



ELSEVIER

Computational Statistics & Data Analysis 42 (2003) 569–593

COMPUTATIONAL
STATISTICS
& DATA ANALYSIS

www.elsevier.com/locate/csda

Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs

T. Mark Beasley^{a,*}, Bruno D. Zumbo^b

^a*Department of Biostatistics, University of Alabama at Birmingham, Ryals Public Health Building
343C, 1665 University Boulevard, Birmingham, AL 35294, USA*

^b*Department of Educational and Counselling Psychology and Special Education, University of British
Columbia, 2125 Main Mall, Scarfe Building, Vancouver, BC, Canada V6T 1Z4*

Received 1 February 2002; received in revised form 1 June 2002

Abstract

Parametric methods are commonly used despite evidence that model assumptions are often violated. Various statistical procedures have been suggested for analyzing data from multiple-group repeated measures (i.e., split-plot) designs when parametric model assumptions are violated (e.g., Akritas and Arnold (J. Amer. Statist. Assoc. 89 (1994) 336); Brunner and Langer (Biometrical J. 42 (2000) 663)), including the use of Friedman ranks. The effects of Friedman ranking on data and the resultant test statistics for single sample repeated measures designs have been examined (e.g., Harwell and Serlin (Comput. Statist. Data Anal. 17 (1994) 35; Comm. Statist. Simulation Comput. 26 (1997) 605); Zimmerman and Zumbo (J. Experiment. Educ. 62 (1993) 75)). However, there have been fewer investigations concerning Friedman ranks applied to multiple groups of repeated measures data (e.g., Beasley (J. Educ. Behav. Statist. 25 (2000) 20); Rasmussen (British J. Math. Statist. Psych. 42 (1989) 91)). We investigate the use of Friedman ranks for testing the interaction in a split-plot design as a robust alternative to parametric procedures. We demonstrated that the presence of a repeated measures main effect may reduce the power of interaction tests performed on Friedman ranks. Aligning the data before applying Friedman ranks was shown to produce more statistical power than simply analyzing Friedman ranks. Results from a simulation study showed that aligning the data (i.e., removing main effects) before applying Friedman ranks and then performing either a univariate or multivariate test can provide more statistical power than parametric tests if the error distributions are skewed.

© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Aligned ranks; Friedman ranks; Split-plot design; Repeated measures; Interaction tests

* Corresponding author. Tel.: +1-205-975-4957; fax: +1-205-975-2540.

E-mail address: mbeasley@uab.edu (T.M. Beasley).

1. Introduction

Repeated measures designs involving two or more independent groups (i.e., split-plot designs) are among the most common experimental designs in a variety of research settings (e.g., Keselman et al., 1998; Koch et al., 1980). Various statistical procedures have been suggested for analyzing data from split-plot designs when parametric model assumptions are violated (e.g., Akritas and Arnold, 1994; Brunner and Langer, 2000). Our focus will be Friedman (1937) rank procedures for testing the interaction. The effects of Friedman ranking on data and the resultant test statistics for single sample repeated measures designs have been examined (e.g., Harwell and Serlin, 1994, 1997; Zimmerman and Zumbo, 1993). However, there have been fewer investigations concerning Friedman ranking in split-plot designs (e.g., Beasley, 2000; Rasmussen, 1989; Rasmussen et al., 1989) and how alignment procedures will affect the robustness and power of Friedman ranks.

Parametric methods are commonly used despite evidence that model assumptions are often violated. Therefore, we review univariate and multivariate approaches for parametric models and two Friedman rank methods for testing the interaction in split-plot designs. A simulation study was conducted to compare these parametric and non-parametric procedures under conditions violating model assumptions.

1.1. Parametric models for split-plot designs

1.1.1. Univariate approach

The univariate analysis of variance (ANOVA) approach to the split-plot design has the following linear model:

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

where j is referenced to the J groups of the between-subjects factor, i to the n_j subjects nested within the j th group, k to the K levels of the within-subjects (repeated measures) factor, ε_{ijk} is a random error vector, and $N = \sum n_j$ is the total number of subjects. The interaction of the between-subjects (i.e., independent grouping or treatment variable) and the within-subjects (i.e., repeated measures) factors is of interest in many applications (Boik, 1993; Koch et al., 1980). In educational experiments, the interaction typically represents differential gains in achievement for a treatment group. In psychological and developmental research, the interaction indicates that independent groups do not have parallel profiles or do not exhibit identical growth curves (Winer et al., 1991). In genetics experiments, the interaction typically indicates differential growth rates for organisms of different genotypes (Lynch and Walsh, 1998).

For the parametric F -ratio for testing the interaction ($F_{(Y)}$) from the univariate model (1), the random error components for each of the JK cells (ε_{ijk}) are assumed to be independent and sampled from identical normal distributions with a mean of zero and a common variance (i.e., $NID[0, \sigma_\varepsilon^2]$ for all j and k). For $K > 2$, the univariate $F_{(Y)}$ also requires an additional assumption concerning the sphericity of the pooled covariance matrix. If the pooled covariance matrix is non-spherical, the degrees-of-freedom (dfs)

are corrected by a factor epsilon and the resultant statistic, $F_{\epsilon(Y)}$, is valid. Methods for estimating epsilon have been investigated for over four decades (e.g., Box, 1954; Greenhouse and Geisser, 1959; Huynh and Feldt, 1970, 1976; Lecoutre, 1991). Also, general approximate methods to correct the *dfs* have been developed (Huynh, 1978). However, these *df*-correction procedures tend to be less powerful than multivariate approaches to analyzing repeated measures designs (e.g., Algina and Keselman, 1998; Algina and Oshima, 1994; Keselman and Algina, 1996) and thus will not be elaborated.

1.1.2. Multivariate approach

The multivariate approach to analyzing repeated measures designs (i.e., multivariate profile analysis) is often suggested because the multivariate tests do not require the additional sphericity assumption. This is of great concern for repeated measures (e.g., longitudinal) designs because it seems unreasonable to make assumptions about the consistency of covariances (i.e., correlational structure) among measures taken over an extended period of time (Koch et al., 1980). One approach to conducting the multivariate profile analysis is to take pairwise differences among the K repeated measures in order to compute $(K - 1)$ transformed scores, $\mathbf{Y}^* = \mathbf{YD}$, where \mathbf{Y} is the $N \times K$ data matrix of scores (Y_{ijk}) and \mathbf{D} is a $K \times (K - 1)$ difference matrix of the general form:

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix}.$$

These transformed scores are then submitted to a MANOVA with the following multivariate linear model:

$$\mathbf{Y}_j^* = \mathbf{M}_{**} + \mathbf{B}_j + \mathbf{E}_j; \tag{2}$$

where \mathbf{M}_{**} is a $(K - 1)$ vector of grand means (centroids), \mathbf{B}_j is a $(K - 1)$ vector of between-subjects effects, and \mathbf{E}_j is a random error matrix.

The interaction null hypothesis ($H_{0(J \times K)} : \mathbf{B}_1 = \dots = \mathbf{B}_j = \dots = \mathbf{B}_J$) can be tested with multivariate statistics such as the Hotelling–Lawley trace, $H_{(Y)}$. Based on Hotelling (1951), $H_{(Y)}$ can be transformed to an F approximation statistic by

$$F_{H(R)} = [2(sn + 1)/(s^2(2m + s + 1))]H_{(Y)}, \tag{3}$$

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 $df_e = [2(sn + 1)]$. When sample sizes are small, using a critical value for $H_{(Y)}$ from the sampling distribution of the Hotelling–Lawley trace using the s , m , and n parameters can provide more statistical power. Unfortunately, few multivariate texts have these critical values tabled.

The homogeneity of variance assumption underlying the multivariate approach to repeated measures data requires that the J covariance matrices (Σ_j) are equivalent so that they can be combined to form the pooled covariance matrix, Σ . Parametric tests for the multivariate model (2) assume that the random error components are independent and multivariate normal with means of zero and a common covariance matrix (i.e., $NID[\mathbf{0}_{(K-1)}, \mathbf{D}'\Sigma\mathbf{D}]$, where $\mathbf{0}_{(K-1)}$ is a $(K-1)$ vector of zeros). In contrast to the univariate approach (1), which requires that the diagonal elements of $\mathbf{D}'\Sigma\mathbf{D}$ to be equal, the multivariate model (2) does not require homogeneity of the variances for each of the K repeated measures.

1.2. Friedman model of ranks

Regardless of whether (1) the univariate ANOVA model or (2) the multivariate approach to analyzing repeated measures design is employed, there are normality assumptions for parametric models. Unfortunately, the normality assumption is violated frequently in a variety of research fields including genetics (e.g., Allison et al., 1999) and behavioral research (e.g., Micceri, 1989; Zumbo and Coulombe, 1997). Rank-based approaches can be used in order to relax the normality assumptions. Friedman (1937) ranks have been applied to related samples data as well as to data originating from single-sample repeated measures designs (Zimmerman and Zumbo, 1993). Friedman ranks have also been suggested when the assumptions of the split-plot ANOVA are violated (e.g., Beasley, 2000; Rasmussen, 1989). To apply the Friedman ranks to data from a split-plot design, let R_{ijk} be the rank assigned to measure k for the i th subject in group j . Also, let \bar{R}_{jk} be the mean of the ranks assigned to measure k by the subjects in group j , \bar{R}_{*k} be the mean of the ranks assigned to measure k averaged over all N subjects, and $\bar{R}_{**} = (K+1)/2$, which is the average of all NK Friedman ranks.

1.2.1. Univariate approach

Based on Beckett and Schucany's (1979) multiple comparison tests, Beasley (2000) demonstrated an omnibus test for the Friedman model with two or more independent groups of subjects. Based on the χ^2 analog of Scheffe's (1959) theorem (see Marascuilo, 1966), the Friedman model for $J \geq 2$ independent samples can be extended as

$$F_{\Gamma(R)} = \frac{\sum_{j=1}^J \sum_{k=1}^K n_j (\bar{R}_{jk} - \bar{R}_{*k})^2}{K(K+1)/12}. \quad (4)$$

This test approximates a χ^2 distribution with $df = (J-1)(K-1)$ asymptotically (Beasley, 2000). However, with smaller samples sizes computing a univariate F -ratio on R_{ijk} may be more appropriate.

Friedman ranks cannot be simply applied because of *any* violation of model assumptions, however. For example, Zimmerman and Zumbo (1993) demonstrated that rank transformed scores inherit the heterogeneity of variance in the original data. Similarly, the non-sphericity present in repeated measures data can also transmit to Friedman ranks (Beasley and Zumbo, 1998; Harwell and Serlin, 1994). Thus in order to have robust univariate tests, the assumptions of independence, homogeneity of variance, identical shape, and sphericity must still preside (Agresti and Pendergast, 1986; Serlin

and Harwell, 2001). Therefore if normality assumptions are not met, then univariate approaches applied to Friedman ranks can be used in order to relax the normality assumptions by assuming that the error components are identically distributed random variables from some continuous distribution, not necessarily the normal (i.e., $IID[0, \sigma_\epsilon^2]$ for all j and k). Furthermore, if normality *and* sphericity conditions do not hold, then Friedman ranks can be analyzed with either (1) the univariate ANOVA model with df -corrections (e.g., Huynh, 1978; Huynh and Feldt, 1976; Lecoutre, 1991) or (2) the multivariate approach.

