excluded from further elaboration.

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).

In practice, it is likely that both the sphericity and normality assumptions are violated. Regardless of whether (a) the univariate ANOVA test with possible *df*-corrections or (b) the multivariate approach to analyzing repeated measures design is employed, there are normality assumptions. For the univariate $F_{(Y)}$ from model 1, the assumption on the random error component is that ζ_{ijk} is NID$(0, \sigma^2_\zeta)$ for each of the JK cells. 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 (see Marascuilo & Levin, 1983), the multivariate normality assumption applied to split-plot designs implies that multivariate parametric tests assume that the random error components are independent and multivariate normal with means of zero and a common covariance matrix, that is, NID$[\mathbf{0}_{(K-1)}, \mathbf{C}_K \boldsymbol{\Sigma} \mathbf{C}_K']$, where $\mathbf{0}_{(K-1)}$ is a $(K - 1)$ vector of zeros, $\mathbf{C}_K$ is a $(K - 1) \times K$ transformation matrix, and $\boldsymbol{\Sigma}$ is the $K \times K$ pooled within-group covariance matrix. 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 et al., 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 et al., 1993).

Rank-Based Alternatives

Rank-based competitors relax the normality assumptions by assuming that the random error components are independent identically distributed random variables from some continuous distribution, not necessarily the normal. However, Sawilowsky (1990) has contended that good nonparametric tests for interactions do not exist. Thus, there have been few recent studies (e.g., Akritas & Arnold, 1994; Beasley, 2000) concerning the use of rank-based tests as nonparametric alternatives for analyzing non-normal, non-spherical data from split-plot designs.

Rank Transform Procedure

Univariate Approach. One approach to transforming the data from a repeated measures experiment is to rank the entire data matrix regardless of measure or group membership. In a $J \times K$ split-plot design, scores (Y_{ijk}) from model 1 are replaced with ranks (R_{ijk}) ranging from 1 to NK, the total number of observations. Iman, Hora, and Conover (1984) suggested calculating the ANOVA F from model 1 on the Rank Transformed scores [$F_{(R)}$] as a test of the interaction in a split-plot design.

Consistent with Zimmerman and Zumbo's (1993) contention that ranks often inherit the properties of the original data, the data can still be non-spherical, although to a lesser degree, after rank transformation (Harwell & Serlin, 1994). Under the complete null hypothesis of no effects, Beasley and Zumbo (1998) showed that the ANOVA performed on R_{ijk} in a split-plot design inflated Type I errors when the original data was non-spherical. They also demonstrated that deriving $\hat{\varepsilon}$ from the Rank Transformed scores (R_{ijk}) and applying the Lecoutre (1991) ε-adjustment (3) to the split-plot ANOVA F-test [$F_{\varepsilon(R)}$] held the Type I error rate near the nominal alpha.

Multivariate Approach. 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. Unfortunately, there has been a paucity of research on rank-based interaction tests in the split-plot design.

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

$$e_{kk'} = \sum_{j=1}^{J} \sum_{i=1}^{nj} (R_{ijk} - \overline{R}_{jk})(R_{ijk} - \overline{R}_{jk'}) . \tag{4}$$

Let $\mathbf{E}^*$ be a $JK \times JK$ block diagonal matrix where the j^{th} block of the main "diagonal" for $\mathbf{E}^*$ is defined as $\mathbf{E}/n_j$, and all other off-diagonal blocks are zero. That is, $\mathbf{E}^*$ is the Kronecker product of a diagonal matrix $\mathbf{n} = \text{diag}\{1/n_1, 1/n_2, \ldots, 1/n_J)$ and $\mathbf{E}$, $\mathbf{E}^* = \mathbf{n} \otimes \mathbf{E}$. Also, define $\mathbf{R}_{JK} = (\bar{R}_{11}, \bar{R}_{12}, \ldots \bar{R}_{1K}, \bar{R}_{21}, \ldots \bar{R}_{2K}, \ldots \bar{R}_{J1}, \ldots \bar{R}_{JK})'$ as a JK-dimensional vector of mean ranks and $\mathbf{C}_{JK}$ as a $(J - 1)(K - 1) \times JK$ 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 $(J - 1) \times J$ contrast matrix for the between-subjects effect and $\mathbf{C}_K$ is a $(K - 1) \times K$ contrast matrix for the repeated measures effect. For example, in a $J = 3 \times K = 4$ split-plot design, define:

$$\mathbf{C}_J = \begin{bmatrix} 2 & -1 & -1 \\ 0 & 1 & -1 \end{bmatrix} \text{ and } \mathbf{C}_K = \begin{bmatrix} -3 & -1 & 1 & 3 \\ -1 & 1 & 1 & -1 \\ -1 & 3 & -3 & 1 \end{bmatrix}$$

as orthogonal contrast matrices. Thus, the interaction contrast matrix is:

$$\mathbf{C}_{JK} = \begin{bmatrix} -6 & -2 & 2 & 6 & 3 & 1 & -1 & -3 & 3 & 1 & -1 & -3 \\ -2 & 2 & 2 & -2 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ -2 & 6 & -6 & 2 & 1 & -3 & 3 & -1 & 1 & -3 & 3 & -1 \\ 0 & 0 & 0 & 0 & -3 & -1 & 1 & 3 & 3 & 1 & -1 & -3 \\ 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & -1 & 3 & -3 & 1 & 1 & -3 & 3 & -1 \end{bmatrix}.$$

It should be noted, however, that $\mathbf{C}_J$ and $\mathbf{C}_K$ need not be orthogonal, only full row rank.

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

$$H = (\mathbf{C}_{JK}\,\mathbf{R}_{JK})'\,(\mathbf{C}_{JK}\,\mathbf{E}^*\mathbf{C}'_{JK})^{-1}(\mathbf{C}_{JK}\,\mathbf{R}_{JK}) \tag{5}$$

multiplied by $(N - 1)$, should approximate a χ^2 distribution with $df = (J - 1)(K - 1)$ asymptotically. It should be noted that H 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 $(N - 1)H$ to a chi-square distribution with $df = (J - 1)(K - 1)$, especially with smaller sample sizes (Harwell & Serlin, 1997). Based on Hotelling (1951), H in Equation 5 is transformed to an F approximation statistic by:

$$(6) \qquad F_H = \{2(sn + 1)/[s^2(2m + s + 1)]\}H,$$

where $s = \min[(J - 1),(K - 1)]$, $m = [(|K - J| - 1)/2]$, and $n = [(N - J - K)/2]$. This F approximation has numerator dfs of $df_h = [s(2m + s + 1)] = [(J - 1)(K - 1)]$ and denominator dfs of $df_e = [2(sn + 1)]$. Alternatively, a researcher could obtain a critical value for H in Equation 5 from the sampling distribution of the Hotelling-Lawley trace using the s, m, and n parameters. Unfortunately, few multivariate texts have these critical values tabled.

Aligned Rank Transform Procedure

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, 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 transformation 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:

$$(7) \qquad A_{ijk} = \text{Rank}(Y_{ijk} - \overline{Y}_{ij*} - \overline{Y}_{*k} + \overline{Y}_{**}),$$

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.

Higgins and Tashtoush (1994) recommended applying the split-plot ANOVA from model 1 to the Aligned Ranks [$F_{(A)}$], thus replacing Y_{ijk} with A_{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). Therefore, the Aligned Ranks may inherit some of the distributional properties of the original data as well. Thus, when performing the split-plot ANOVA *F* on Aligned Ranks, the *df*-correction methods may be employed if the pooled covariance matrix is non-spherical (e.g., ε-adjusted *F*) or if the between-subjects covariance matrices are heterogeneous (e.g., GA, IGA). As an alternative to the univariate *df*-correction procedures, a multivariate test based on Agresti and Pendergast (1986) may be used. That is, the multivariate tests (Equations 5 or 6) could be performed on the Aligned Ranks (Equation 9), replacing R_{ijk} with A_{ijk}.

