Journal of Modern Applied Statistical Methods

Volume 8 | Issue 1 Article 4

5-1-2009

Aligned Rank Tests for Interactions in Split-Plot Designs: Distributional Assumptions and Stochastic Heterogeneity

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

Bruno D. Zumbo
University of British Columbia, bruno.zumbo@ubc.ca

Journal of Modern Applied Statistical Methods
May 2009, Vol. 8, No. 1, 16-50

1538 – 9472/09/$95.00

Aligned Rank Tests for Interactions in Split-Plot Designs: Distributional Assumptions and Stochastic Heterogeneity

T. Mark Beasley
University of Alabama at Birmingham

Bruno D. Zumbo
University of British Columbia

Three aligned rank methods for transforming data from multiple group repeated measures (split-plot) designs are reviewed. Univariate and multivariate statistics for testing the interaction in split-plot designs are elaborated. Computational examples are presented to provide a context for performing these ranking procedures and statistical tests. SAS/IML and SPSS syntax code to perform the procedures is included in the Appendix.

Key words: nonparametrics, aligned ranks, split-plot design, repeated measures, stochastic heterogeneity.

Introduction

Measuring pre-treatment or baseline levels of behavior, aptitude, achievement, or pre-existing status is often necessary as a means of assessing the internal validity of applied research (Cook & Campbell, 1979). Therefore, repeated measures designs involving two or more independent groups (split-plot designs) are among the most common experimental designs in educational, psychological, developmental, and many other fields of scientific research (e.g., Keselman et al., 1998; Koch, Amara, Stokes, & Gillings, 1980). Various statistical procedures have been suggested for analyzing data from split-plot designs when parametric model assumptions are violated. The focus here is aligned rank procedures for testing the interaction.

The effects of ranking on data and the resultant test statistics for one- and two-factor designs involving only between-subjects factors (e.g., Blair, Sawilowsky, & Higgins, 1987; Sawilowsky, Blair, & Higgins, 1989; Vargha & Delaney, 1998; Toothaker & Newman, 1994; Wilcox, 1993; Zimmerman, 1996) and single sample within-subjects designs (e.g., Agresti & Pendergast, 1986; Harwell & Serlin, 1994, 1997; Zimmerman & Zumbo, 1993) are well known. However, there have been fewer investigations concerning the effects of ranking in split-plot designs (e.g., Akritas & Arnold, 1994; Beasley, 2000, in press; Brunner & Langer, 2001; Higgins & Tashtoush, 1994; Koch, 1969).

T. Mark Beasley is Associate Professor of Public Health in the Department of Biostatistics in the School of Business. Email: mbeasley@uab.edu. Bruno D. Zumbo is Professor of Measurement, Evaluation and Research Methodology, as well as a member of the Department of Statistics and the Institute of Applied Mathematics. Email: bruno.zumbo@ubc.ca.

Methodology

Parametric Models for Split-Plot Designs: Univariate Approach

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

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

where, j is referenced to the J groups of the between-subjects factor, i is referenced to the n_j subjects nested within the j^{th} group, k is referenced to the K levels of the within-subjects (repeated measures) factor, ε_{ijk} is a random

error vector, and $N = \Sigma n_j$ 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, Brown, & Michels, 1991). In genetics experiments, the interaction typically indicates differential growth rates for organisms of different genotypes (Lynch & Walsh, 1998).

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

$$H_{0(JxK)} : \sum_{j=1}^{J} \sum_{k=1}^{K} (\beta\tau_{jk})^2 = 0 \qquad (2)$$

In using the parametric *F*-ratio for testing the interaction, the random error components (ε_{ijk}) are assumed to be independent and identically distributed with a mean of zero, a common variance (σ_ε^2), and normal shape for each of the *JK* cells (i.e., NID[$0, \sigma_\varepsilon^2$] for all *j* and *k*). By requiring identical error distributions, it can be assured that a rejection of the null hypothesis in (2) is due to shifts (differences) among location parameters. Furthermore, by assuming normal error distributions means as estimates of location will yield the maximum statistical power for rejecting (2).

For $K > 2$, there is an additional assumption concerning the sphericity of the pooled covariance matrix. If the pooled covariance matrix is non-spherical, the *F*-ratio is valid if the degrees-of-freedom (*df*s) are corrected by a factor epsilon (see Huynh & Feldt, 1970). Methods for estimating epsilon have been investigated for over four decades (e.g., Box, 1954; Greenhouse & Geisser, 1959; Huynh & Feldt, 1970, 1976; Lecoutre, 1991). Also, general approximate methods to correct the *df*s 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 & Keselman, 1998; Algina & Oshima, 1994; Keselman & Algina, 1996) and thus will not be elaborated.