1.2.2. Multivariate approach

Hollander and Sethuraman (1978) developed a multivariate statistic to test for discordance in ranking patterns for $J = 2$ groups of raters. Beasley (2000) proposed an extension of this statistic for $J \geq 2$ groups. For the j th group, let $\mathbf{m}_j = [(\bar{R}_{j1} - \bar{R}_{*1}), \dots, (\bar{R}_{jk} - \bar{R}_{*k}), \dots, (\bar{R}_{jK} - \bar{R}_{*K})]'$, for $j = 1, \dots, J$, be a K -dimensional column vector of deviations for the k th measure for each group j . Let \mathbf{S}_R be the sample within-group error covariance matrix of the Friedman ranks. Also, define \mathbf{S}_R^* as the Kronecker product of a diagonal matrix $\mathbf{n} = \text{diag}\{1/n_1, \dots, 1/n_J\}$ and \mathbf{S}_R , $\mathbf{S}_R^* = \mathbf{n} \otimes \mathbf{S}_R$. Then, the following statistic takes the general quadratic form:

$$H_{(R)} = \mathbf{M}'\mathbf{S}_R^{*-1}\mathbf{M}, \tag{5}$$

where $\mathbf{M} = [\mathbf{m}'_1, \dots, \mathbf{m}'_j, \dots, \mathbf{m}'_J]'$ is a JK column vector. Because the data matrix has a fixed mean of $(K + 1)/2$, both \mathbf{S}_R and \mathbf{S}_R^* will be singular. Therefore, a generalized inverse must be employed to compute \mathbf{S}_R^{*-1} . The distribution of $(N - 1)H_{(R)}$ approximates a χ^2 distribution with $df = (J - 1)(K - 1)$ asymptotically (Beasley, 2000; Hollander and Sethuraman, 1978). It should be noted that $H_{(R)}$ is the Hotelling–Lawley trace for the interaction effect from a multivariate profile analysis performed on the Friedman ranks. Thus, this procedure could also be accomplished by computing $\mathbf{R}^* = \mathbf{R}\mathbf{D}$, where \mathbf{R} is the $(N \times K)$ data matrix for the Friedman ranks, and then replacing \mathbf{Y}^* with \mathbf{R}^* in the multivariate model (2). Tests from the multivariate model (2) applied to Friedman ranks assume that the J groups have independent and identically distributed error components with means of zero and a common variance matrix for each of the K measures separately (i.e., $IID[\mathbf{0}_{(K-1)}, \mathbf{D}'\mathbf{\Sigma}\mathbf{D}]$). Consistent with Agresti and Pendergast (1986), transforming $H_{(R)}$ 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 and Serlin, 1997). Although using an exact critical value from the Hotelling–Lawley trace distribution would be the most powerful approach with small sample sizes (Beasley, 2002), critical values for many designs with small sample sizes are often not tabled and thus are not readily available to applied researchers.

1.3. Aligned ranks

Interaction tests performed on Wilcoxon ranks applied to data from between-subjects factorial designs have performed poorly compared with their normal theory counterparts

(e.g., Salter and Fawcett, 1993; Toothaker and Newman, 1994). Interaction tests for the rank transform (Conover and Iman, 1981) have also performed poorly for a variety of other designs (Akritas, 1990; Headrick and Rotou, 2001; Headrick and Vineyard, 2001; Thompson, 1991, 1993). This is because the expected value of ranks for an observation in one cell has a non-linear dependence on the original means of the other cells (Headrick and Sawilowsky, 2000; Thompson, 1991). Consequently, interaction and main effect relationships are not maintained after rank transformations are performed (Blair et al., 1987). As a result, a parametric test for interaction applied to ranks lacks an invariance property, which produces distorted Type I and Type II error rates. Thus, additivity in the original data does not imply additivity of the ranks, nor does additivity in the ranks imply additivity in the original data. Therefore, Hora and Conover (1984) warned that simply ranking the data does not result in an adequate test for non-additivity (i.e., interaction).

Several studies have shown that aligning the data before ranking yields better tests of the interactions in factorial designs (Beasley, 2002; Higgins and Tashtoush, 1994). Data from a split-plot design have *three* nuisance parameters that must be removed in order to align the scores for ranking and subsequent analysis of interaction effects. Specifically, the three nuisance parameters from model (1) are the repeated measures main effect (τ_k), the between-subjects main effect (β_j), and subjects' individual differences effect that is nested in the between-subjects factor, $\pi_{i(j)}$. After applying Friedman ranks to data from a split-plot design, all subjects have the same marginal mean of $(K+1)/2$. Thus, it is an attempt to eliminate the between-subjects variance (β_j) and the nested subjects variance ($\pi_{i(j)}$) in model (1) (Hollander and Wolfe, 1973, p. 143). However, the Friedman model rank method does not remove the repeated measures main effect (τ_k) from model (1). Higgins and Tashtoush (1994) proposed the following method of alignment:

$$Y_{ijk}^* = [Y_{ijk} - \bar{Y}_{**k} - \bar{Y}_{ij*} + \bar{Y}_{***}], \quad (6)$$

where \bar{Y}_{**k} is the marginal mean of the k th measure averaged over all N subjects, \bar{Y}_{ij*} is the mean for the i th subject averaged across the K measures, and \bar{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 regressing Y_{ijk} on a set of $(N-1)$ dummy codes that represent the subject effect ($\pi_{i(j)}$) and a set of $(K-1)$ contrast codes that represent the repeated-measures main effect (τ_k) from model (1). Define A_{ijk} as the Friedman ranks applied to the aligned scores (6). These aligned scores have the nuisance effects removed so that a subsequent test performed on A_{ijk} will be sensitive to detecting interaction effects among location parameters.

Beasley (2000) demonstrated that test statistics for the unaligned Friedman ranks (R_{ijk}) maintained the expected Type I error rate when a slight repeated measures main effect was present; however, without removing the repeated measures main effect through alignment, the statistics for testing the interaction suggested by Beasley (2000) may demonstrate low statistical power when a strong repeated measures main effect is present in each group. Therefore, the purpose of this investigation is to examine whether aligning the data before applying Friedman ranks results in (a) Type I error rates that are more consistent with the nominal alpha and (b) more statistical power.

2. Method

2.1. Design

A 4 (sample size: $n_j=10, 15, 20,$ and 30) $\times 2$ (spherical and non-spherical covariance structure) $\times 4$ (shape of error distribution: normal, double exponential, and exponential, chi-square with $df = 1$) $\times 3$ (degree of a main effect: $c = 0, 0.25,$ and 0.50) factorial design was employed for this simulation study of Type I error rates. For each of these 96 conditions, 10,000 replications were generated using SAS/IML 8.2 (SAS Institute, 2001). Comparisons were made among 14 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 unaligned Friedman ranks (R_{ijk}), and the aligned Friedman ranks (A_{ijk}), the following four statistics were calculated: (a) the conventional F -test; (b) the Lecoutre (1991) ϵ -adjusted F ; (c) the F approximate test (3) for the Hotelling–Lawley trace (H) from a multivariate profile analysis; and (d) H using a critical value from the Hotelling–Lawley trace distribution. For a $J = 3 \times K = 4$ split-plot design, the parameters for the Hotelling–Lawley trace distribution are: $s = 2$; $m = 0$; and $n = 11.5$ for $n_j = 10$, $n = 19$ for $n_j = 15$, $n = 26.5$ for $n_j = 20$, and $n = 41.5$ for $n_j = 30$. The $\alpha = 0.05$ critical values for H are 0.605, 0.350, 0.246, and 0.156 for $n_j = 10, 15, 20,$ and 30 , respectively. The extension of the Friedman χ^2 approximate test, Fr (4), was performed on R_{ijk} and A_{ijk} but not on the original scores (Y_{ijk}). The $n_j = 10$ condition was chosen because it has been used in other studies (e.g., Agresti and 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 performed on ranks 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. 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 and Keselman, 1998; Algina and Oshima, 1994; Keselman et al., 1993). The χ^2 with $df = 1$ distribution ($\chi_{(1)}^2$) was selected as a boundary condition where the errors were more skewed ($\gamma_1 = \sqrt{8}$) and extremely heavy-tailed ($\gamma_2 = 12$). Moreover, it is representative of skewed, heavy-tailed distributions found in experimental psychology, most notably reaction time data (Zumbo and Coulombe, 1997). Micceri (1989) reported that 30.9% of the data from educational and psychological research had asymmetry as extreme as the exponential and $\chi_{(1)}^2$ distributions. The normal distribution was simulated for comparison purposes.

2.2. Simulation procedures and conditions

Using the SAS/IML RANNOR function, an 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 separately. A covariance structure was 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 ($\boldsymbol{\Sigma}_j$) and pre-multiplying it by the transpose of \mathbf{X}_j to create a data matrix \mathbf{Y}_j that simulates $\boldsymbol{\Sigma}_j$:

$$\mathbf{Y}'_j = \mathbf{F}\mathbf{X}'_j \quad (7)$$

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

In the first condition, all population correlations between measures (i.e., off-diagonal elements of $\boldsymbol{\Sigma}_j$) 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 second condition, covariance structures with $\varepsilon = 0.64$ were imposed. The pairwise intercorrelations 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 of an autoregressive process.

This a simpler structure in which 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.

Three conditions of error non-normality were simulated: double exponential, exponential, and $\chi^2_{(1)}$. To simulate the error distributions for both non-normal conditions, intermediate population correlation values were derived (see Headrick and 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 (7). Then, data transformations using an extended Fleishman (1978) power method were performed (Headrick and 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 distributions non-normal. Thus, these values were transformed so that the variances and shapes of each of the K error distributions 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 (ε_{ijk}) were $IID(0, \sigma_\varepsilon^2)$ for each of the JK cells, which permitted an investigation of the 14 statistics as tests of interaction in terms of a univariate shift model for location parameters (Beasley and Zumbo, 2002). Under conditions in which the covariance structures were not spherical, however, only the less restrictive multivariate assumption (i.e., $IID[\mathbf{0}_{(K-1)}, \mathbf{D}_K \boldsymbol{\Sigma} \mathbf{D}'_K]$) was valid, thus creating a violation of the underlying assumptions for the univariate 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 $j = 1$, a vector of constants, $\mathbf{c}_1 = [0 \ 0 \ 2c \ 0]$, was

added to each observation (Y_{ilk}) for the $K = 4$ repeated measures. For group $j = 2$, $\mathbf{c}_2 = [-c \ -c \ c \ -c]$, and for the third group ($j = 3$), $\mathbf{c}_3 = [-2c \ -2c \ 0 \ -2c]$. 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 or any other distributional differences.