Assumptions and Hypotheses for Interaction Tests Performed on Ranks

It is important to note that statistically significant values of the univariate and multivariate tests performed on Rank Transformed scores (R_{ijk}) or Aligned Ranks (A_{ijk}) do not necessarily imply that the interaction is due to differences in location parameters unless additional assumptions are made. That is, because ranks inherit the distributional properties of the original data, a significant test statistic may reflect differences in other distributional characteristics (i.e., variance or shape) rather than differences in location (Serlin & Harwell, 2001). Fortunately, significant test statistics can generally be attributed to differences in location parameters (Marascuilo & McSweeney, 1977, pp. 304-305). However, credible inferences about means require the assumption that the population distributions are symmetric

(Serlin & Harwell, 2001); whereas, credible inferences concerning location parameters in general require the assumption that the population distributions are of identical shape, not necessarily symmetric (Varga & Delaney, 1998).

Strictly, statistical tests performed on the Rank Transformed scores (R_{ijk}) involve inferences concerning the *distribution of the original data* (Y_{ijk}) because ranks are "placeholders" for the percentiles of the original data (M. R. Harwell, personal communication, April 24, 2001). For Aligned Ranks, the major purpose of the alignment process (Equation 7) 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). However, 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 (A_{ijk}) are "placeholders" for the percentiles of the original data (Y_{ijk}) with the nuisance location parameters removed (M. R. Harwell, personal communication, April 24, 2001). In either case, there is no guarantee that test statistics performed on R_{ijk} or A_{ijk} will reflect differences in location parameters.

Akritas and Arnold (1994) have argued that hypotheses should be expressed in a manner that does not place additional distributional assumptions on the data. These *fully nonparametric hypotheses* differ because they do not attribute the rejection to location parameters alone but rather to any distributional differences (Marascuilo & McSweeney, 1977), a concept recently referred to as "stochastic heterogeneity" (Varga & Delaney, 1998). Thus, hypotheses of this form reduce the risk of drawing incorrect conclusions about the likely sources of the significant interaction, but do so at the cost of not being able to characterize precisely how population distributions differ (Serlin & Harwell, 2001).

To elaborate, the univariate statistics, $F_{(R)}$ and $F_{(A)}$, for a split-plot design actually test a restrictive null hypothesis of "exchangeability" or *permutational equivalence*:

$$(8) \qquad H_{0(J \times K)}\text{: } \mathbf{G}_1(\mathbf{Y}_1) = \mathbf{G}_2(\mathbf{Y}_2) = \ldots = \mathbf{G}_j(\mathbf{Y}_j) = \ldots = \mathbf{G}_J(\mathbf{Y}_J)\ ,$$

where $\mathbf{G}_j(\mathbf{Y}_j)$ is the K-dimensional distribution function of the original scores for the j^{th} group (Agresti & Pendergast, 1986, p. 1418). This implies that under the null hypothesis in Equation 8 not only are all J groups expected have identical error distributions, but the error distributions for the K repeated measures are also expected to be identically distributed: IID$(0, \sigma^2_\zeta)$ for all j and k. This is similar to the NID$(0, \sigma^2_\zeta)$ assumption for univariate parametric tests except normal error distributions are not required.

The multivariate procedures (Equations 5 or 6) test a broader null hypothesis of between-group marginal homogeneity:

$$(9)\ H_{0(J \times K)}: G_1(Y_{1k}) = G_2(Y_{2k}) = \ldots = G_j(Y_{jk}) = \ldots = G_J(Y_{Jk})_{\,,} \quad \text{for } k = 1, \ldots K\,,$$

where $G_j(Y_{jk})$ is the one-dimensional distribution function of the k^{th} repeated measure for the j^{th} group. Strictly, this is a null hypothesis of *distributional equivalence* across the J groups for each of the K measures. That is, each of the K repeated measures may have different distributions, but as long as there are no distributional differences across the J groups, (Equation 9) is true. Thus, to obtain the asymptotic null distributions of the test statistics (Equations 5 or 6), it is only necessary to assume the null hypothesis of between-group distributional equivalence of error terms in Equation 9: [IID(0, σ^2_ζ) for all j for each k separately] or IID[$\mathbf{0}_{(K-1)}, \mathbf{C}_K\mathbf{\Sigma}\mathbf{C}_K'$]. Thus, it is not necessary to make stronger assumptions concerning joint (or permutational) distributions (i.e., common correlations between pairs of measures) as in Equation 8. 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.

Strictly, rejections of these null hypotheses (Equations 8 and 9) typically imply a pattern in which one of the J groups is stochastically larger that the other(s) on at least one of the K repeated measures and that this “stochastic superiority” is not constant across all K repeated measures. To illustrate, imagine a $J = 2$ groups (e.g., Control and Treatment) by $K = 3$ repeated measures (e.g., Pretest, Posttest, Follow-up) design. Suppose that for the first measure ($k = 1$) the two groups are stochastically identical, $G_1(Y_{11}) = G_2(Y_{21})$, which would be expected on a pretest if the groups were randomly assigned. Thus for all real values, u, the probability of scores larger than u is the same in both groups, $P(Y_{11} > u) = P(Y_{21} > u)$. Now imagine that the posttest ($k = 2$) was measured after some treatment had been administered to second group ($j = 2$) while the first group remained a control. If the treatment “worked,” then the second group should have higher scores, and thus, $G_1(Y_{12}) \neq G_2(Y_{22})$. Because the Treatment group has scores (Y_{21}) that are stochastically larger than the scores for the Control group (Y_{11}) the between-group probabilities of scores larger than all real values (u) are not equal, $P(Y_{11} > u) \leq P(Y_{21} > u)$.

Shift Model for Ranks in Split-Plot Designs

If the univariate assumption that all JK cells have identically shaped error distributions with a common variance [i.e., IID(0, σ^2_ζ) for all j and k] is tenable, then rejection of the null hypothesis (Equation 8) must be 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 in Equation 8 as:

$$(10) H_{0(J \times K)}: \mathbf{G}_1(\mathbf{Y}_1 - \mathbf{1}\boldsymbol{\Delta}_1) = \mathbf{G}_2(\mathbf{Y}_2 - \mathbf{1}\boldsymbol{\Delta}_2) = \ldots = \mathbf{G}_j(\mathbf{Y}_j - \mathbf{1}\boldsymbol{\Delta}_j) = \ldots = \mathbf{G}_J(\mathbf{Y}_J - \mathbf{1}\boldsymbol{\Delta}_J)$$

where $\mathbf{Y}_j$ is the $N \times K$ data matrix for the j^{th} group, $\boldsymbol{\Delta}_j = (\delta_{j1}\ \delta_{j2} \ldots \delta_{jk} \ldots \delta_{jK})$ is a $1 \times K$ vector of location parameters for the j^{th} group, and $\mathbf{1}$ is an $N \times 1$ vector of ones. By requiring the univariate IID$(0, \sigma^2_\zeta)$ assumption, if Equation 10 is true then a statistically significant test statistics [i.e., $F_{(R)}$ or $F_{(A)}$] and a rejection of Equation 8 implies that the interaction is due to shifts in location parameters, a result conceptually similar to a rejection of the parametric null hypothesis in Equation 2.

To illustrate the shift model for the multivariate approach to the split-plot design, define the null hypothesis in Equation 9 as:

$$(11) H_{0(J \times K)}: G_1(\mathbf{Y}_{1k} - \delta_{1k}) = G_2(\mathbf{Y}_{2k} - \delta_{2k}) = \ldots = G_j(\mathbf{Y}_{jk} - \delta_{jk}) = \ldots = G_J(\mathbf{Y}_{Jk} - \delta_{Jk})$$

,

for $k = 1, \ldots K$,

where $\mathbf{Y}_{jk}$ is the $N \times 1$ 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. Thus, under the multivariate model assumption that the random error vectors are IID$[\mathbf{0}_{(K-1)}, \mathbf{C}_K \boldsymbol{\Sigma} \mathbf{C}_K']$ across the J groups, if Equation 11 is true then a statistically significant multivariate test statistics (Equations 5 or 6) performed on R_{ijk} or A_{ijk} and a rejection of Equation 9 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 Equation 2.