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 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 NxK data matrix of scores (Y_{ijk}) and $\mathbf{D}$ is a Kx($K-1$) difference matrix of the general form:

$$\mathbf{D} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix} \qquad (3)$$

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

$$\mathbf{Y}^*_j = \mathbf{M}** + \mathbf{B}_j + \mathbf{E}_j, \qquad (4)$$

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. Testing the null hypothesis ($H_{0(K)}$: $\mathbf{M}** = \mathbf{0}_{(K-1)}$,where $\mathbf{0}_{(K-1)}$ is a ($K-1$) vector of zeros) is equivalent to testing the repeated measures main effect. With the original scores expressed as difference scores, the multivariate model (4)

contains only between-subjects effects. Thus, the null hypothesis in (2) can be expressed as:

$$H_{0(JxK)}: \mathbf{B}_1 = \mathbf{B}_2 = \ldots = \mathbf{B}_j = \ldots = \mathbf{B}_J, \quad (5)$$

where $\mathbf{B}_j$ is a (K-1) vector of between-subjects effects (i.e., mean differences) for the j^{th} group. Thus, the variables from the $\mathbf{Y}^*$ matrix are defined as difference scores and the null hypothesis in (2) can also be expressed as:

$$H_{0(JxK)}: (\mu_{1k}-\mu_{1k'}) = (\mu_{2k}-\mu_{2k'}) = \ldots = (\mu_{jk}-\mu_{jk'}) = \ldots = (\mu_{Jk}-\mu_{Jk'}), \text{ for } k \neq k' ; k = 1, \ldots K. \quad (6)$$

To illustrate the assumptions underlying the multivariate approach to repeated measures data, define $\mathbf{\Sigma}_j$ as the K x K covariance matrix of $\mathbf{Y}_j$. The homogeneity of covariance assumption requires that the J covariance matrices ($\mathbf{\Sigma}_j$) are equivalent so that they can be combined to form the pooled covariance matrix, $\mathbf{\Sigma}$. Parametric tests for the multivariate model (4) assume that the random error components are independent and multivariate normal with means of zero and a common covariance matrix (i.e., NID[$\mathbf{0}_{(K-1)}$, $\mathbf{D'\Sigma D}$]).

In contrast to the univariate approach (1), the multivariate model (4) does not require homogeneity of the variances for each of the K repeated measures. That is, the multivariate approach does not require the diagonal elements of $\mathbf{\Sigma}$ to be equal. By taking difference scores this also translates into not requiring the (K−1) transformed variables ($\mathbf{Y}^*$) to have the same variances. For example, with K=3 repeated measures, the variance of the first pairwise difference, $\sigma^2(Y_{j1}-Y_{j2})$, is not assumed to be equivalent to the variance of the second pairwise difference, $\sigma^2(Y_{j1}-Y_{j2})$, under the multivariate model (4); however, this variance homogeneity, which is equivalent to the sphericity requirement (see Winer, et al., 1991, pp. 240-243), is assumed implicitly in the univariate model (1).

Rank-Based Tests

Regardless of whether (a) the univariate ANOVA test with possible *df*-corrections (e.g., Huynh, 1978; Huynh & Feldt, 1976; Lecoutre, 1991), or (b) 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., Bradley, 1968; Cliff, 1996; Micceri, 1989; Zumbo & Coulombe, 1997).

Rank-based approaches can be used in order to relax the normality assumptions by assuming that the error components are random variables from some continuous distribution, not necessarily the normal. However, rank-based approaches cannot be simply applied due to violations of model assumptions. For example, Zimmerman and Zumbo (1993) demonstrated that rank transformed scores inherit the heterogeneity of variance in the original data. Likewise, ranks can also inherit the non-sphericity present in repeated measures data (Beasley & Zumbo, 1998; Harwell & Serlin, 1994). Thus, to test hypotheses concerning shifts in location parameters the assumptions of independence, homogeneity of variance, and identical shape must still preside (Serlin & Harwell, 2001).