3. 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 -ratio, F_H refers to the F approximation (3) for the Hotelling–Lawley trace, H refers to testing the Hotelling–Lawley trace with a critical value from its referent distribution, and Fr refers to the extension of the Friedman test (4). Subscripts of Y , R , and A refer to the tests performed on the original data (Y_{ijk}), Friedman ranks (R_{ijk}), and aligned Friedman ranks (A_{ijk}), respectively. The results for the condition in which the $K = 4$ repeated measures were generated with equicorrelated ($\rho = 0.60$) and thus a spherical covariance structure is denoted by $\varepsilon = 1.00$ and $\varepsilon = 0.64$ refers to the non-spherical condition.

3.1. Type I error rates

Given $\alpha = 0.05$ and 10,000 replications, a simulated estimate has a standard error of $SE(\hat{\alpha}) = 0.0022$. Thus for empirical estimates of Type I error rates, any rejection rate 2 standard errors above 0.05 (i.e., 0.0544) was considered unacceptably 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 . The letter (a) is used to denote empirical values that were significantly above the nominal alpha (i.e., liberal). Likewise, any rejection rate below 0.0456 was considered conservative, but acceptable for power comparisons, and is denoted with the letter (b).

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 (Y_{ijk}) with non-normal error distributions and $n_j = 10$, the Type I error rates were below Bradley's (1978) criterion for a nominal alpha of 0.05, especially for data with skewed error distributions (i.e., Exponential, $\chi^2_{(1)}$). 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 or $\chi^2_{(1)}$ error distributions was somewhat conservative. By contrast, when F_ε was performed on the Friedman (R_{ijk}) and aligned Friedman ranks (A_{ijk}) of data with exponential or

Table 1
Empirical Type I error rates for the interaction tests in the absence of a repeated measures main effect ($c=0$)

$n_j = 10$	Normal		Double exponential		Exponential		Chi-square $df = 1$	
	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0474	0.0756 ^a	0.0480	0.0768 ^a	0.0478	0.0729 ^a	0.0454 ^b	0.0709 ^a
$F_{\varepsilon(Y)}$	0.0453 ^b	0.0536	0.0440 ^b	0.0515	0.0420 ^b	0.0461	0.0371 ^b	0.0419 ^b
$F_{H(Y)}$	0.0538	0.0526	0.0517	0.0458	0.0442 ^b	0.0392 ^b	0.0393 ^b	0.0345 ^b
$H_{(Y)}$	0.0428 ^b	0.0441 ^b	0.0404 ^b	0.0373 ^b	0.0353 ^b	0.0332 ^b	0.0311 ^b	0.0263 ^b
$Fr_{(R)}$	0.0442 ^b	0.0531	0.0466	0.0576 ^a	0.0435 ^b	0.0528	0.0454 ^b	0.0529
$F_{(R)}$	0.0498	0.0592 ^a	0.0521	0.0629 ^a	0.0509	0.0586 ^a	0.0520	0.0593 ^a
$F_{\varepsilon(R)}$	0.0492	0.0531	0.0511	0.0551 ^a	0.0505	0.0523	0.0513	0.0541
$F_{H(R)}$	0.0596 ^a	0.0556 ^a	0.0588 ^a	0.0576 ^a	0.0604 ^a	0.0600 ^a	0.0612 ^a	0.0579 ^a
$H_{(R)}$	0.0494	0.0462	0.0474	0.0474	0.0496	0.0499	0.0496	0.0490
$Fr_{(A)}$	0.0505	0.0598 ^a	0.0518	0.0606 ^a	0.0453 ^b	0.0549 ^a	0.0412 ^b	0.0474
$F_{(A)}$	0.0507	0.0598 ^a	0.0530	0.0618 ^a	0.0514	0.0609 ^a	0.0511	0.0585 ^a
$F_{\varepsilon(A)}$	0.0503	0.0525	0.0520	0.0550 ^a	0.0505	0.0528	0.0496	0.0522
$F_{H(A)}$	0.0600 ^a	0.0562 ^a	0.0599 ^a	0.0559 ^a	0.0600 ^a	0.0600 ^a	0.0581 ^a	0.0553 ^a
$H_{(A)}$	0.0502	0.0475	0.0504	0.0465	0.0515	0.0507	0.0479	0.0441 ^b
$n_j = 15$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0529	0.0763 ^a	0.0495	0.0784 ^a	0.0464	0.0788 ^a	0.0455 ^b	0.0750 ^a
$F_{\varepsilon(Y)}$	0.0515	0.0520	0.0477	0.0533	0.0412 ^b	0.0487	0.0386 ^b	0.0469
$F_{H(Y)}$	0.0524	0.0470	0.0507	0.0511	0.0419 ^b	0.0437 ^b	0.0381 ^b	0.0429 ^b
$H_{(Y)}$	0.0469	0.0476	0.0450 ^b	0.0457	0.0366 ^b	0.0374 ^b	0.0337 ^b	0.0383 ^b
$Fr_{(R)}$	0.0470	0.0569 ^a	0.0493	0.0593 ^a	0.0463	0.0568 ^a	0.0456	0.0575 ^a
$F_{(R)}$	0.0512	0.0604 ^a	0.0544	0.0633 ^a	0.0507	0.0610 ^a	0.0507	0.0619 ^a
$F_{\varepsilon(R)}$	0.0506	0.0537	0.0539	0.0555 ^a	0.0503	0.0540	0.0504	0.0567 ^a
$F_{H(R)}$	0.0551 ^a	0.0546 ^a	0.0594 ^a	0.0588 ^a	0.0569 ^a	0.0579 ^a	0.0547 ^a	0.0578 ^a
$H_{(R)}$	0.0511	0.0498	0.0540	0.0539	0.0501	0.0506	0.0492	0.0524
$Fr_{(A)}$	0.0526	0.0603 ^a	0.0517	0.0625 ^a	0.0440 ^b	0.0591 ^a	0.0431 ^b	0.0572 ^a
$F_{(A)}$	0.0529	0.0606 ^a	0.0528	0.0638 ^a	0.0482	0.0646 ^a	0.0521	0.0655 ^a
$F_{\varepsilon(A)}$	0.0526	0.0528	0.0527	0.0535	0.0479	0.0536	0.0511	0.0576 ^a
$F_{H(A)}$	0.0543	0.0535	0.0618 ^a	0.0572 ^a	0.0535	0.0575 ^a	0.0565 ^a	0.0569 ^a
$H_{(A)}$	0.0498	0.0480	0.0541	0.0508	0.0476	0.0527	0.0510	0.0525

^aLiberal Type I error rates that are significantly above the nominal alpha of 0.05.

^bConservative Type I error rates that are significantly below the nominal alpha of 0.05.

$\chi^2_{(1)}$ error distributions, however, Type I error rates were more consistent with the nominal alpha even with smaller sample sizes (see Tables 1–6).

The multivariate approach using F_H (3) was liberal, especially for R_{ijk} and A_{ijk} with $n_j \leq 20$. 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 $n_j = 20$, testing H (5) with an exact critical value was generally effective in controlling Type I errors as compared to

Table 2
Empirical Type I error rates for the interaction tests in the absence of a repeated measures main effect ($c=0$)

$n_j = 20$	Normal		Double exponential		Exponential		Chi-square $df = 1$	
	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0513	0.0801 ^a	0.0498	0.0731 ^a	0.0495	0.0768 ^a	0.0459	0.0756 ^a
$F_{\varepsilon(Y)}$	0.0505	0.0523	0.0468	0.0478	0.0461	0.0470	0.0398 ^b	0.0454 ^b
$F_{H(Y)}$	0.0532	0.0501	0.0484	0.0455 ^b	0.0473	0.0444 ^b	0.0392 ^b	0.0407 ^b
$H_{(Y)}$	0.0499	0.0473	0.0456	0.0426 ^b	0.0450	0.0428 ^b	0.0359 ^b	0.0383 ^b
$Fr_{(R)}$	0.0510	0.0570 ^a	0.0454 ^b	0.0549 ^a	0.0490	0.0573 ^a	0.0508	0.0550 ^a
$F_{(R)}$	0.0540	0.0600 ^a	0.0482	0.0583 ^a	0.0510	0.0601 ^a	0.0533	0.0582 ^a
$F_{\varepsilon(R)}$	0.0537	0.0520	0.0479	0.0504	0.0507	0.0520	0.0530	0.0521
$F_{H(R)}$	0.0577 ^a	0.0501	0.0535	0.0516	0.0543	0.0563 ^a	0.0539	0.0549 ^a
$H_{(R)}$	0.0548 ^a	0.0473	0.0494	0.0487	0.0512	0.0534	0.0515	0.0531
$Fr_{(A)}$	0.0531	0.0603 ^a	0.0491	0.0588 ^a	0.0492	0.0534	0.0430 ^b	0.0523
$F_{(A)}$	0.0527	0.0603 ^a	0.0496	0.0591 ^a	0.0536	0.0587 ^a	0.0520	0.0591 ^a
$F_{\varepsilon(A)}$	0.0526	0.0526	0.0495	0.0490	0.0529	0.0499	0.0508	0.0523
$F_{H(A)}$	0.0582 ^a	0.0545 ^a	0.0531	0.0500	0.0562 ^a	0.0533	0.0561 ^a	0.0553 ^a
$H_{(A)}$	0.0547 ^a	0.0521	0.0498	0.0473	0.0538	0.0515	0.0524	0.0514
$n_j = 30$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0526	0.0794 ^a	0.0506	0.0765 ^a	0.0468	0.0800 ^a	0.0481	0.0714 ^a
$F_{\varepsilon(Y)}$	0.0522	0.0531	0.0494	0.0490	0.0433 ^b	0.0514	0.0442 ^b	0.0448 ^b
$F_{H(Y)}$	0.0522	0.0512	0.0494	0.0489	0.0447 ^b	0.0482	0.0461	0.0435 ^b
$H_{(Y)}$	0.0497	0.0483	0.0453 ^b	0.0454 ^b	0.0427 ^b	0.0444 ^b	0.0431 ^b	0.0396 ^b
$Fr_{(R)}$	0.0490	0.0597 ^a	0.0470	0.0578 ^a	0.0492	0.0625 ^a	0.0475	0.0575 ^a
$F_{(R)}$	0.0520	0.0616 ^a	0.0494	0.0606 ^a	0.0515	0.0647 ^a	0.0500	0.0594 ^a
$F_{\varepsilon(R)}$	0.0519	0.0536	0.0494	0.0514	0.0512	0.0555 ^a	0.0497	0.0529
$F_{H(R)}$	0.0544	0.0527	0.0510	0.0527	0.0543	0.0544	0.0533	0.0506
$H_{(R)}$	0.0506	0.0489	0.0476	0.0485	0.0507	0.0517	0.0496	0.0470
$Fr_{(A)}$	0.0519	0.0600 ^a	0.0480	0.0593 ^a	0.0493	0.0599 ^a	0.0432 ^b	0.0518
$F_{(A)}$	0.0516	0.0603 ^a	0.0483	0.0600 ^a	0.0522	0.0635 ^a	0.0496	0.0580 ^a
$F_{\varepsilon(A)}$	0.0513	0.0523	0.0480	0.0494	0.0517	0.0531	0.0486	0.0505
$F_{H(A)}$	0.0529	0.0514	0.0525	0.0501	0.0512	0.0506	0.0541	0.0509
$H_{(A)}$	0.0494	0.0479	0.0487	0.0468	0.0493	0.0484	0.0510	0.0472

^aLiberal Type I error rates that are significantly above the nominal alpha of 0.05.