Note that the null hypotheses (Equations 10 and 11) are equivalent in terms of location parameters. Also, the location parameters within a group (δ_{jk}) are not required to be equal under Equations 10 or 11. In other words, repeated measures main effects may exist in the absence of an interaction. Furthermore, it is important to note that if Equation 10 or 11 is false, then Equations 8 and 9 are also false. However, a false Equation 8 or 9 does not imply that Equation 10 (or 11) is necessarily false. That is, a significant test statistic may reflect differences in other distributional characteristics (i.e., variance or shape) rather than differences in location (Serlin & Harwell, 2001), unless additional distributional assumptions are met. Also, a null hypothesis for the interaction implies that differences in the location parameters for any two groups ($j = 1 \ldots J$) are equal for all K measures.

Furthermore, if Equation 10 is true so is Equation 11; however, if Equation 11 is true, it does not imply that Equation 10 is true. Likewise, a false Equation 10 does not imply a false Equation 11. These distinctions are important because in order to test a null hypothesis of shifts in location parameters analogous to the null hypothesis in Equation 2, the univariate null model for ranks (Equation 10) 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 (Equation 11) only requires an assumption that the error distribution for each of the *K* repeated measures is 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).

Summary and Research Purpose

There is no consensus concerning the analysis of data from split-plot designs when both the sphericity and normality assumptions are violated. A simulation study compared the performance of the univariate *F*-ratio from model 1, the Lecoutre (1991) ε-adjusted *F* (Equation 3), and statistics (Equations 5 and 6) from a multivariate approach to repeated measures for testing the interaction hypothesis concerning *location parameters* in a split-plot design under several conditions. These procedures were applied to Rank Transformed data (R_{ijk}) and the original data without any transformation (Y_{ijk}). Because tests for interactions applied to the Rank Transform scores have been noted to perform poorly when main effects are present in the original data (e.g., Blair et al., 1987), these tests performed on the Aligned Ranks (A_{ijk} from Equation 7) proposed by Higgins and Tashtoush (1994) were also investigated.

Method

Design

A 2 (sample size: $n_j = 10$ and 30) × 3 (covariance structure: independent, correlated spherical, and correlated non-spherical) × 3 (shape of error distribution: Normal, Double Exponential, and Exponential) × 3 (degree of a main effect: $c = 0$, 0.25, and 0.50) factorial design was employed for this simulation study. For each of these 54 conditions, 10,000 replications were generated using SAS/IML 6.12 (SAS Institute, 1996). Comparisons were made among 12 procedures for testing the interaction effect in a $J = 3 \times K = 4$ split-plot design at the $\alpha = 0.05$ significance level. For the original data (Y_{ijk}), the Rank Transformed data (R_{ijk}), and the Aligned Ranks (A_{ijk}), the following four statistics were calculated: (a) the conventional F-test; (b) the Lecoutre (1991) ε-adjusted F from Equation 3; (c) the F approximate test (Equation 6) for the Hotelling-Lawley trace (H) from a multivariate profile analysis; and (d) H (Equation 5) using a critical value from the Hotelling-Lawley trace distribution. Since the Rank Transform procedure proposed by Agresti and Pendergast (1986) was based on Hotelling's (1931) T^2, it seemed reasonable to calculate the Hotelling-Lawley trace on the original data and Aligned Ranks for the purposes of consistency. For a $J = 3 \times K = 4$ split-plot design, the parameters for the Hotelling-Lawley trace distribution are $s = 2$, $m = 0$, $n = 11.5$ for $n_j = 10$, and $n = 41.5$ for $n_j = 30$. Therefore, the $\alpha = .05$ critical values for H are 0.587 and 0.155 for $n_j = 10$ and 30, respectively.

The $n_j = 10$ condition was chosen because it has been used in other 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 total sample sizes of $N = 30$.

The Double Exponential distribution was chosen as a condition where the errors were symmetric but heavy-tailed with skewness and kurtosis values of $\gamma_1 = 0$ and $\gamma_2 = 3$, respectively. The Exponential distribution was selected as a condition where the errors were skewed ($\gamma_1 = 2$) and extremely heavy-tailed ($\gamma_2 = 6$). 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 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 ($\gamma_1 = 1.75$; $\gamma_2 = 5.90$) 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).

Simulation Procedures and Conditions

Using the SAS/IML RANNOR function, a (n_j = 10 or 30) 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 $K \times K$ 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 $\mathbf{\Sigma}_j$:

$$\mathbf{Y}_j' = \mathbf{F}\,\mathbf{X}_j' \tag{12}$$

(Beasley, 1994; Kaiser & Dickman, 1962). Because only constants were added later to create fixed effects (i.e., main effects, interactions), the values of $\mathbf{Y}_j$ are the error components.

In the first covariance structure condition, the repeated measures were independent. That is, the expected value for all pairwise correlations was zero ($\rho = 0$), and thus, $\mathbf{\Sigma}_j$ and $\mathbf{F}$ were identity matrices. The results of this condition were expected to be similar to what would happen in a between-subjects factorial design, a replication of Blair et al. (1987).

In the second condition, all population correlations between measures (i.e., off-diagonal elements of $\mathbf{\Sigma}_j$) were $\rho = 0.60$. This condition yielded results for a spherical covariance structure ($\varepsilon = 1$) in which case the univariate F-tests should not inflate Type I error rates. In the third condition, covariance structures with $\varepsilon = 0.64$ were imposed. The pairwise inter-correlations were ρ_{12} and $\rho_{34} = 0.70$ 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 Equation 12. 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 ($\boldsymbol{\Sigma}_j$) 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 structures were spherical, the random error components (ζ_{ijk}) were IID(0, σ^2_ζ) for each of the *JK* cells, which permitted an investigation of the 12 statistics as tests of interaction in terms of a univariate shift model for location parameters (Equation 10). Under condition in which the covariance structures were not spherical, however, only the less restrictive multivariate was valid assumption (i.e., IID[$\mathbf{0}_{(K-1)}$,$\mathbf{C}_K\boldsymbol{\Sigma}\mathbf{C}_K'$]), thus creating a violation of the underlying assumptions for the univariate parametric *F*-tests.

Using a balanced $J = 3 \times K = 4$ split-plot design from model 1, a repeated measures main effect pattern resulting in no interaction was imposed (see Blair et al., 1987, p. 1143). Specifically for group 1, a vector of constants, $\mathbf{c}_1 = [0\ 0\ 2c\ 0]$, was added to each observation for the $K = 4$ repeated measures. For group 2, $\mathbf{c}_2 = [-c\ -c\ c\ -c]$, and for group 3, $\mathbf{c}_3 = [-2c\ -2c\ 0\ -2c]$. Consistent with Blair et al. (1987), three values of *c* were used: $c = 0$, 0.25, and 0.50. For $c = 0$, both the repeated measures main effect and interaction effect null hypotheses were true. For all other values of *c*, a repeated measures main effect of $[-c\ -c\ c\ -c]$ was present in terms of location parameters, but there was no interaction nor any other distributional differences. Blair et al. (1987) demonstrated that when data with this pattern are ranked without alignment, the ranks (R_{ijk}) exhibit a non-zero interaction effect when $c \neq 0$. Correlation among the repeated measures was expected to exacerbate this problem. Again, when the correlation structure was independent, the results of this study were expected to be similar to those of Blair et al. (1987).

Results

For all tables, *F* refers to the univariate ANOVA *F*-test for model 1, F_ε refers to the Lecoutre (1991) ε-adjusted *F* (Equation 3), F_{H} refers to the *F* approximation (Equation 6) for the Hotelling-Lawley trace (Equation 5), and *H* refers to testing the Hotelling-Lawley trace (Equation 5) with a critical value from its referent distribution. Subscripts of *Y*, *R*, and *A* refer to the tests performed on the Original Data (Y_{ijk}), Rank Transform scores (R_{ijk}), and

Aligned Ranks (A_{ijk}) from Equation 9, respectively. The results for the condition in which the K = 4 repeated measures were generated independently is denoted by $\rho = 0$; $\varepsilon = 1.00$ refers to the condition where the repeated measures were equicorrelated ($\rho = 0.60$) and thus spherical; and $\varepsilon = 0.64$ refers to the non-spherical condition.