Specifically, credible inferences about means require the assumption that the population distributions are symmetric (Koch, 1969; Serlin & Harwell, 2001); whereas, credible inferences concerning location parameters generally require the assumption that the population distributions are of identical shape, not necessarily symmetric (i.e., IID $[0, \sigma_\varepsilon^2]$ or IID[$\mathbf{0}_{(K-1)}$, $\mathbf{D'\Sigma D}$]). This frequently overlooked detail is one reason why so much attention has been given to rank-based procedures such as tests of stochastic homogeneity (Vargha & Delaney, 1998), distributional equivalence (Agresti & Pendergast, 1986; Beasley, 2000), or fully nonparametric hypotheses (Akritas & Arnold, 1994).

As a departure from parametric models that test differences among means, general

nonparametric models specifying only that observations in different cells which are governed by different distribution functions (Akritas & Arnold, 1994; Akritas, Arnold, & Brunner, 1997) have been developed for a variety of factorial designs including split-plot designs (Akritas & Arnold, 1994; Brunner & Langer, 2000). For a split-plot design, the fully nonparametric approach would involve ranking the data from 1 to *NK* and computing the appropriate test statistics (e.g., Serlin & Harwell, 2001).

Brunner, Domhof, and Langer (2002) warn that this practice should not be regarded as a technique for the derivation of statistics but rather as a property that can be useful for computational purposes. Therefore, fully nonparametric tests are not viewed as robust alternatives to normal theory methods, allowing direct inference concerning location parameters (Akritas, et al., 1997). Rather, statistically significant fully nonparametric tests are attributed to differences among any distributional characteristic (e.g., location, dispersion, shape). 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).

Rank-based tests, however, are especially sensitive to shifts in location parameters because they are computed using mean ranks. Therefore, even if assumptions concerning identical distributions and homogeneous variances are not tenable, the researcher may still conclude that one or more groups are stochastically dominant over another group(s). For an interaction in a multiple group repeated measures design, this concept of stochastic heterogeneity (Vargha & Delaney, 1998) implies that one or more groups tends to have higher scores on some measurement and that this stochastic dominance is not constant over the *K* measurements (Agresti & Pendergast, 1986; Brunner & Langer, 2000).

Aligned Rank Transform Procedures

Because the Rank Transform is monotonic, it is commonly believed that the null hypothesis for the parametric test of interaction (2) from model (1) is similar to the null hypothesis for similar tests performed on ranks, except statistical inferences concern mean ranks (i.e., location parameters). However, interaction tests performed on ranked data from factorial designs have performed poorly compared with their normal theory counterparts. 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 & Sawilowsky, 2000). For example, consider a two-factor model where ranks are assigned regardless of cell membership. The result is that if one of the effects is large then other effects must (because of the ranking) be small, thus producing distorted Type I and Type II error rates. Thus, a parametric test for interaction applied to ranks lacks an invariance property. Hence, interaction and main effect relationships are not expected to be maintained after rank transformations are performed (Blair, et al., 1987).

Headrick and Sawilowsky (2000) demonstrated computationally that in the presence of main effects the expected mean ranks for the cells in a factorial design can indicate an interaction when the original data do not. Moreover, Salter and Fawcett (1993) demonstrated conditions in which an interaction effect in the original data is lost in the ranking process. These situations illustrate that additivity in the original data does not imply additivity of the ranks, nor does additivity in the ranks imply additivity in the original data. Thus, Hora and Conover (1984) warned that simply ranking the data does not provide an adequate test for non-additivity (i.e., interaction) in the conventional sense of testing shifts among location parameters.

Several studies have shown that aligning the data before ranking yields better tests of the interactions among location parameters in factorial designs. Based on the work of Hodges and Lehmann (1962), 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. Higgins and Tashtoush (1994) and Koch (1969) have

proposed aligned rank procedures for testing interactions in split-plot designs. Based on Hollander and Sethuraman (1978), statistics for the Friedman (1937) model of ranks have been suggested as tests for interactions (Beasley, 2000; Rasmussen, 1989). Each of these procedures aligns the data in different ways.

Higgins and Tashtoush Alignment Procedure

Both the McSweeney (1967) and Hettmansperger (1984) 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. To elaborate, the data from a split-plot design has 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, $\pi_{i(j)}$. In terms of population effects, model (1) can be expressed as:

$$(Y_{ijk} - \mu_{***}) = \beta_j + \pi_{i(j)} + \tau_k + \beta\tau_{jk}$$

(see Winer, et al., 1991). Solving for the interaction yields:

$$\beta\tau_{jk} = (Y_{ijk} - \mu_{***}) - \beta_j - \pi_{i(j)} - \tau_k.$$

Using sample estimates of the effects yields:

$$\beta\tau_{jk} = (Y_{ijk} - \bar{Y}_{***}) - (\bar{Y}_{*j*} - \bar{Y}_{***}) - (\bar{Y}_{ij*} - \bar{Y}_{*j*}) - (\bar{Y}_{**k} - \bar{Y}_{***}), \quad (7)$$

where $\bar{Y}_{**k}$ is the marginal mean of the k^{th} measure averaged over all N subjects, $\bar{Y}_{*j*}$ is the marginal mean of the j^{th} measure averaged over all K measures and 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. Thus, to create scores aligned for effects other than the interaction ($\beta\tau_{jk}$) in model (1), equation (7) reduces to:

$$Y^*_{ijk} = [Y_{ijk} - \bar{Y}_{**k} - \bar{Y}_{ij*} + \bar{Y}_{***}], \quad (8)$$

These aligned scores have the nuisance effects removed so that a subsequent test performed on the ranks of Y^*_{ijk} will be sensitive only to detecting interaction effects. Higgins and Tashtoush (1994) proposed using this method of alignment and then ranking the aligned data from 1 to NK as follows:

$$A_{ijk} = \text{Rank}[Y_{ijk} - \bar{Y}_{**k} - \bar{Y}_{ij*} + \bar{Y}_{***}] \quad (9)$$

(see Table 1). 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). As can be inferred from (8) a set of (J−1) contrast codes that represent the between-subjects main effect (β_j) is not necessary for the residualization.

Univariate Approach

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}. As previously mentioned, 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, it is possible that the aligned ranks could also inherit some of the distributional properties of the original data as well. Thus, when performing the split-plot ANOVA F on aligned ranks, df-correction methods (e.g., Huynh & Feldt, 1976) may be employed if the pooled covariance matrix is non-spherical or if the between-subjects covariance matrices are heterogeneous (e.g., Huynh, 1978). These methods performed on ranks hold the Type I error rate near the nominal

alpha but have low statistical power in a variety of conditions (Beasley & Zumbo, 1998).

Multivariate Approach

Agresti and Pendergast (1986) proposed a multivariate rank-based test for testing repeated measures effects in a single-sample design. Beasley (2002) extended this approach for testing the interaction in a split-plot design using aligned ranks (9). Define **E** as a K x K pooled-sample cross-product error matrix with elements:

$$e_{kk'} = \sum_{j=1}^{J}\sum_{i=1}^{n_j} (A_{ijk} - \bar{A}_{jk})(A_{ijk} - \bar{A}_{jk}') . \quad (10)$$

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

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

and

$$\mathbf{C}_K = \begin{bmatrix} -3 & -1 & 1 & 3 \\ -1 & 1 & 1 & -1 \\ -1 & 3 & -3 & 1 \end{bmatrix}$$

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

It should be noted, however, that $\mathbf{C}_J$ and $\mathbf{C}_K$ need not be orthogonal, only linearly independent. For example, this matrix could be constructed by defining $\mathbf{C}_J$ and $\mathbf{C}_K$ as difference matrices in the general form of **D** in (3), and thus,

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

Based on Agresti and Pendergast (1986), Beasley (2002) proposed the statistic,

$$H_{(A)} = (\mathbf{C}_{JK}\,\mathbf{A}_{JK})' (\mathbf{C}_{JK}\,\mathbf{E^*C'}_{JK})^{-1} (\mathbf{C}_{JK}\,\mathbf{A}_{JK}). \quad (11)$$

It should be noted that $H_{(A)}$ is the Hotelling's (1931) trace for the interaction effect from a multivariate profile analysis of model (4) performed on A_{ijk}. Thus, this procedure could also be accomplished by computing $\mathbf{A}^*=\mathbf{AD}$, where **A** is the (NxK) data matrix of aligned ranks (9), and then replacing $\mathbf{Y}^*$ with $\mathbf{A}^*$ in the multivariate model (4).

Because it is a rank-based version of the Hotelling's trace, $H_{(A)}$ multiplied by $(N-1)$ should approximate a χ^2 distribution with $df = (J-1)(K-1)$, asymptotically. Consistent with Agresti and Pendergast (1986), transforming $H_{(A)}$ to an F-test may provide better control of Type I error rates as opposed to comparing $H_{(A)}(N-1)$ to a chi-square distribution with $df = (J-1)(K-1)$, especially with smaller sample sizes (Beasley, 2002; Harwell & Serlin, 1997). Based on Hotelling (1951), $H_{(A)}$ is transformed to an F approximation statistic by:

$$F_{H(A)} = [2(sn+1)/(s^2(2m+s+1))]H_{(A)}, \quad (12)$$

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 critical value for $H_{(A)}$ could be obtained from the sampling distribution of the Hotelling's trace using the s, m, and n parameters. This approach has been shown to maintain the expected Type I error rate better than the F approximate test (12) with a relatively small sample size of N = 30 (Beasley, 2002). Unfortunately, few multivariate texts have extensive tables of these critical values.