^bConservative Type I error rates that are significantly below the nominal alpha of 0.05.

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 1–6).

In Tables 3–6, it is evident that the univariate extension of the Friedman statistic for testing interactions (4) performed on R_{ijk} was inadequate when main effects were present. For example, the empirical Type I error rates for these interaction tests (Fr) were well below Bradley’s criterion for a nominal alpha when a main effect of $c =$

Table 3
Empirical Type I error rates for the interaction tests in the presence of a repeated measures main effect ($c = 0.25$)

$n_j = 10$	Normal		Double exponential		Exponential		Chi-square $df = 1$	
	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0481	0.0819 ^a	0.0507	0.0778 ^a	0.0431 ^b	0.0756 ^a	0.0455 ^b	0.0706 ^a
$F_{\varepsilon(Y)}$	0.0461	0.0562 ^a	0.0475	0.0511	0.0372 ^b	0.0458	0.0362 ^b	0.0425 ^b
$F_{H(Y)}$	0.0534	0.0534	0.0539	0.0512	0.0433 ^b	0.0430 ^b	0.0412 ^b	0.0384 ^b
$H_{(Y)}$	0.0430 ^b	0.0413 ^b	0.0451 ^b	0.0418 ^b	0.0347 ^b	0.0345 ^b	0.0293 ^b	0.0310 ^b
$Fr_{(R)}$	0.0253 ^b	0.0270 ^b	0.0213 ^b	0.0189 ^b	0.0143 ^b	0.0175 ^b	0.0120 ^b	0.0110 ^b
$F_{(R)}$	0.0503	0.0679 ^a	0.0527	0.0640 ^a	0.0512	0.0697 ^a	0.0579 ^a	0.0646 ^a
$F_{\varepsilon(R)}$	0.0494	0.0576 ^a	0.0511	0.0549 ^a	0.0483	0.0572 ^a	0.0547 ^a	0.0543
$F_{H(R)}$	0.0552 ^a	0.0584 ^a	0.0572 ^a	0.0568 ^a	0.0556 ^a	0.0566 ^a	0.0582 ^a	0.0543
$H_{(R)}$	0.0453 ^b	0.0471	0.0475	0.0475	0.0467	0.0473	0.0487	0.0454 ^b
$Fr_{(A)}$	0.0480	0.0614 ^a	0.0540	0.0604 ^a	0.0482	0.0573	0.0442 ^b	0.0551 ^a
$F_{(A)}$	0.0483	0.0616 ^a	0.0543	0.0619 ^a	0.0539	0.0637 ^a	0.0550	0.0658 ^a
$F_{\varepsilon(A)}$	0.0479	0.0547 ^a	0.0544	0.0549 ^a	0.0533	0.0554	0.0538	0.0575 ^a
$F_{H(A)}$	0.0594 ^a	0.0584 ^a	0.0628 ^a	0.0618 ^a	0.0611 ^a	0.0604 ^a	0.0611 ^a	0.0615 ^a
$H_{(A)}$	0.0496	0.0484	0.0536	0.0509	0.0505	0.0484	0.0497	0.0509
$n_j = 15$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0537	0.0783 ^a	0.0455 ^b	0.0768 ^a	0.0502	0.0713 ^a	0.0454 ^b	0.0746 ^a
$F_{\varepsilon(Y)}$	0.0525	0.0534	0.0421 ^b	0.0532	0.0452 ^b	0.0459	0.0381 ^b	0.0447 ^b
$F_{H(Y)}$	0.0574 ^a	0.0495	0.0470	0.0529	0.0464	0.0456 ^b	0.0401 ^b	0.0419 ^b
$H_{(Y)}$	0.0524	0.0438 ^b	0.0423 ^b	0.0481	0.0421 ^b	0.0388 ^b	0.0356 ^b	0.0378 ^b
$Fr_{(R)}$	0.0269 ^b	0.0280 ^b	0.0183 ^b	0.0224 ^b	0.0163 ^b	0.0147 ^b	0.0120 ^b	0.0118 ^b
$F_{(R)}$	0.0500	0.0626 ^a	0.0487	0.0629 ^a	0.0522	0.0605 ^a	0.0537	0.0637 ^a
$F_{\varepsilon(R)}$	0.0491	0.0542	0.0473	0.0521	0.0504	0.0497	0.0512	0.0512
$F_{H(R)}$	0.0560 ^a	0.0532	0.0504	0.0536	0.0536	0.0496	0.0535	0.0549 ^a
$H_{(R)}$	0.0509	0.0485	0.0450 ^b	0.0491	0.0469	0.0452 ^b	0.0480	0.0491
$Fr_{(A)}$	0.0510	0.0614 ^a	0.0485	0.0584 ^a	0.0490	0.0555 ^a	0.0440 ^b	0.0518
$F_{(A)}$	0.0515	0.0618 ^a	0.0488	0.0588 ^a	0.0532	0.0619 ^a	0.0534	0.0599 ^a
$F_{\varepsilon(A)}$	0.0512	0.0541	0.0483	0.0527	0.0525	0.0528	0.0522	0.0537
$F_{H(A)}$	0.0555 ^a	0.0527	0.0548 ^a	0.0538	0.0568 ^a	0.0510	0.0547 ^a	0.0580 ^a
$H_{(A)}$	0.0502	0.0471	0.0493	0.0479	0.0517	0.0469	0.0500	0.0526

^aLiberal Type I error rates that are significantly above the nominal alpha of 0.05.

^bConservative Type I error rates that are significantly below the nominal alpha of 0.05.

0.25 was present (see Tables 3 and 4). This “conservatism” worsened with rejection rates less than 1% as the main effect increased to $c = 0.50$. This indicates that the Friedman ranking procedure obscures an interaction effect when a repeated measures (within-subjects) main effect is present in the original data.

The univariate F test performed on R_{ijk} yielded rejection rates more consistent with the nominal alpha; however, these empirical Type I error rates were inflated when the

Table 4
 Empirical Type I error rates for the interaction tests in the presence of a repeated measures main effect
 ($c = 0.25$)

$n_j = 20$	Normal		Double exponential		Exponential		Chi-square $df = 1$	
	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0530	0.0783 ^a	0.0469	0.0763 ^a	0.0481	0.0786 ^a	0.0451 ^b	0.0768 ^a
$F_{\varepsilon(Y)}$	0.0522	0.0529	0.0449 ^b	0.0483	0.0442 ^b	0.0492	0.0408 ^b	0.0442 ^b
$F_{H(Y)}$	0.0574 ^a	0.0542	0.0463	0.0487	0.0449 ^b	0.0436 ^b	0.0428 ^b	0.0424 ^b
$H_{(Y)}$	0.0551 ^a	0.0502	0.0438 ^b	0.0459	0.0423 ^b	0.0402 ^b	0.0399 ^b	0.0402 ^b
$Fr_{(R)}$	0.0310 ^b	0.0247 ^b	0.0226 ^b	0.0213 ^b	0.0163 ^b	0.0167 ^b	0.0127 ^b	0.0122 ^b
$F_{(R)}$	0.0533	0.0608 ^a	0.0525	0.0625 ^a	0.0552 ^a	0.0653 ^a	0.0555 ^a	0.0669 ^a
$F_{\varepsilon(R)}$	0.0527	0.0496	0.0518	0.0522	0.0532	0.0534	0.0529	0.0557 ^a
$F_{H(R)}$	0.0554 ^a	0.0492	0.0551 ^a	0.0537	0.0544	0.0528	0.0513	0.0561 ^a
$H_{(R)}$	0.0519	0.0463	0.0522	0.0503	0.0512	0.0494	0.0498	0.0531
$Fr_{(A)}$	0.0522	0.0584 ^a	0.0504	0.0596 ^a	0.0467	0.0548 ^a	0.0458	0.0524
$F_{(A)}$	0.0520	0.0587 ^a	0.0506	0.0601 ^a	0.0514	0.0595 ^a	0.0527	0.0601 ^a
$F_{\varepsilon(A)}$	0.0516	0.0512	0.0504	0.0528	0.0510	0.0495	0.0517	0.0509
$F_{H(A)}$	0.0580 ^a	0.0531	0.0557 ^a	0.0532	0.0544	0.0528	0.0551 ^a	0.0540
$H_{(A)}$	0.0542	0.0494	0.0534	0.0494	0.0518	0.0495	0.0520	0.0516
$n_j = 30$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0490	0.0815 ^a	0.0469	0.0751 ^a	0.0460	0.0763 ^a	0.0519	0.0781 ^a
$F_{\varepsilon(Y)}$	0.0481	0.0534	0.0453 ^b	0.0491	0.0431 ^b	0.0481	0.0478	0.0491
$F_{H(Y)}$	0.0515	0.0565 ^a	0.0476	0.0476	0.0467	0.0427 ^b	0.0487	0.0446 ^b
$H_{(Y)}$	0.0475	0.0531	0.0446 ^b	0.0451 ^b	0.0434 ^b	0.0400 ^b	0.0461	0.0421 ^b
$Fr_{(R)}$	0.0267 ^b	0.0302 ^b	0.0223 ^b	0.0207 ^b	0.0138 ^b	0.0175 ^b	0.0132 ^b	0.0120 ^b
$F_{(R)}$	0.0499	0.0641 ^a	0.0522	0.0614 ^a	0.0495	0.0646 ^a	0.0567 ^a	0.0666 ^a
$F_{\varepsilon(R)}$	0.0490	0.0530	0.0522	0.0495	0.0480	0.0519	0.0541	0.0538
$F_{H(R)}$	0.0515	0.0531	0.0514 ^a	0.0525 ^a	0.0485 ^a	0.0503	0.0537	0.0561 ^a
$H_{(R)}$	0.0471	0.0505	0.0492	0.0490	0.0453 ^b	0.0462	0.0506	0.0528
$Fr_{(A)}$	0.0497	0.0629 ^a	0.0506	0.0593 ^a	0.0452 ^b	0.0602 ^a	0.0451 ^b	0.0546 ^a
$F_{(A)}$	0.0496	0.0633 ^a	0.0507	0.0600 ^a	0.0489	0.0627 ^a	0.0509	0.0601 ^a
$F_{\varepsilon(A)}$	0.0497	0.0537	0.0507	0.0501	0.0486	0.0539	0.0502	0.0526
$F_{H(A)}$	0.0535	0.0562 ^a	0.0515	0.0528	0.0505	0.0525	0.0537	0.0538
$H_{(A)}$	0.0503	0.0523	0.0488	0.0501	0.0468	0.0495	0.0511	0.0496

^aLiberal Type I error rates that are significantly above the nominal alpha of 0.05.