Table 1
Type I Error Rates for the Interaction Tests in the Absence of a Repeated Measures Main Effect ($c = 0$)

	Normal			Double Exponential			Exponential		
$n_j = 10$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	.0511	.0527	.0829 *	.0480	.0483	.0817 *	.0472	.0453 +	.0746 *
$F_{\varepsilon(Y)}$	.0383+	.0418 +	.0523	.0351 +	.0359 +	.0475	.0324 +	.0310 +	.0451 +
$F_{H(Y)}$	.0535	.0557 *	.0559 *	.0492	.0535	.0496	.0426 +	.0402 +	.0421 +
$H_{(Y)}$	.0483	.0515	.0508	.0450 +	.0482	.0444 +	.0386 +	.0368 +	.0385 +
$F_{(R)}$	.0514	.0523	.0765 *	.0507	.0534	.0803 *	.0485	.0468	.0797 *
$F_{\varepsilon(R)}$	.0500	.0504	.0570 *	.0496	.0524	.0539	.0470	.0447	.0564 *
$F_{H(R)}$	.0576*	.0546 *	.0546 *	.0552 *	.0575 *	.0548 *	.0525	.0533	.0532
$H_{(R)}$	.0520	.0506	.0499	.0508	.0530	.0499	.0478	.0488	.0487
$F_{(A)}$	.0525	.0542	.0756 *	.0507	.0532	.0755 *	.0518	.0499	.0745 *
$F_{\varepsilon(A)}$	.0513	.0531	.0579 *	.0498	.0521	.0572 *	.0501	.0481	.0569 *
$F_{H(A)}$	.0558*	.0595 *	.0564 *	.0554 *	.0560 *	.0547 *	.0531	.0529	.0552 *
$H_{(A)}$	.0522	.0542	.0535	.0515	.0544	.0506	.0476	.0494	.0499
$n_j = 30$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	.0501	.0487	.0800 *	.0489	.0528	.0777 *	.0466	.0472	.0751 *
$F_{\varepsilon(Y)}$	.0456	.0465	.0499	.0429+	.0469	.0480	.0399 +	.0416 +	.0464
$F_{H(Y)}$	.0526	.0493	.0510	.0495	.0558 *	.0490	.0453 +	.0460	.0436 +
$H_{(Y)}$	.0507	.0473	.0497	.0476	.0533	.0480	.0441 +	.0444 +	.0423 +
$F_{(R)}$	.0487	.0485	.0773 *	.0497	.0529	.0776 *	.0490	.0474	.0810 *
$F_{\varepsilon(R)}$	.0483	.0479	.0512	.0493	.0524	.0517	.0485	.0468	.0542
$F_{H(R)}$	.0518	.0482	.0516	.0529	.0514	.0522	.0496	.0478	.0496
$H_{(R)}$	.0497	.0468	.0500	.0513	.0491	.0502	.0480	.0454 +	.0472
$F_{(A)}$	.0487	.0489	.0730 *	.0490	.0540	.0712 *	.0468	.0500	.0728 *
$F_{\varepsilon(A)}$	.0485	.0498	.0515	.0487	.0538	.0505	.0464	.0489	.0530
$F_{H(A)}$	.0527	.0510	.0532	.0518	.0557 *	.0508	.0482	.0496	.0503
$H_{(A)}$	.0503	.0498	.0515	.0494	.0539	.0494	.0468	.0477	.0488

Note. * refers to liberal Type I error rates. + refers to conservative Type I error rates.

Type I Error Rates

For this study, tests that demonstrated a Type I error rate considerably 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 $\alpha = 0.05$ 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 2 standard errors above 0.05 (i.e., 0.0544) was considered "significantly liberal." An asterisk (*) is used to denote empirical values that were significantly above the nominal alpha (i.e., liberal). This is consistent with Bradley's (1978) stringent 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) and is denoted with a plus (+).

Consistent with Toothaker and Newman (1994) and Wilcox (1993), the effects of violating the normality assumption were a "dampening" of empirical alpha rates with small samples. For example, when the univariate F, the Lecoutre *df*-correction procedure, and the multivariate tests were performed on the original data with non-normal error distributions (Y_{ijk}) for $n_j = 10$, the Type I error rates were below Bradley's (1978) stringent criterion for a nominal alpha of 0.05, especially for data with Exponential error distributions. The general effects of non-sphericity on the univariate F-test whether performed on Y_{ijk}, R_{ijk}, or A_{ijk} are also evident in all of these results. That is, when the covariance structure was not spherical ($\varepsilon = 0.64$), univariate F-tests demonstrated drastically inflated Type I error rates regardless of sample size, shape of the error distribution, or whether the data were ranked.

The Lecoutre (1991) ε-adjusted F-test (F_ε) performed on data with Exponential error distributions was somewhat conservative, especially with $n_j = 10$. When F_ε was performed on the Aligned Ranks of data with Exponential error distributions, however, Type I error rates were more consistent with the nominal alpha even with the smaller sample size. In most other conditions, excluding the Rank Transformed scores in the presence of a repeated measures main effect (see Tables 2 and 3), F_ε performed on the original data and the Aligned Ranks held the Type I error rate near the nominal alpha of 0.05 with departures from normality and sphericity.

The multivariate approach using F_{H} (Equation 6) was liberal, especially for the Rank Transform scores and Aligned Ranks with $n_j = 10$ (see Table 1). However, these results may be a function of sample size in that the empirical Type I error rates were more consistent with the nominal alpha of 0.05 when sample size was increased to $n_j = 30$. Also for $n_j = 10$, testing H (Equation

5) with an exact critical value was generally effective in controlling Type I errors as compared to F_H. The rejections for these two multivariate tests were similar with $n_j = 30$, although testing H with an exact critical value was slightly more conservative in general (see Tables 1, 2, and 3).

Table 2

Type I Error Rates for the Interaction Tests in the Presence of a Repeated Measures Main Effect (c = 0.25)

	Normal			Double Exponential			Exponential		
$n_j = 10$	ρ = 0	ε = 1.00	ε = 0.64	ρ = 0	ε = 1.00	ε = 0.64	ρ = 0	ε = 1.00	ε = 0.64
$F_{(Y)}$	.0510	.0490	.0774 *	.0480	.0482	.0805 *	.0455 +	.0445 +	.0743 *
$F_{\varepsilon(Y)}$	.0371+	.0383 +	.0487	.0348 +	.0409 +	.0486	.0329 +	.0298 +	.0421 +
$F_{H(Y)}$	.0574*	.0496	.0559 *	.0482	.0510	.0515	.0455 +	.0396 +	.0404 +
$H_{(Y)}$	.0526	.0448 +	.0510	.0432 +	.0463	.0472	.0412 +	.0350 +	.0368 +
$F_{(R)}$	.0517	.0472	.0726 *	.0505	.0525	.0792 *	.0558 *	.0565 *	.0883 *
$F_{\varepsilon(R)}$	.0500	.0459	.0530	.0490	.0517	.0575 *	.0541	.0528	.0603 *
$F_{H(R)}$	.0586*	.0495	.0522	.0564 *	.0552 *	.0550 *	.0599 *	.0572 *	.0748 *
$H_{(R)}$	.0548*	.0450	.0473	.0494	.0510	.0506	.0549 *	.0551 *	.0690 *
$F_{(A)}$	.0515	.0485	.0735 *	.0491	.0507	.0776 *	.0515	.0500	.0712 *
$F_{\varepsilon(A)}$	.0504	.0478	.0550 *	.0475	.0512	.0581 *	.0502	.0478	.0536
$F_{H(A)}$	.0611*	.0517	.0571 *	.0540	.0540	.0548 *	.0547 *	.0527	.0540
$H_{(A)}$	.0537	.0477	.0532	.0493	.0518	.0506	.0500	.0487	.0504
$n_j = 30$	ρ = 0	ε = 1.00	ε = 0.64	ρ = 0	ε = 1.00	ε = 0.64	ρ = 0	ε = 1.00	ε = 0.64
$F_{(Y)}$	.0490	.0528	.0794 *	.0501	.0484	.0771 *	.0502	.0515	.0761 *
$F_{\varepsilon(Y)}$	.0457	.0506	.0527	.0469	.0448 +	.0480	.0445 +	.0469	.0478
$F_{H(Y)}$	.0505	.0546 *	.0546 *	.0505	.0489	.0476	.0471	.0487	.0478
$H_{(Y)}$	.0479	.0532	.0531	.0487	.0472	.0461	.0458	.0468	.0452 +
$F_{(R)}$	.0497	.0545 *	.0781 *	.0527	.0516	.0748 *	.0595 *	.0646 *	.0922 *
$F_{\varepsilon(R)}$	.0491	.0547 *	.0529	.0527	.0516	.0514	.0574 *	.0630 *	.0625 *
$F_{H(R)}$	.0501	.0569 *	.0532	.0546 *	.0545 *	.0551 *	.0647 *	.0715 *	.0970 *
$H_{(R)}$	.0490	.0550 *	.0515	.0522	.0518	.0529	.0631 *	.0692 *	.0947 *
$F_{(A)}$	.0490	.0527	.0745 *	.0502	.0472	.0729 *	.0532	.0521	.0733 *
$F_{\varepsilon(A)}$	.0488	.0528	.0531	.0506	.0471	.0519	.0523	.0517	.0530
$F_{H(A)}$	.0505	.0544	.0530	.0524	.0498	.0528	.0532	.0540	.0525
$H_{(A)}$	.0486	.0531	.0514	.0503	.0484	.0514	.0514	.0518	.0511