^bConservative Type I error rates that are significantly below the nominal alpha of 0.05.

covariance structure was non-spherical or when a large repeated measures main effect ($c=0.50$) was present. By contrast, the multivariate approach to analyzing the Friedman ranks performed reasonably well, even in the presence of a large repeated measures main effect; however, these results show that the multivariate statistic $H(5)$ should be tested with an exact critical value for the Hotelling–Lawley trace distribution instead of using $F_H(3)$, even with sample sizes as large as $n_j = 30$.

Table 5
Empirical Type I error rates for the interaction tests in the presence of a repeated measures main effect ($c = 0.50$)

$n_j = 10$	Normal		Double exponential		Exponential		Chi-square $df = 1$	
	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0506	0.0772 ^a	0.0483	0.0811 ^a	0.0433 ^b	0.0741 ^a	0.0431 ^b	0.0695 ^a
$F_{s(Y)}$	0.0490	0.0498 ^a	0.0449 ^b	0.0549 ^a	0.0383 ^b	0.0455 ^b	0.0343 ^b	0.0395 ^b
$F_{H(Y)}$	0.0557 ^a	0.0542	0.0483	0.0507	0.0411 ^b	0.0437 ^b	0.0371 ^b	0.0370 ^b
$H_{(Y)}$	0.0442 ^b	0.0438 ^b	0.0386	0.0423	0.0332 ^b	0.0344 ^b	0.0286 ^b	0.0266 ^b
$Fr_{(R)}$	0.0042 ^b	0.0046 ^b	0.0024 ^b	0.0029 ^b	0.0030 ^b	0.0032 ^b	0.0034 ^b	0.0015 ^b
$F_{(R)}$	0.0561 ^a	0.0718 ^a	0.0593 ^a	0.0738 ^a	0.0612 ^a	0.0706 ^a	0.0635 ^a	0.0720 ^a
$F_{s(R)}$	0.0517	0.0530 ^a	0.0532	0.0534	0.0534	0.0544	0.0538	0.0545 ^a
$F_{H(R)}$	0.0531	0.0521	0.0530	0.0474	0.0543	0.0504	0.0490	0.0500
$H_{(R)}$	0.0434 ^b	0.0449	0.0434 ^b	0.0390 ^b	0.0447 ^b	0.0408	0.0394 ^b	0.0401 ^b
$Fr_{(A)}$	0.0486	0.0611 ^a	0.0508	0.0650 ^a	0.0492	0.0590 ^a	0.0433 ^b	0.0469
$F_{(A)}$	0.0492	0.0614 ^a	0.0515	0.0656 ^a	0.0548 ^a	0.0659 ^a	0.0538	0.0595 ^a
$F_{s(A)}$	0.0486	0.0562 ^a	0.0501	0.0577 ^a	0.0540	0.0586 ^a	0.0522	0.0522
$F_{H(A)}$	0.0577 ^a	0.0562 ^a	0.0597 ^a	0.0599 ^a	0.0641 ^a	0.0602 ^a	0.0606 ^a	0.0553 ^a
$H_{(A)}$	0.0472	0.0457	0.0501	0.0500	0.0529	0.0494	0.0512	0.0452
$n_j = 15$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0510	0.0834 ^a	0.0493	0.0770 ^a	0.0440 ^b	0.0753 ^a	0.0441 ^b	0.0768 ^a
$F_{s(Y)}$	0.0497	0.0533	0.0468	0.0510	0.0394 ^b	0.0477	0.0367 ^b	0.0463
$F_{H(Y)}$	0.0522	0.0526	0.0510	0.0504	0.0433 ^b	0.0425 ^b	0.0404 ^b	0.0423 ^b
$H_{(Y)}$	0.0473	0.0478	0.0463	0.0441 ^b	0.0364 ^b	0.0390 ^b	0.0354 ^b	0.0364 ^b
$Fr_{(R)}$	0.0047 ^b	0.0039 ^b	0.0042 ^b	0.0045 ^b	0.0031 ^b	0.0031 ^b	0.0026 ^b	0.0023 ^b
$F_{(R)}$	0.0556 ^a	0.0711 ^a	0.0614 ^a	0.0734 ^a	0.0605 ^a	0.0716 ^a	0.0635 ^a	0.0714 ^a
$F_{s(R)}$	0.0513	0.0515	0.0545 ^a	0.0564 ^a	0.0524	0.0526	0.0552 ^a	0.0543
$F_{H(R)}$	0.0501	0.0486	0.0516	0.0510	0.0505	0.0462	0.0502	0.0503
$H_{(R)}$	0.0442 ^b	0.0438 ^b	0.0473	0.0457	0.0459	0.0425 ^b	0.0452	0.0455
$Fr_{(A)}$	0.0518	0.0592 ^a	0.0536	0.0599 ^a	0.0471	0.0577 ^a	0.0424 ^b	0.0508
$F_{(A)}$	0.0520	0.0595 ^a	0.0542	0.0613 ^a	0.0517	0.0630 ^a	0.0521	0.0615 ^a
$F_{s(A)}$	0.0518	0.0522	0.0540	0.0534	0.0511	0.0541	0.0511	0.0533
$F_{H(A)}$	0.0577 ^a	0.0566 ^a	0.0577 ^a	0.0517	0.0575 ^a	0.0555 ^a	0.0551 ^a	0.0576 ^a
$H_{(A)}$	0.0523	0.0497	0.0524	0.0463	0.0520	0.0507	0.0497	0.0524

^aLiberal Type I error rates that are significantly above the nominal alpha of 0.05.

^bConservative Type I error rates that are significantly below the nominal alpha of 0.05.

Tests performed on the aligned Friedman ranks (A_{ijk}) generally maintained the expected Type I error rate in the presence of a strong repeated measures main effect (see Tables 3–6). The only problem exhibited was that with smaller sample sizes ($n_j \leq 20$) F_H (3) performed on A_{ijk} inflated the Type I error rates, which was more effectively controlled by testing H with a critical value from the Hotelling–Lawley trace distribution. With a sample size of $n_j = 30$, however, both multivariate tests for A_{ijk} held the Type I error rates near 0.05.

Table 6
Empirical Type I error rates for the interaction tests in the presence of a repeated measures main effect
($c = 0.50$)

$n_j = 20$	Normal		Double exponential		Exponential		Chi-square $df = 1$	
	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0480	0.0813 ^a	0.0460	0.0760 ^a	0.0484	0.0777 ^a	0.0476	0.0783 ^a
$F_{\varepsilon(Y)}$	0.0472	0.0551 ^a	0.0449 ^b	0.0494	0.0446 ^b	0.0460	0.0419 ^b	0.0472
$F_{H(Y)}$	0.0509	0.0550 ^a	0.0469	0.0494	0.0440 ^b	0.0432 ^b	0.0436 ^b	0.0428 ^b
$H_{(Y)}$	0.0486	0.0526	0.0438 ^b	0.0465	0.0423 ^b	0.0404 ^b	0.0405 ^b	0.0403 ^b
$Fr_{(R)}$	0.0062 ^b	0.0057 ^b	0.0040 ^b	0.0026 ^b	0.0029 ^b	0.0028 ^b	0.0028 ^b	0.0028 ^b
$F_{(R)}$	0.0598 ^a	0.0726 ^a	0.0558 ^a	0.0694 ^a	0.0605 ^a	0.0698 ^a	0.0606 ^a	0.0738 ^a
$F_{\varepsilon(R)}$	0.0557 ^a	0.0542	0.0503	0.0492	0.0529	0.0504	0.0528	0.0537
$F_{H(R)}$	0.0520	0.0518	0.0498	0.0493	0.0527	0.0477	0.0508	0.0487
$H_{(R)}$	0.0494	0.0498	0.0461	0.0467	0.0503	0.0453 ^b	0.0467	0.0460
$Fr_{(A)}$	0.0497	0.0603 ^a	0.0518	0.0572 ^a	0.0482	0.0564 ^a	0.0412 ^b	0.0528
$F_{(A)}$	0.0497	0.0605 ^a	0.0519	0.0572 ^a	0.0526	0.0601 ^a	0.0490	0.0611 ^a
$F_{\varepsilon(A)}$	0.0496	0.0514	0.0517	0.0495	0.0523	0.0516	0.0482	0.0524
$F_{H(A)}$	0.0543	0.0561 ^a	0.0536	0.0528	0.0543	0.0519	0.0516	0.0525
$H_{(A)}$	0.0520	0.0525	0.0508	0.0505	0.0515	0.0490	0.0488	0.0494
$n_j = 30$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.0540	0.0822 ^a	0.0487	0.0812 ^a	0.0467	0.0767 ^a	0.0444 ^b	0.0825 ^a
$F_{\varepsilon(Y)}$	0.0537	0.0522	0.0470	0.0521	0.0442 ^b	0.0481	0.0412 ^b	0.0476
$F_{H(Y)}$	0.0544	0.0483	0.0506	0.0516	0.0462	0.0509	0.0408 ^b	0.0440 ^b
$H_{(Y)}$	0.0513 ^a	0.0447	0.0466	0.0481	0.0431 ^b	0.0478	0.0378 ^b	0.0415 ^b
$Fr_{(R)}$	0.0063 ^b	0.0046 ^b	0.0034 ^b	0.0038 ^b	0.0037 ^b	0.0045 ^b	0.0029 ^b	0.0036 ^b
$F_{(R)}$	0.0558 ^a	0.0719 ^a	0.0565 ^a	0.0738 ^a	0.0630 ^a	0.0723 ^a	0.0580 ^a	0.0735 ^a
$F_{\varepsilon(R)}$	0.0513	0.0551 ^a	0.0512	0.0535	0.0540	0.0531	0.0484	0.0561 ^a
$F_{H(R)}$	0.0481	0.0539	0.0485	0.0495	0.0525	0.0528	0.0477	0.0490
$H_{(R)}$	0.0442 ^b	0.0510	0.0460	0.0467	0.0480	0.0489	0.0433 ^b	0.0446 ^b
$Fr_{(A)}$	0.0508	0.0589 ^a	0.0515	0.0622 ^a	0.0488	0.0596 ^a	0.0447 ^b	0.0551 ^a
$F_{(A)}$	0.0514	0.0592 ^a	0.0520	0.0626 ^a	0.0529	0.0647 ^a	0.0517	0.0619 ^a
$F_{\varepsilon(A)}$	0.0513	0.0518	0.0519	0.0532	0.0524	0.0531	0.0505	0.0529
$F_{H(A)}$	0.0541	0.0508	0.0535	0.0525	0.0522	0.0530	0.0514	0.0513
$H_{(A)}$	0.0505	0.0465	0.0505	0.0491	0.0485	0.0501	0.0484	0.0500

^aLiberal Type I error rates that are significantly above the nominal alpha of 0.05.

^bConservative Type I error rates that are significantly below the nominal alpha of 0.05.

3.2. Power comparison

When data with non-normal error distributions were analyzed, the empirical Type I error rates for tests performed on Y_{ijk} , R_{ijk} , and A_{ijk} , were often unstable with both conservative and liberal rejection rates. Therefore, an additional simulation study with 10,000 replications per condition was conducted to investigate whether any of the

procedures that maintained the Type I error rate would demonstrate an advantage in statistical power. The sample size, shape of the error distributions, and covariance structure conditions previously described in Section 2 were used. To simulate an interaction effect among the location parameters of the original data (Y_{ijk}), a vector of $\Delta_1 = \Delta_2 = [-\delta \quad -\delta \quad \delta \quad -\delta]$ was added to each $1 \times K$ observation vector in groups $j = 1$ and 2. The third group was not transformed; thus, $\Delta_3 = [0 \quad 0 \quad 0 \quad 0]$. 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 Tables 3 and 4. To investigate situations with lower statistical power, smaller interaction effects ($\delta = 0.125$ and 0.250) were also simulated. Although interactions can exist in the absence of main effects, situations where both effects exist are more common in practice. Therefore, other interaction patterns were not investigated. Because variances and shapes of the error distributions were held constant across the $J = 3$ groups and $K = 4$ repeated measures, the rejection rates reported in Tables 7–10 represent empirical estimates of statistical power in terms of location parameters rather than other distributional differences.

For each sample size condition, the Type I error rates of the 14 tests were evaluated under 24 conditions (Tables 1–6). Thus, approximately 1.2 (0.05×24) of the empirical Type I error rates for any test would be expected to exceed $\alpha + 2SE(\hat{\alpha}) = 0.0544$. Therefore, any test that had two or more rejection rates above 0.0544 across the 24 conditions was excluded from the power comparisons. Furthermore, because many of the tests performed on R_{ijk} 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 A_{ijk} with $n_j = 20$ were also excluded because of liberal empirical alpha rates (see Tables 1–6). The rejection rates for all tests under the $\delta = 0.375$ and $n_j \geq 20$ conditions were near unity and thus not reported.

As expected, the results show that tests performed on the original scores (Y_{ijk}) with normal error distributions demonstrated a power advantage over the rank-based tests. For data with symmetric, heavy-tailed (i.e., double exponential) error distributions, the parametric procedures exhibited a slight power advantage over tests performed on R_{ijk} and A_{ijk} . 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 skewed, heavy-tailed (i.e., exponential or $\chi^2_{(1)}$) error distributions, however, there was a considerable advantage to using aligned Friedman ranks. For the multivariate approach, a power advantage of $H_{(A)}$ over $H_{(R)}$ and $H_{(Y)}$ was evident with $n_j = 10$. For example, with an exponential error distribution, non-spherical covariance structure ($\varepsilon = 0.64$), and smaller effect size ($\delta = 0.25$), $H_{(A)}$ exhibited an empirical power estimate of approximately 67% rejection at $\alpha = 0.05$; whereas, $H_{(R)}$ and $H_{(Y)}$ exhibited lower empirical power estimates of approximately 54% and 52% rejection, respectively (see Table 7). This power advantage was larger for the more skewed $\chi^2_{(1)}$ error distribution. For example, with $n_j = 20$, $\varepsilon = 0.64$, and $\delta = 0.125$, $H_{(A)}$ exhibited an empirical power estimate of approximately 68% rejection at $\alpha = 0.05$; whereas, $H_{(R)}$ and $H_{(Y)}$ exhibited lower empirical power estimates of approximately 59% and 31% rejection, respectively (see Table 10).

Table 7
Rejection rates for the interaction tests in the presence of a repeated measures main effect and an interaction effect ($c = 0.25$)

$n_j = 10$	Normal		Double exponential		Exponential		Chi-square $df = 1$	
	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.2028	0.2241*	0.2185	0.2635*	0.2079	0.2556*	0.2199	0.2665*
$F_{\varepsilon(Y)}$	0.1982	0.1636	0.2098	0.1874	0.1876	0.1758	0.1924	0.1798
$F_{H(Y)}$	0.1921	0.4719	0.2175	0.5592	0.2207	0.5512	0.2452	0.5840
$H_{(Y)}$	0.1701	0.4330	0.1912	0.5231	0.1909	0.5160	0.2157	0.5499
$Fr_{(R)}$	0.1302	0.1611*	0.1527	0.2134*	0.2055	0.2671*	0.2511	0.3081*
$H_{(R)}$	0.1621	0.3121	0.2011	0.4154	0.3004	0.5358	0.3719	0.5879
$Fr_{(A)}$	0.1664	0.2324*	0.2113	0.3176*	0.3700	0.5472*	0.4815	0.6664*
$F_{(A)}$	0.1667	0.2330*	0.2134	0.3201*	0.3849	0.5641*	0.5099	0.6947*
$H_{(A)}$	0.1562	0.3094	0.2015	0.4255	0.3580	0.6719	0.4853	0.7654
$n_j = 15$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.2996	0.3450*	0.3415	0.4050*	0.3234	0.3842*	0.3221	0.3882*
$F_{\varepsilon(Y)}$	0.2963	0.2513	0.3338	0.2958	0.3049	0.2800	0.2993	0.2843
$F_{H(Y)}$	0.2881	0.7022	0.3333	0.7740	0.3336	0.7418	0.3566	0.7615
$H_{(Y)}$	0.2716	0.6851	0.3176	0.7586	0.3178	0.7265	0.3381	0.7483
$Fr_{(R)}$	0.1921	0.2773*	0.2415	0.3685*	0.3549	0.4632*	0.4310	0.5202*
$H_{(R)}$	0.2331	0.4937	0.3044	0.6373	0.4749	0.7730	0.5882	0.8198
$Fr_{(A)}$	0.2306	0.3709*	0.3152	0.5164*	0.5732	0.7951*	0.7221	0.8853*
$F_{(A)}$	0.2313	0.3721*	0.3168	0.5180*	0.5850	0.8025*	0.7384	0.8965*
$F_{\varepsilon(A)}$	0.2298	0.3368	0.3157	0.4810	0.5838	0.7773	0.7369	0.8826
$H_{(A)}$	0.2255	0.5012	0.3025	0.6683	0.5698	0.8840	0.7221	0.9390

*Indicates that the rejection rate for the tests had a Type I error rate significantly above the nominal alpha of 0.05 and should not be considered valid.

The advantage of using the aligned Friedman ranks can also be inferred from the results of the df -correction procedure (i.e., F_{ε}). Table 10 shows that with $n_j = 30$, the $\chi^2_{(1)}$ error distribution, and the non-spherical covariance structure ($\varepsilon = 0.64$), F_{ε} performed on the aligned ranks, $F_{\varepsilon(A)}$, exhibited empirical power estimates of approximately 78% rejection. $F_{\varepsilon(R)}$ and $F_{\varepsilon(Y)}$ exhibited much lower empirical power estimates of approximately 62% and 13% rejection, respectively. 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 dfs .

4. Discussion

After completing a study with a multiple group repeated measures design, a researcher may be confronted with data for which he cannot assume normality of the

Table 8
 Rejection rates for the interaction tests in the presence of a repeated measures main effect and an interaction effect ($c = 0.25$)

$n_j = 20$	Normal		Double exponential		Exponential		Chi-square $df = 1$	
	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.4113	0.4794*	0.4462	0.5468*	0.4266	0.5037*	0.4326	0.5103*
$F_{\varepsilon(Y)}$	0.4078	0.3625	0.4385	0.4183	0.4113	0.3890	0.4082	0.3893
$F_{H(Y)}$	0.3991	0.8555	0.4412	0.8869	0.4347	0.8595	0.4603	0.8765
$H_{(Y)}$	0.3866	0.8492	0.4310	0.8823	0.4241	0.8540	0.4501	0.8724
$Fr_{(R)}$	0.2591	0.4030*	0.3393	0.5144*	0.4946	0.6499*	0.5847	0.7202*
$F_{\varepsilon(R)}$	0.3019	0.4405	0.3992	0.5715	0.5741	0.7292	0.6862	0.8060
$H_{(R)}$	0.3142	0.6562	0.4139	0.7877	0.6245	0.9086	0.7385	0.9365
$Fr_{(A)}$	0.3172	0.5034*	0.4212	0.6736*	0.7309	0.9245*	0.8577	0.9706*
$F_{(A)}$	0.3169	0.5040*	0.4220	0.6746*	0.7376	0.9279*	0.8671	0.9741*
$F_{\varepsilon(A)}$	0.3161	0.4643	0.4208	0.6331	0.7369	0.9118	0.8659	0.9680
$H_{(A)}$	0.3139	0.6629	0.4156	0.8154	0.7245	0.9679	0.8545	0.9877
$n_j = 30$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.5976	0.7209*	0.6513	0.7875*	0.6181	0.7343*	0.6164	0.7374*
$F_{\varepsilon(Y)}$	0.5957	0.6033	0.6471	0.6838	0.6062	0.6139	0.6001	0.6163
$F_{H(Y)}$	0.5839	0.9715	0.6462	0.9822	0.6258	0.9653	0.6333	0.9653
$H_{(Y)}$	0.5712	0.9696	0.6359	0.9804	0.6156	0.9633	0.6219	0.9637
$Fr_{(R)}$	0.4094	0.6261*	0.5280	0.7715*	0.7297	0.8836*	0.8159	0.9212*
$F_{\varepsilon(R)}$	0.4557	0.6620	0.5934	0.8136	0.8029	0.9201	0.8774	0.9570
$H_{(R)}$	0.4627	0.8575	0.6040	0.9417	0.8323	0.9861	0.9073	0.9919
$Fr_{(A)}$	0.4660	0.7276*	0.6221	0.8829*	0.9206	0.9937*	0.9690	0.9982*
$F_{(A)}$	0.4661	0.7278*	0.6232	0.8838*	0.9236	0.9941*	0.9709	0.9983*
$F_{\varepsilon(A)}$	0.4654	0.6929	0.6231	0.8611	0.9233	0.9919	0.9708	0.9980
$F_{H(A)}$	0.4676	0.8660	0.6228	0.9583	0.9169	0.9989	0.9676	0.9993
$H_{(A)}$	0.4545	0.8599	0.6112	0.9558	0.9125	0.9988	0.9666	0.9993

*Indicates that the rejection rate for the tests had a Type I error rate significantly above the nominal alpha of 0.05 and should not be considered valid.

error distributions. If the error distributions are symmetric, then standard parametric approaches will usually suffice in terms of Type I error control and statistical power. If the error distributions are skewed, however, applying Friedman ranks and performing either univariate or multivariate test can offer a more powerful alternative. However, if the researcher is interested in testing the interaction, then the presence of a repeated measures main effect can reduce the power of interaction tests performed on Friedman ranks. That is, if the repeated measures main effect is large, then the interaction effects must be small due to ranking within the i th case, which in turn produces distorted Types I and II error rates. Aligning the data (6) before applying Friedman ranks was shown to be a more powerful alternative to simply analyzing Friedman ranks.