Note. * refers to liberal Type I error rates. + refers to conservative Type I error rates.

Table 3
Type I Error Rates for the Interaction Tests in the Presence of a Repeated Measures Main Effect ($c = 0.50$)

	Normal			Double Exponential			Exponential		
$n_j = 10$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	.0506	.0527	.0808 *	.0454 +	.0486	.0746 *	.0454 +	.0421 +	.0759 *
$F_{\varepsilon(Y)}$	.0379+	.0468	.0510	.0349 +	.0362 +	.0467	.0316 +	.0289 +	.0437 +
$F_{\mathrm{H}(Y)}$	.0544	.0553 *	.0543	.0511	.0517	.0523	.0476	.0384 +	.0441 +
$H_{(Y)}$	.0498	.0510	.0496	.0467	.0485	.0487	.0441 +	.0347 +	.0402 +
$F_{(R)}$	.0502	.0579 *	.0757 *	.0550 *	.0650 *	.0854 *	.0689 *	.0907 *	.1232 *
$F_{\varepsilon(R)}$	.0487	.0573 *	.0558 *	.0521	.0635 *	.0656 *	.0629 *	.0862 *	.0854 *
$F_{\mathrm{H}(R)}$	.0562*	.0617 *	.0640 *	.0628 *	.0656 *	.0883 *	.0827 *	.1073 *	.1640 *
$H_{(R)}$	.0507	.0583 *	.0581 *	.0582 *	.0621 *	.0800 *	.0767 *	.0990 *	.1548 *
$F_{(A)}$	.0498	.0545 *	.0746 *	.0518	.0527	.0732 *	.0539	.0475	.0743 *
$F_{\varepsilon(A)}$	.0488	.0557 *	.0558 *	.0498	.0513	.0550 *	.0523	.0455 +	.0559 *
$F_{\mathrm{H}(A)}$	.0571*	.0603 *	.0586 *	.0573 *	.0598 *	.0584 *	.0587 *	.0504	.0531
$H_{(A)}$	.0512	.0553 *	.0543	.0522	.0541	.0534	.0546 *	.0449 +	.0472
$n_j = 30$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	.0497	.0499	.0775 *	.0509	.0459	.0811 *	.0478	.0480	.0766 *
$F_{\varepsilon(Y)}$	.0453+	.0470	.0489	.0462	.0414 +	.0510	.0420 +	.0414 +	.0482
$F_{\mathrm{H}(Y)}$	.0504	.0509	.0510	.0517	.0471	.0526	.0472	.0460	.0484
$H_{(Y)}$	.0487	.0492	.0497	.0499	.0451 +	.0506	.0449 +	.0443 +	.0469
$F_{(R)}$	.0548*	.0638 *	.0870 *	.0690 *	.1007 *	.1226 *	.0942 *	.1899 *	.2097 *
$F_{\varepsilon(R)}$	.0544	.0634 *	.0602 *	.0678 *	.0995 *	.0857 *	.0903 *	.1860 *	.1478 *
$F_{\mathrm{H}(R)}$	.0577*	.0664 *	.0868 *	.0723 *	.1049 *	.1626 *	.1306 *	.2357 *	.4166 *
$H_{(R)}$	.0558*	.0640 *	.0842 *	.0690 *	.1013 *	.1587 *	.1270 *	.2313 *	.4109 *
$F_{(A)}$	.0506	.0498	.0712 *	.0515	.0477	.0733 *	.0503	.0501	.0716 *
$F_{\varepsilon(A)}$	.0504	.0499	.0510	.0513	.0474	.0514	.0496	.0498	.0522
$F_{\mathrm{H}(A)}$	.0507	.0510	.0526	.0541	.0485	.0543	.0500	.0513	.0507
$H_{(A)}$	.0487	.0494	.0508	.0519	.0469	.0522	.0479	.0503	.0491

Note. * refers to liberal Type I error rates. + refers to conservative Type I error rates.

In Tables 2 and 3, it is evident that tests for interactions performed on Rank Transform scores were inadequate when main effects were present. For example, the empirical Type I error rates for the interaction tests were well above Bradley's stringent criterion for a nominal alpha when a large main effect (c = 0.50) was present (see Table 3). This indicates that the Rank Transform procedure imposes an interaction effect in the expected values of the ranked data when main effects are present in the original data even in the absence of an interaction effect (Blair et al., 1987). Furthermore, this implies that tests for interactions performed on unaligned ranks (R_{ijk}) are sensitive to between-group distributional differences (see Equations 8 and 9) in the form of location parameters, even if these location differences are the same across the K repeated measures and thus not indicative of an interaction.

The empirical Type I error rates found in this study were extremely similar to those reported for between-subjects designs (see Blair et al., 1987, p. 1138). This problem with Rank Transform scores worsened under conditions where the repeated measures were correlated (i.e., $\varepsilon = 1.00$; $\varepsilon = 0.64$). That is, the empirical Type I error rates were considerably higher for tests performed on the Rank Transformed scores as compared to the results for the uncorrelated covariance structure condition ($\rho = 0$). The Type I error rate inflation for tests performed on Rank Transformed scores also worsened with skewed (i.e., Exponential) error distributions and a larger sample size $n_j = 30$. By contrast, tests for the Aligned Ranks generally maintained the expected Type I error rate in the presence of a strong repeated measures main effect (see Table 3). The only problem exhibited was that the F_{H} (Equation 6) performed on the Aligned Ranks inflated the Type I error rates with a small sample size ($n_j = 10$). Again, testing H with a critical value from the Hotelling-Lawley trace distribution was more effective in controlling Type I errors with the smaller sample size of $n_j = 10$. But with a sample size of $n_j = 30$, both multivariate tests for Aligned Ranks held the Type I error rates near 0.05.

Power Comparison

Because many of the tests exhibited conservative empirical Type I error rates when data with Exponential and Double Exponential error distributions were analyzed, an additional simulation study with 10,000 replications per condition was conducted to investigate whether any of these procedures would demonstrate an advantage in statistical power. The sample size, shape of the error distributions, and covariance structure conditions previously described in the Method section were used. To simulate an interaction effect among the location parameters, a vector of $\mathbf{\Delta}_1 = \mathbf{\Delta}_2 = [-\delta\ -\delta\ \delta\ -\delta]$ was added to each $1 \times K$ observation in groups $j = 1$ and $j = 2$, respectively. The third group was not transformed; thus, $\mathbf{\Delta}_3 = [0\ 0\ 0\ 0]$. A value of $\delta = 0.375$ was chosen because it created of an interaction effect while also imposing a repeated measures main effect that was equivalent to the $c = 0.25$ Type I error rate condition previously reported in Table 2. To investigate a situation with lower statistical power, a smaller interaction effect ($\delta = 0.25$) was also simulated. It should be noted that interactions can exist in the absence of main effects, but situations where both effects exist are more common in behavioral research.