Table 9

Rejection rates for the interaction tests in the presence of a repeated measures main effect and an interaction effect ($c = 0.375$)

$n_j = 10$	Normal		Double exponential		Exponential		Chi-square $df = 1$	
	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.4438	0.5283*	0.4949	0.6011*	0.4760	0.5628*	0.4772	0.5728*
$F_{\varepsilon(Y)}$	0.4381	0.4101	0.4813	0.4775	0.4524	0.4474	0.4390	0.4534
$F_{H(Y)}$	0.4038	0.8605	0.4735	0.9060	0.4832	0.8751	0.5095	0.8805
$H_{(Y)}$	0.3647	0.8376	0.4358	0.8878	0.4480	0.8577	0.4760	0.8631
$Fr_{(R)}$	0.2208	0.3200*	0.2865	0.3848*	0.3361	0.3973*	0.3617	0.4250*
$H_{(R)}$	0.3159	0.6438	0.4147	0.7420	0.5093	0.7717	0.5644	0.7819
$Fr_{(A)}$	0.3380	0.5364*	0.4597	0.6892*	0.6648	0.8577*	0.7689	0.9085*
$F_{(A)}$	0.3378	0.5368*	0.4619	0.6914*	0.6795	0.8671*	0.7901	0.9183*
$H_{(A)}$	0.3175	0.6654	0.4320	0.8037	0.6438	0.9161	0.7524	0.9436
$n_j = 15$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.6422	0.7741*	0.7096	0.8332*	0.6625	0.7756*	0.6750	0.7827*
$F_{\varepsilon(Y)}$	0.6380	0.6544	0.7017	0.7329	0.6444	0.6635	0.6508	0.6753
$F_{H(Y)}$	0.6119	0.9823	0.6941	0.9850	0.6705	0.9704	0.6933	0.9691
$H_{(Y)}$	0.5943	0.9798	0.6779	0.9835	0.6553	0.9675	0.6791	0.9672
$Fr_{(R)}$	0.3768	0.5467*	0.4715	0.6351*	0.5548	0.6705*	0.6045	0.7026*
$H_{(R)}$	0.4939	0.8670	0.6278	0.9183	0.7386	0.9421	0.8033	0.9501
$Fr_{(A)}$	0.5088	0.7728*	0.6727	0.8977*	0.8704	0.9758*	0.9371	0.9926*
$F_{(A)}$	0.5089	0.7747*	0.6750	0.8995*	0.8754	0.9777*	0.9432	0.9936*
$F_{\varepsilon(A)}$	0.5080	0.7398	0.6738	0.8785	0.8749	0.9718	0.9426	0.9919
$H_{(A)}$	0.4962	0.8857	0.6564	0.9583	0.8566	0.9902	0.9296	0.9972

*Indicates that the rejection rate for the tests had a Type I error rate significantly above the nominal alpha of 0.05 and should not be considered valid.

If the skewed error distributions have a spherical covariance structure, then univariate tests performed on aligned Friedman ranks tend to have more power. However, the sphericity condition rarely holds in longitudinal data (Koch et al., 1980). Furthermore, the alignment process affects the dependency among the resultant scores (A_{ijk}), and therefore, a spherical covariance structure cannot be assumed (Thompson, 1991). Univariate df -correction procedures ($F_{\varepsilon(A)}$) and multivariate tests performed on the aligned Friedman ranks are viable alternatives if the skewed error distributions have a non-spherical covariance structure; the multivariate approach generally showed more statistical power. However, with a small sample size of $n_j = 10$ cases per group, only the multivariate test evaluated with an exact critical value ($H_{(A)}$) maintained the Type I error rate. For slightly larger samples, both $H_{(A)}$ and $F_{\varepsilon(A)}$ maintained the Type I error rate, but $H_{(A)}$ demonstrated more statistical power. Unfortunately, few texts have extensive tables of critical values for multivariate statistics, and thus, testing interactions with $H_{(A)}$ may not be readily available to applied researchers. Thus, for smaller sample sizes ($n_j \leq 20$) with skewed, non-spherical error distributions, the applied research

Table 10

Rejection rates for the interaction tests in the presence of a repeated measures main effect and an interaction effect ($c = 0.125$)

$n_j = 20$	Normal		Double exponential		Exponential		Chi-square $df = 1$	
	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.1178	0.1420*	0.1192	0.1529*	0.1193	0.1452*	0.1237	0.1490*
$F_{\varepsilon(Y)}$	0.1160	0.0948	0.1160	0.1021	0.1119	0.0935	0.1112	0.0972
$F_{H(Y)}$	0.1193	0.2617	0.1229	0.3058	0.1265	0.2927	0.1331	0.3147
$H_{(Y)}$	0.1147	0.2547	0.1170	0.2979	0.1206	0.2834	0.1272	0.3056
$Fr_{(R)}$	0.0894	0.1253*	0.1085	0.1544*	0.2073	0.2746*	0.2749	0.3615*
$F_{\varepsilon(R)}$	0.0993	0.1204	0.1204	0.1564	0.2368	0.2858	0.3177	0.4005
$H_{(R)}$	0.1023	0.1853	0.1206	0.2518	0.2514	0.4781	0.3541	0.5902
$Fr_{(A)}$	0.1022	0.1433*	0.1208	0.1793*	0.2586	0.3911*	0.3702	0.5425*
$F_{(A)}$	0.1022	0.1431*	0.1210	0.1810*	0.2700	0.4006*	0.3930	0.5649*
$F_{\varepsilon(A)}$	0.1017	0.1259	0.1206	0.1582	0.2686	0.3652	0.3912	0.5369
$H_{(A)}$	0.1027	0.1868	0.1201	0.2439	0.2665	0.5405	0.3910	0.6819
$n_j = 30$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$	$\varepsilon = 1.00$	$\varepsilon = 0.64$
$F_{(Y)}$	0.1646	0.1801*	0.1813	0.2061*	0.1663	0.1916*	0.1635	0.1980*
$F_{\varepsilon(Y)}$	0.1641	0.1214	0.1769	0.1404	0.1588	0.1307	0.1559	0.1335
$F_{H(Y)}$	0.1596	0.3952	0.1799	0.4598	0.1706	0.4297	0.1759	0.4550
$H_{(Y)}$	0.1522	0.3845	0.1718	0.4469	0.1630	0.4173	0.1688	0.4426
$Fr_{(R)}$	0.1239	0.1705*	0.1619	0.2355*	0.3138	0.4537*	0.4383	0.5796*
$F_{\varepsilon(R)}$	0.1322	0.1644	0.1757	0.2288	0.3459	0.4670	0.4859	0.6169
$H_{(R)}$	0.1314	0.2680	0.1721	0.3752	0.3662	0.6846	0.5212	0.8034
$Fr_{(A)}$	0.1356	0.1892*	0.1738	0.2613*	0.4180	0.6347*	0.5776	0.7854*
$F_{\varepsilon(A)}$	0.1360	0.1661	0.1746	0.2323	0.4258	0.6032	0.5926	0.7761
$F_{H(A)}$	0.1391	0.2768	0.1772	0.3822	0.4282	0.7805	0.5941	0.8859
$H_{(A)}$	0.1328	0.2652	0.1705	0.3689	0.4165	0.7737	0.5838	0.8819

*Indicates that the rejection rate for the tests had a Type I error rate significantly above the nominal alpha of 0.05 and should not be considered valid.

may have to settle for the less powerful univariate df -correction procedure applied to aligned Friedman ranks ($F_{\varepsilon(A)}$). If the sample size is larger ($n_j \geq 30$), the researcher can rely on the F approximation test for the multivariate statistic (3) applied to aligned Friedman ranks, which was substantially more powerful than parametric procedures with smaller effect sizes (see Table 10).

Reasons for rejecting an interaction null hypothesis are of more interest than the simple conclusion that it is false; therefore, contrast testing procedures are of great utility. Given that the aligned Friedman rank procedure is a viable approach to analyzing repeated measures data, then contrast procedures based on these methods should hold quite generally (Agesti and Pendergast, 1986; Beasley, 2000; Koch, 1969). Methods for conducting contrast tests with Friedman ranks are detailed in Beasley (2000), Beasley and Zumbo (2002), and Marascuilo and McSweeney (1967). For issues in conducting multiple planned comparisons or simultaneous test procedures, there are several

excellent references for both the univariate and multivariate approaches references (see Klockars and Hancock (2000) and Sheehan-Holt (1998) for review).

5. Interpretation of aligned Friedman ranks

It is commonly believed that the null hypotheses for the parametric tests of interaction from models (1) or (2) are similar to the null hypotheses for similar tests performed on ranks, except statistical inferences concern mean ranks. However the parametric tests and tests performed on Friedman ranks evaluate two distinctly different, although conceptually related, hypotheses concerning the similarity of ranking patterns among multiple groups. Although the null hypotheses underlying tests performed on Friedman ranks can be expressed as differences in the probability of each of the $K!$ permutations of ranks (Beasley, 2000), the concept of stochastic homogeneity applies (Randles and Wolfe, 1979; Vargha and Delaney, 1998).

To elaborate, the process of aligning the scores before ranking permits test statistics to focus on interactions among location parameters. By removing main effects, the aligned ranks should not inherit any effects due to marginal location differences (i.e., main effects). However, the alignment does not remove other marginal distributional effects, and therefore, aligned ranks may still inherit the distributional properties of the original data (e.g., heterogeneity of variance). Therefore, as analogs to parametric procedures, aligned rank tests are likely to be sensitive to variance heterogeneity, especially with unequal sample sizes (Algina and Keselman, 1998). Thus, one may conclude that the aligned Friedman rank procedures as tests of location parameters would be somewhat robust to heterogeneous variance and differences in shape when sample sizes are equal.

Similarly, Wilcox (1993) noted that parametric tests are not robust to differences in skew when sample sizes are not equal; however, they are more sensitive to mean differences when there are differences in shape and equal sample sizes. Specifically, credible inferences about means require the assumption that the population distributions are symmetric (Koch, 1969; Serlin and Harwell, 2001); whereas, credible inferences concerning location parameters in general require the assumption that the population distributions are of identical shape, not necessarily symmetric. If the assumption that the data for each of the J groups are sampled from identically shaped distributions is tenable, then a statistically significant test statistic implies an interaction due to location parameters. However, if the assumptions of identical shape and constant variance are not met, test statistics based on aligned Friedman ranks may become more sensitive to detecting any distributional difference and thus should strictly be considered tests of stochastic homogeneity (Beasley, 2000; Serlin and Harwell, 2001; Vargha and Delaney, 1998), especially with a large disparity among sample sizes.