Table 4 shows the rejection rates for several tests under conditions where there is an interaction effect in the location parameters of the original data (Y_{ijk}). Because variances and shapes of the error distributions were held constant across the $J = 3$ groups and $K = 4$ repeated measures, these values represent empirical estimates of statistical power in terms of location parameters rather other distributional differences. Because most of the tests performed on the Rank Transformed scores inflated the Type I error rate when a repeated measures main effect was present, these procedures were excluded from the power comparison. Results for F_{H} performed on the Aligned Ranks with $n_j = 10$ were also excluded because of liberal empirical alpha rates (see Tables 1-3). The rejection rates for all tests under the $\delta = 0.375$, $n_j = 30$ condition were near unity and thus not reported.

The results show that tests performed on the original scores (Y_{ijk}) with Normal error distributions demonstrated a slight power advantage over the rank-based tests. For spherical covariance structures, the univariate $F_{(Y)}$, was slightly more powerful than the multivariate tests. For the non-spherical covariance structure, the multivariate approach was more powerful than the univariate tests. For data with symmetric, heavy-tailed (i.e., Double Exponential) error distributions, tests performed on the Aligned Ranks exhibited a slight power advantage over parametric procedures. Again, for spherical covariance structures, the univariate $F_{(A)}$ was slightly more powerful than multivariate tests performed on Aligned Ranks. For the non-spherical covariance structure, the multivariate approach was more powerful than the univariate tests. Thus, these results indicate that if the errors are identically distributed with symmetric shape there is no clear advantage to using Aligned Ranks over the original data, especially with cell sizes of $n_j = 30$ or more.

For data with skewed, heavy-tailed (i.e., Exponential) error distributions, however, there was a considerable advantage to using Aligned Ranks. For example. with the smaller sample size of $n_j = 10$, using a critical value from the Hotelling-Lawley Trace distribution for H (Equation 5) performed on the Aligned Ranks [$H_{(A)}$] was generally more powerful than the other tests

Table 4
Rejection Rates for the Interaction Tests in the Presence of a Repeated Measures Main Effect and a Interaction Effect

	Normal			Double Exponential			Exponential		
$\delta = 0.375$									
$n_j = 10$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	.1843	.4344	—	.2083	.4952	—	.1885	.4619	—
$F_{\varepsilon(Y)}$	.1557	.3867	.3851	.1704	.4371	.4491	.1482	.3992	.4297
$F_{\mathrm{H}(Y)}$	—	.3987	—	.2040	.4729	.9061	.1875	.4733	.8758
$H_{(Y)}$	.1676	.3809	.8498	.1933	.4577	.8972	.1768	.4546	.8670
$F_{(A)}$	.1776	.4181	—	.2331	.5295	—	.2570	.6270	—
$F_{\varepsilon(A)}$	.1750	.4145	—	.2300	.5245	—	.2522	.6200	.7421
$H_{(A)}$	.1709	.3922	.8362	.2232	.4975	.9128	.2574	.6011	.9269
$\delta = 0.250$									
$n_j = 10$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	.0960	.1978	—	.1090	.2223	—	.1074	.2146	—
$F_{\varepsilon(Y)}$	.0778	.1623	.1534	.0865	.1804	.1774	.0833	.1680	.1597
$F_{\mathrm{H}(Y)}$	—	.1879	—	.1128	.2199	.5596	.1064	.2197	.5539
$H_{(Y)}$	.0890	.1778	.4549	.1049	.2084	.5436	.0996	.2060	.5388
$F_{(A)}$	.0952	.1906	—	.1226	.2397	—	.1426	.3185	—
$F_{\varepsilon(A)}$	.0941	.1870	—	.1208	.2354	—	.1383	.3132	.3673
$H_{(A)}$	.0957	.1882	.4454	.1249	.2325	.5643	.1380	.3203	.6622
$\delta = 0.250$									
$n_j = 30$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\rho = 0$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	.2469	.5968	—	.2803	.6489	—	.2565	.6033	—
$F_{\varepsilon(Y)}$	.2318	.5789	.5954	.2644	.6306	.6802	.2389	.5788	.6081
$F_{\mathrm{H}(Y)}$	.2417	—	—	.2814	.6444	.9822	.2611	.6067	.9676
$H_{(Y)}$	.2364	.5767	.9725	.2762	.6401	.9816	.2563	.6008	.9670
$F_{(A)}$	.2421	.5780	—	.3231	.6960	—	.3715	.8307	—
$F_{\varepsilon(A)}$	.2415	.5766	.7205	.3223	.6954	.8641	.3696	.8300	.9504
$F_{\mathrm{H}(A)}$	.2404	.5743	.9648	.3213	.6927	.9893	.3744	.8253	.9966
$H_{(A)}$	.2355	.5688	.9631	.3163	.6885	.9887	.3692	.8219	.9965

Note. Rejection Rates for Tests with Type I error rates above the nominal alpha of 0.05 are omitted.

performed on data with skewed error distributions. A power advantage of $H_{(A)}$ over H performed on the original data with Exponential error distributions [$H_{(Y)}$] was evident with a smaller sample size of $n_j = 10$. For example, with the non-spherical covariance structure ($\varepsilon = 0.64$) and smaller effect size ($\delta = 0.25$), the $H_{(A)}$ exhibited an empirical power estimate of approximately 66% rejection; whereas, $H_{(Y)}$ exhibited a lower empirical power estimate of approximately 54% rejection at $\alpha = 0.05$.

This power advantage of $H_{(A)}$ over $H_{(Y)}$ was also evident with a larger sample size of $n_j = 30$ for both independent ($\rho = 0$) and correlated spherical covariance structures ($\varepsilon = 1.00$). For example, with $n_j = 30$ and the spherical covariance structure ($\varepsilon = 1.00$), the $H_{(A)}$ exhibited an empirical power estimate of approximately 82% rejection. $H_{(Y)}$ performed on the original data with Exponential error distributions exhibited a much lower empirical power estimate of approximately 60% rejection. With a non-spherical covariance structure ($\varepsilon = 0.64$) and $n_j = 30$, however, the distinct advantage of $H_{(A)}$ over $H_{(Y)}$ diminished to some extent.

The advantage of using the Aligned Ranks can also be inferred from the results of the *df*-correction procedure (i.e., F_ε). For example, with $n_j = 30$ and the non-spherical covariance structure ($\varepsilon = 0.64$), the F_ε performed on the Aligned Ranks exhibited empirical power estimates of approximately 95% rejection. F_ε performed on the original data with Exponential error distributions exhibited much lower empirical power estimates of approximately 61% rejection (see Table 4). However, this may be attributed to the ranking process reducing the degree of non-sphericity. That is, Aligned Ranks, although inheriting some of the non-sphericity present in the original data, did have smaller departures from sphericity with higher estimates of ε and thus larger *df*s.