Vargha and Delaney (1998) explicate this issue by showing that the null hypotheses of stochastic homogeneity and a null hypothesis of equal mean ranks are equivalent for non-identical, but symmetric distributions. They also demonstrated that stochastic homogeneity and a null hypothesis of equal location parameters are equivalent for identical, asymmetric distributions. Therefore, statistically significant values for interaction

tests performed on aligned Friedman ranks, typically imply a pattern in which one of the J groups is stochastically larger than the other(s) on at least one of the K repeated measures and that this stochastic dominance is not constant across all K repeated measures (Brunner and Langer, 2000; Vargha and Delaney, 1998). 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 the groups were randomly assigned. Then for the pretest measure ($k = 1$), one would expect that the two groups to be stochastically identical, $G_1(Y_{11}) = G_2(Y_{21})$, where $G_j(Y_{jk})$ is the distribution function of the k th repeated measure for the j th group. 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_{12}) that are stochastically larger than the scores for the control group (Y_{22}) the between-group probabilities of scores larger than all real values (u) are no longer equal, $P(Y_{12} > u) \leq P(Y_{22} > u)$. This conclusion that the stochastic dominance of one group over another is not constant over time is consistent with the answers that aligned rank tests provide to the ordinal question (Cliff, 1996) of “Did the groups respond differently after treatment?” Specifically, the treatment group tends to have stochastically larger gains than the control group.

Although statistically significant results may be attributed to other distributional differences, these aligned rank tests are especially sensitive to shifts in location parameters because they use mean ranks in their computation. Therefore, statistically significant test statistics performed on aligned Friedman ranks can generally be attributed to differences in location parameters (Marascuilo and McSweeney, 1977, pp. 304, 305). Hence, the univariate and multivariate aligned Friedman rank procedures can be considered robust alternatives to normal theory methods, allowing inferences concerning location parameters, as opposed to “fully non-parametric” models which specify only that observations in different cells are governed by different distribution functions (Akritas and Arnold, 1994; Akritas et al., 1997). However, given the difficulty of testing model assumptions especially with small samples (i.e., small sample estimates of skew and kurtosis are typically unstable) and the potential influence of between-group differences in variance and shape with unbalanced data, it would be more prudent to interpret results from these procedures in terms of stochastic heterogeneity (Beasley, 2000; Vargha and Delaney, 1998). That is, statistically significant tests performed on aligned Friedman ranks may not be attributed solely to differences in location parameters. In this case, aligned Friedman rank procedures produce what may be considered a more ambiguous formulation of the underlying null hypothesis that is of interest conceptually. Yet, the conclusions are consistent with the ordinal answers that Cliff (1996) has extolled as the effect of actual interest to many researchers.

References

- Agresti, A., Pendergast, J., 1986. Comparing mean ranks for repeated measures data. *Comm. Statist. Theory Methods* 15, 1417–1433.

- Akritis, M.G., 1990. The rank transform method on some two factor designs. *J. Amer. Statist. Assoc.* 85, 73–78.
- Akritis, M.G., Arnold, S.F., 1994. Fully non-parametric hypotheses for factorial designs, I: multivariate repeated-measures designs. *J. Amer. Statist. Assoc.* 89, 336–343.
- Akritis, M.G., Arnold, S.F., Brunner, E., 1997. Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *J. Amer. Statist. Assoc.* 92, 258–265.
- Algina, J., Keselman, H.J., 1998. A power comparison of the Welch James and improved general approximation tests in the split plot design. *J. Educ. Behav. Statist.* 23, 152–169.
- Algina, J., Oshima, T.C., 1994. Type I error rates for Huynh's general approximation and improved general approximation tests. *British J. Math. Statist. Psych.* 47, 151–165.
- Allison, D.B., Neale, M.C., Zannolli, R., Schork, N.J., Amos, C.I., Blangero, J., 1999. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *Am. J. Hum. Genet.* 65, 531–544.
- Beasley, T.M., 1994. CORRMTX: generating correlated data matrices in SAS/IML. *Appl. Psych. Measure.* 18, 95.
- Beasley, T.M., 2000. Nonparametric tests for analyzing interactions among intra-block ranks in multiple group repeated measures designs. *J. Educ. Behav. Statist.* 25, 20–59.
- Beasley, T.M., 2002. Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multiv. Behav. Res.* 37, 197–226.
- Beasley, T.M., Zumbo, B.D., 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.
- Beasley, T.M., Zumbo, B.D., 2002. Aligned rank tests for interactions in split-plot designs: distributional assumptions and stochastic homogeneity. Manuscript submitted for publication.
- Beckett, J., Schucany, W.R., 1979. Concordance among categorized groups of judges. *J. Educ. Statist.* 4, 125–137.
- Blair, R.C., Sawilowsky, S.S., Higgins, J.J., 1987. Limitations of the rank transform statistic in test for interactions. *Comm. Statist. Simulation Comput.* 16, 1133–1145.
- Boik, R.J., 1993. The analysis of two-factor interactions in fixed effects linear models. *J. Educ. Statist.* 18, 1–40.
- Box, G.E.P., 1954. Some theorems on quadratic forms applied in the study of analysis of variance problems, I: effect of inequality of variance in the one-way classification. *Ann. Math. Statist.* 25, 290–302.
- Bradley, J.V., 1978. Robustness? *British J. Math. Statist. Psych.* 31, 144–152.
- Brunner, E., Langer, F., 2000. Nonparametric analysis of ordered categorical data in designs with longitudinal observations and small sample sizes. *Biometrical J.* 42, 663–675.
- Cliff, N., 1996. Answering ordinal questions with ordinal data using ordinal statistics. *Multiv. Behav. Res.* 31, 331–350.
- Conover, W.J., Iman, R.L., 1981. Rank transformations as a bridge between parametric and non-parametric statistics. *Amer. Statist.* 35, 124–133.
- Fleishman, A.I., 1978. A method for simulating non-normal distributions. *Psychometrika* 43, 521–532.
- Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Assoc.* 32, 675–701.
- Greenhouse, S.W., Geisser, S., 1959. On methods in the analysis of profile data. *Psychometrika* 24, 95–112.
- 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. *Comput. Statist. Data Anal.* 17, 35–49.
- Harwell, M.R., Serlin, R.C., 1997. An empirical study of five multivariate tests for the single-factor repeated measures model. *Comm. Statist. Simulation Comput.* 26, 605–618.
- Headrick, T.C., Rotou, O., 2001. An investigation of the rank transformation in multiple regression. *Comput. Statist. Data Anal.* 38, 203–215.
- 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. *Comm. Statist. Simulation Comput.* 29, 1059–1088.
- Headrick, T.C., Vineyard, G., 2001. An empirical investigation of four tests of interaction in the context of factorial analysis of covariance. *Multiv. Linear Regress. View* 27, 3–15.

- Hettmansperger, T.P., 1984. *Statistical Inference Based on Ranks*. Wiley, New York.
- Higgins, J.J., Tashouh, S., 1994. An aligned rank transform test for interaction. *Nonlinear World* 1, 201–211.
- Hollander, M., Sethuraman, J., 1978. Testing for agreement between two groups of judges. *Biometrika* 65, 403–411.
- Hollander, M., Wolfe, D.A., 1973. *Nonparametric Statistical Methods*. Wiley, New York.
- Hora, S.C., Conover, W.J., 1984. The F -statistic in the two-way layout with rank-score transformed data. *J. Amer. Statist. Assoc.* 79, 668–673.
- Hotelling, H., 1951. A generalized T-test and measure of multivariate dispersion. *Proceedings of the Second Berkeley Symposium in Mathematics, Statistics and Probability*, Vol. 2, pp. 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. *J. Amer. Statist. Assoc.* 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. *J. Educ. Statist.* 1, 69–82.
- 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. JAI Press, Greenwich, CT, pp. 45–70.
- 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. *Rev. Educ. Res.* 68, 350–386.
- Klockars, A.J., Hancock, G.R., 2000. Scheffé's more powerful F -protected post hoc procedure. *J. Educ. Behav. Statist.* 25, 13–19.
- Koch, G.G., 1969. Some aspects of the statistical analysis of "split-plot" experiments in completely randomized layouts. *J. Amer. Statist. Assoc.* 64, 485–506.
- Koch, G.G., Amara, I.A., Stokes, M.E., Gillings, D.B., 1980. Some views on parametric and non-parametric analysis for repeated measurements and selected bibliography. *Internat. Statist. Rev.* 48, 249–265.
- Lecoutre, B., 1991. A correction for the e approximate test in repeated measures designs with two or more independent groups. *J. Educ. Statist.* 16, 371–372.
- Lynch, M., Walsh, B., 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.
- Marascuilo, L.A., 1966. Large-sample multiple comparisons. *Psych. Bull.* 65, 280–290.
- Marascuilo, L.A., McSweeney, M., 1967. Nonparametric and post hoc comparisons for trend. *Psych. Bull.* 67, 401–412.
- Marascuilo, L.A., McSweeney, M., 1977. *Nonparametric and Distribution-Free Methods for the Social Sciences*. Brooks-Cole, Monterey, CA.
- Micceri, T., 1989. The unicorn, the normal curve, and other improbable creatures. *Psych. Bull.* 105, 156–166.
- Randles, R.H., Wolfe, D.A., 1979. *Introduction to the Theory of Non-Parametric Statistics*. Wiley, New York.
- Rasmussen, J.L., 1989. Parametric and non-parametric analysis of groups by trials design under variance-covariance inhomogeneity. *British J. Math. Statist. Psych.* 42, 91–102.
- Rasmussen, J.L., Heumann, K.A., Heumann, M.T., Botzum, M., 1989. Univariate and multivariate groups by trials analysis under violation of variance-covariance and normality assumptions. *Multiv. Behav. Res.* 24, 93–105.
- Salter, K.C., Fawcett, R.F., 1993. The ART test of interaction: a robust and powerful test of interaction in factorial models. *Comm. Statist. Simulation Comput.* 22, 137–153.
- SAS Institute, 2001. *SAS/IML user's guide* (Release 8.2). Cary, NC.
- Scheffé, H., 1959. *The Analysis of Variance*. Wiley, New York.
- Serlin, R.C., Harwell, M.R., 2001. A review of non-parametric test for complex experimental designs in educational research. Paper presented at the American Educational Research Association, Seattle, WA.
- Sheehan-Holt, J., 1998. MANOVA simultaneous test procedures: the power and robustness of restricted multivariate contrasts. *Educ. Psych. Measure.* 58, 861–881.
- Thompson, G.L., 1991. A note on the rank transform for interactions. *Biometrika* 78, 697–701.
- Thompson, G.L., 1993. A correction note on the rank transform for interactions. *Biometrika* 80, 711.

- Toothaker, L.E., Newman, D., 1994. A. Nonparametric competitors to the two way ANOVA. *J. Educ. Behav. Statist.* 19, 237–273.
- Vargha, A., Delaney, H.D., 1998. The Kruskal–Wallis test and stochastic homogeneity. *J. Educ. Behav. Statist.* 23, 170–192.
- Wilcox, R., 1993. Robustness in ANOVA. In: Edwards, E. (Ed.), *Applied Analysis of Variance in the Behavioral Sciences*. Marcel Dekker, New York, pp. 345–374.
- Winer, B.J., Brown, D.R., Michels, K.M., 1991. *Statistical Principles in Experimental Design*, 3rd Edition. McGraw-Hill, New York.
- Zimmerman, D., Zumbo, B.D., 1993. Relative power of the Wilcoxon test, the Friedman test, and the repeated-measures ANOVA on ranks. *J. Experiment. Educ.* 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. *Canad. J. Experiment. Psych.* 51, 139–149.