Multiple Comparison Procedures for Aligned Ranks

Given that the Aligned Ranks procedure is a viable approach to analyzing repeated measures data, then contrast procedures based on this method should hold quite generally (Agresti & Pendergast, 1986). The most typical form of a contrast used by behavioral researchers is a product interaction contrast (Hochberg & Tamhane, 1987, pp. 294-303; Marascuilo & Levin, 1970) defined as:

$$
\begin{aligned}
\hat{\psi} = {} & a_1(b_1\overline{A}_{11} + b_2\overline{A}_{12} + \ldots + b_k\overline{A}_{1k} + \ldots + b_K\overline{A}_{1K}) \\
& + a_2(b_1\overline{A}_{21} + b_2\overline{A}_{22} + \ldots + b_k\overline{A}_{2k} + \ldots + b_K\overline{A}_{2K}) \\
& \ldots + a_j(b_1\overline{A}_{j1} + b_2\overline{A}_{j2} + \ldots + b_k\overline{A}_{jk} + \ldots + b_K\overline{A}_{jK}) \\
& \ldots + a_J(b_1\overline{A}_{J1} + b_2\overline{A}_{J2} + \ldots + b_k\overline{A}_{Jk} + \ldots + b_K\overline{A}_{JK});
\end{aligned}
\tag{13}
$$

where $\overline{A}_{jk}$ is the mean Aligned Rank from Equation 9 for the j^{th} group on the k^{th} repeated measure, $(a_1 + a_2 + \ldots + a_j \ldots + a_J)'$ is a vector of contrast coefficients, **a**, that compares the J independent samples, and $(b_1 + b_2 + \ldots + b_k + \ldots + b_K)'$ is a vector of contrast coefficients, **b**, that involves the K repeated measures with the restriction that $\Sigma a_j = 0$ and $\Sigma b_k = 0$. For comparing the J independent groups, a set of pairwise or group combination contrasts would most likely be of interest for defining **a**. For comparing the K repeated measures either pairwise, polynomial, or trend contrasts would most typically define **b** (Lix & Keselman, 1996; Marascuilo & McSweeney, 1967).

From a univariate perspective, a pooled squared standard error of a contrast in a split-plot design (see Kirk 1982, pp. 516-518) can be calculated by defining:

$$A^*_{ij} = \sum_{k=1}^{K} b_k A_{ijk}\,, \tag{14}$$

$$SS_{A^*_j} = \sum_{i=1}^{n_j} (A^*_{ij} - \overline{A}^*_j)^2 \text{, and} \tag{15}$$

$$SE^2_{\hat{\psi}} = \sum_{j=1}^{J} \left(\frac{a_j^2}{n_j}\right) \frac{SS_{A^*_j}}{(N-J)}. \tag{16}$$

where $\overline{A}^*_j$ is the mean for the j^{th} group for the transformed score A^*_{ij} in Equation 14.

From a multivariate perspective, the covariance matrices are not pooled and all elements of **E** (Equation 4) based on Aligned Ranks (Equation 9) are used. Thus, the squared standard error of a contrast defined in Equation 13 can be calculated by:

$$SE^2_{\hat{\psi}} = \sum_{j=1}^{J} \left(\frac{a_j^2}{n_j}\right) \frac{(\mathbf{b'Eb})}{(N-J)}. \tag{17}$$

Squared standard errors may also be computed from $\mathbf{E}^*$ such that

$$SE^2_{\hat{\psi}} = \frac{(\mathbf{m'E^*m})}{(N-J)} \tag{18}$$

where **m** is a JK column vector of interaction contrasts defined in Equation 13 so that $\mathbf{m}' = \{a_1\mathbf{b}', \ldots, a_j\mathbf{b}', \ldots, a_J\mathbf{b}'\}$. For example, in a $J = 3$ by $K = 4$ design, imagine that a researcher tests whether the linear trend, $\mathbf{b}' = \{-3\ -1\ 1\ 3\}$, of the first group is different from the linear trend of the other two groups combined, $\mathbf{a}' = \{2\ -1\ -1\}$, and thus, $\mathbf{m}' = \{-6\ -2\ 2\ 6\ 3\ 1\ -1\ -3\ 3\ 1\ -1\ -3\}$.

A (1 - α)% confidence interval for the contrast of Aligned Ranks can be formed by:

$$\hat{\psi} \pm S(SE_{\hat{\psi}}), \tag{19}$$

The null hypothesis H_0: $\psi = 0$ is rejected if the confidence interval in Equation 19 does not cover zero. From the univariate approach, $SE_{\hat{\psi}}$ is defined as the square root of (Equation 16). $SE_{\hat{\psi}}$ is defined as the square root of Equation 17 or 18 for the multivariate approach. The definition of S depends on the type of contrast is conducted. For example, to construct a post hoc Scheffé-type confidence interval, S would be defined as:

$$S = \sqrt{df_h F_{\alpha;df_h,df_e}} \tag{20}$$

Constructing Scheffé-type confidence intervals is not usually suggested because it is a generally conservative procedure; however, Klockars and Hancock (2000) have proposed a more powerful modification based on Scheffé (1970) that could be applied. As an alternative approach that also could yield more statistical power, define S as:

$$S = t_{(1-\alpha_{\text{DS}}),df_e}, \tag{21}$$

a critical value from Student's t distribution using the Dunn-Sidák correction, $\alpha_{\text{DS}} = [1\text{-}(1 - \alpha)^{1/d}]/2$, for d contrasts. It should be noted that df_e for Equations 20 and 21 differs for the univariate and multivariate approaches with the univariate approach resulting in larger df_e. For defining S in terms of the sampling distribution of the Hotelling-Lawley trace, the reader is referred to Gabriel (1968) and Sheehan-Holt (1998).

It should also be noted that conducting post hoc analyses is not generally suggested as an optimal procedure to adopt (Marascuilo & Levin, 1970). Rather, a defined set of planned contrasts with an appropriate adjustment for controlling Type I errors is often recommended in which case the omnibus tests previously elaborated should be bypassed. For conducting multiple planned comparisons or simultaneous test procedures, there are several excellent references for both the univariate and multivariate approaches references (e.g., Hochberg & Tamhane, 1987;

Gabriel, 1968; Lix & Keselman, 1996; Maxwell & Delaney, 2000; Sheehan-Holt, 1998).

It is debatable whether the multivariate or univariate approach is better in terms of robustness and power (Maxwell & Delaney, 2000), and thus, this issue should be investigated. However, the multivariate approach would be expected to yield more precise confidence intervals than the univariate approach, especially in situations where the pooled covariance matrix of Aligned Ranks is non-spherical (Boik, 1981).

Discussion

Despite the perspective that good nonparametric tests for interactions do not exist (e.g., Sawilowsky, 1990), this study showed that the proposed multivariate tests of interactions among location parameters using Aligned Ranks in a split-plot design controlled Type I error rates for data with non-spherical and non-normal error distributions. Furthermore, these multivariate procedures, as well as other univariate tests for Aligned Ranks maintained the expected Type I error rate in the presence of a strong repeated measures main effect, whereas tests performed on Rank Transformed scores demonstrated severely inflated Type I error rates. Thus, for testing interactions in a split-plot design, aligning the data before ranking is essential when main effects are present.

Tests performed on the Aligned Ranks also demonstrated more statistical power than parametric tests performed on the original data with non-normal errors, especially with skewed, heavy-tailed (i.e., Exponential) error distributions. The multivariate approach was most powerful; however, multivariate tests have strict sample size requirements based on the number of repeated measures. Thus, with smaller samples sizes (e.g., $N = 30$), using a critical value from the referent distribution of the multivariate test is suggested because this method (a) consistently controlled Type I errors (see Tables 1-3) and (b) demonstrated superior power (see Table 4). For larger sample sizes (e.g., $N \geq 90$), the Hotelling-Lawley trace and its F approximation showed similar rejection rates, and thus, either test could be employed. Unfortunately, few multivariate texts have extensive tables of critical values for multivariate statistics, and thus, F approximations (Equation 6) may be employed out of necessity.

Although the Aligned Rank procedure for testing interactions performed better than the Rank Transform method in the presence of main effects, it is important to reiterate that the Aligned Rank method also involves a monotone transformation of the original data. Therefore, issues concerning the interpretation of rank-based tests are of concern. Namely, multivariate procedures performed on Aligned Ranks test a null hypothesis of *distributional equivalence* (Equation 9) across the J groups for each of the K measures.

Vargha and Delaney (1998) have noted that with $K = 1$ (a univariate one-way design which would not require alignment), the ANOVA performed on ranks [$F_{(R)}$] actually tests a null hypothesis of distributional equivalence or "stochastic homogeneity." A null hypothesis of stochastic homogeneity is equivalent to a null hypothesis of equal expected location parameters only under conditions where variances are homogeneous and distributional shapes are symmetric or identically asymmetric. Otherwise the null hypothesis of equal location parameters is not what is actually tested by rank-based statistics. Thus, when $K > 1$, the Hotelling-Lawley trace performed on Aligned Ranks also evaluates a null hypothesis of distributional equivalence. Therefore, the interaction null hypothesis in Equation 9 is equivalent to a null hypothesis of equal expected location parameters only under conditions where each of the J groups have a symmetric (or identically asymmetric) error distributions with homogeneous variances for each of the K repeated measures separately. Thus, if the error distributions of original data (Y_{ijk}) for any of the J groups differ substantially in terms of variance or shape on any of the K repeated measures, multivariate statistics performed on A_{ijk} should be considered tests of a null hypothesis of between-group distributional equivalence (Equation 9), rather than tests of location parameters only (Serlin & Harwell, 2001).

To elaborate, if the null hypothesis of distributional equivalence (9) is true, it does imply the absence of an interaction in that differences in the location parameters for any two groups ($j = 1 \ldots J$) are equal for all K measures. However, there are situations in which the location parameters (e.g., mean ranks) would not indicate an interaction, but the distributions would not be equivalent, and therefore, the interaction null hypothesis (Equation 9) would be false. This is important because there are situations where the interaction null hypothesis in Equation 9 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 (see Agresti & Pendergast, 1986; Beasley, 2000; Serlin & Harwell, 2001; Vargha & Delaney, 1998). However, situations where distributional equivalence does not hold while location parameters are identical are rare. Furthermore, these test statistics have mean rank differences in the numerator of the formula (e.g., 5). Thus, both univariate and multivariate tests performed on Aligned Ranks would be considered especially sensitive to differences in location parameters (Lehmann, 1998; Marascuilo & McSweeney, 1977).

Future studies into tests for Aligned Ranks should investigate statistical issues that distinguish the multivariate tests from the univariate *df*-correction procedures (e.g., Huynh, 1978; Lecoutre, 1991). Specifically, the multivariate tests used in this study and the Lecoutre (1991) ε-adjusted F are known to be extremely sensitive to heterogeneity of variance. For other multivariate tests, several studies (e.g., Keselman & Keselman, 1990; Olson, 1974) have reported that, as compared to the Hotelling-Lawley trace, the Pillai-Bartlett trace is more robust to heterogeneous covariance matrices. Keselman et al. (1993) have suggested a multivariate Welch-James type statistic (Johansen, 1980) for situations in which the covariance matrices are heterogeneous. This statistic uses separate covariance matrices rather than pooling the covariance matrices over the *J* groups. For a univariate approach, Huynh (1978) developed the GA and IGA specifically for situations where the covariance matrices were non-spherical and heterogeneous.

Also, pairwise multiple comparison procedures that correct for heterogeneous variances (i.e., Games & Howell, 1976; Tamhane, 1979) or simultaneous pairwise multiple comparison procedures developed for repeated measures designs (Alberton & Hochberg, 1984; Keselman, 1994; Keselman, Keselman, & Shaffer, 1991) could be applied to Aligned Ranks. If tests of post hoc contrasts are desired, then the Brown and Forsythe (1974) procedure, which corrects for heterogeneous variances among complex contrasts, could be implemented with similar modifications. These suggestions are in agreement with recommendations to calculate a separate estimate of experimental error for each contrast (Boik, 1981). However, the statistical properties of other multivariate statistics (i.e., Wilks' lambda; Pillai-Bartlett trace; Welch-James test), univariate procedures (i.e., GA; IGA), and multiple comparison procedures performed on Aligned Ranks from split-plot designs have not been evaluated and thus should be investigated under conditions of non-normality, non-sphericity, and heterogeneous covariance matrices.

References

Agresti, A. & Pendergast, J. (1986). Comparing mean ranks for repeated measures data. *Communications in Statistics: Theory and 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.

Alberton, Y. & Hochberg, Y. (1984). Approximations for the distribution of a maximal pairwise *t* in some repeated measures designs. *Communications in Statistics*: *Theory and Method*, *13A*, 2847-2854.

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

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

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

Beasley, T. M. (2000). Nonparametric tests for analyzing interactions among intra-block ranks in multiple group repeated measures designs. *Journal of Educational and Behavioral Statistics*, *25*, 20-59.

Beasley, T. M. & Zumbo, B. D. (April, 1998). *Rank transformation and df-correction procedures for split-plot designs*. Paper presented at the meeting of the American Educational Research Association. San Diego, CA.

Blair, R. C., Higgins, J. J., Karniski, W., & Kromrey, J. D. (1994). A study of multivariate permutation tests which may replace Hotelling's 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: Simulation and Computation*, *16*(4), 1133-1145.

Boik, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika, 46*, 241-255.

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 and Statistical Psychology*, *31*, 144-152.

Brown, M. B. & Forsythe, A. B. (1974). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics*, *30*, 719-724.

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

Gabriel, K. R. (1968). Simultaneous test procedures in multivariate analysis of variance. *Biometrika*, *55*, 489-504.

Games, P. A. & Howell, J. F. (1976). Pairwise multiple comparisons with unequal *n*'s and/ or variances: A Monte Carlo study. *Journal of Educational Statistics*, *1*, 112-125.

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, 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: Computation and Simulation*, *29*(4), 1059-1088.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks.* New York: Wiley.

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

Hochberg, Y. & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: 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. *Proceedings of the Second Berkeley Symposium of Mathematical Statistics and 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. J*ournal 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. (1994). Stepwise and simultaneous multiple comparison procedures of repeated measures means. *Journal of Educational Statistics*, *19*, 127-162.

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

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

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

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

Keselman, H. J., Keselman, J. C., & Shaffer, J. P. (1991). Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity. *Psychological Bulletin*, *111*, 162-170.

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

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont CA: Brooks-Cole.

Klockars, A. J. & Hancock, G. R. (2000). Scheffé's more powerful *F*-protected post hoc procedure. *Journal of Educational and Behavioral Statistics*, *25*, 13-19.

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.

Lix, L. M. & Keselman, H. J. (1996). Interaction contrasts in repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, *49*, 147-162.

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.

Marascuilo, L. A. & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs. The elimination of Type IV errors. *American Educational Research Journal*, *7*, 397-421.

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

Marascuilo, L. A. & McSweeney, M. (1967). Nonparametric and post hoc comparisons for trend. *Psychological Bulletin*, *67*, 401-412.

Maxwell, S. E. & Delaney, H. D. (2000). *Designing experiments and analyzing data: A model comparison perspective*. Mahwah, NJ: Erlbaum.

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.

Robey, R. R. & Barcikowski, R. S. (April, 1995). *The value of* ε *estimates as descriptive statistics*. Paper presented at the meeting of the American Educational Research Association. San Francisco, CA.

SAS Institute. (1996). *SAS/IML user's guide* (Release 6.12). Cary, NC: Author.

Sawilowsky, S. S. (1990). Nonparametric test of interaction in experimental design. *Review of Educational Research*, *60*, 91-126.

Scheffé, H. (1970). Multiple testing versus multiple estimation. Improper confidence sets. Estimation of directions and ratios. *Annals of Mathematical Statistics*, *41*, 1-29.

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

Sheehan-Holt, J. (1998). MANOVA simultaneous test procedures: The power and robustness of restricted multivariate contrasts. *Educational and Psychological Measurement*, *58*, 861-881.

Tamhane, A. C. (1979). A comparison of procedures for multiple comparisons if means with unequal variances. *Journal of the American Statistical Association*, *74*, 471-480.

Thompson, G. L. (1991). A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association*, *86*, 410-419.

Toothaker, L. E. & Newman, D. (1994). A. Nonparametric competitors to the two way ANOVA. *Journal of Educational and 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 E. Edwards (Ed.), Applied analysis of variance in the behavioral sciences (pp. 345-374). New York: Marcel Dekker.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Zimmerman, D. & Zumbo, B. (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*(2), 139-149.

Accepted July, 2001.