Koch Model of Ranking

In the Koch (1969) model, each of the K^2 paired differences among the repeated measures is ranked separately regardless of group membership. These ranks are then summed over the K levels of the repeated measures factor. To elaborate, for each of the K repeated measures, let $T_{ij(k,k')}$ = Rank[$Y_{ijk} - Y_{ijk'}$] using mid-ranks in case of ties. Thus, $T_{ij(k,k')}$ ranges from 1 to N, except when $k = k'$ in which case [$Y_{ijk} - Y_{ijk'}$] = 0, and thus, all values of $T_{ij(k,k)} = (N+1)/2$. Also, many of the K^2 ranked differences are reverse rankings so that the correlation between say $T_{ij(1,2)}$ and $T_{ij(2,1)}$ is -1. The final data set is defined as

$$Q_{ijk} = \sum_{k'=1}^{K} T_{ij(k,k')} \qquad (13)$$

(see Table 2). This procedure aligns the data in a less explicit manner than the Higgins-Tashtoush method (9). Specifically, the subjects' individual differences effect that is nested in the between-subjects factor, $\pi_{i(j)}$ from model (1), is removed by computing pairwise differences. This is analogous to the manner in which $\pi_{i(j)}$ is removed from Y_{ijk} in model (1) by computing $\mathbf{Y}^*=\mathbf{YD}$ and submitting $\mathbf{Y}^*$ to the multivariate model in (4), which only has between-subjects effects. Furthermore, by ranking each pairwise difference separately (i.e., $T_{ij(k,k')}$) before summing, the mean for each of the K measures and for all the Q_{ijk} values must equal $K(N+1)/2$. This eliminates the variance due to the repeated measures main effect (τ_k) from model (1).

To test the interaction, a univariate F-test on this ranked data $F_{(Q)}$ could be performed (Iman, Hora, & Conover, 1984). However, Koch (1969, *p*. 495) proposed performing a nonparametric analog to the multivariate profile analysis, $V_{(Q)}$. Let $\mathbf{Q}_{ij}$ = [Q_{ij1}, …, Q_{ijk}, …, Q_{ijK}]′ be a (Kxn_j) data matrix for the j^{th} group and let $\overline{\mathbf{Q}}_j$ be a K dimensional vector of means for the j^{th} group:

$$\overline{\mathbf{Q}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} \mathbf{Q}_{ij} = [\overline{Q}_{j1}, ..., \overline{Q}_{jk}, \ldots, \overline{Q}_{jK}]'. \qquad (14)$$

Also, let $\widetilde{\mathbf{Q}}_j = \{\overline{\mathbf{Q}}_j - K(N+1)/2\}$ be a vector of mean deviations and define the pooled covariance matrix as $\mathbf{S}_Q = 1/N\ [\mathbf{Q}_{ij} - K(N+1)/2][\mathbf{Q}_{ij} - K(N+1)/2]'$. The test statistic $V_{(Q)}$ is computed as:

$$V_{(Q)} = (N\text{-}1)/N\ [\widetilde{\mathbf{Q}}*'\mathbf{S}_Q^{*\text{-}1}\ \widetilde{\mathbf{Q}}*], \qquad (15)$$

where

$$\mathbf{S}_Q^* = \mathbf{n} \otimes \mathbf{S}_Q,\ \widetilde{\mathbf{Q}}* = [\widetilde{\mathbf{Q}}_1', ..., \widetilde{\mathbf{Q}}_2', ..., \widetilde{\mathbf{Q}}_J']',$$

and $\mathbf{n}$ = diag{$1/n_1$, $1/n_2$, . . . , $1/n_J$}.

This test is a synthesis of a nonparametric multivariate statistic for the repeated measures main effect (Koch & Sen, 1968) and the Kruskal-Wallis test. In fact, it is computationally equivalent to the Pillai's (1960) trace (V) scaled by (N−1). That is, a multivariate profile analysis performed on Q_{ijk} yields a Pillai's trace such that $V_{(Q)} = V(N-1)$. Thus, this procedure could also be accomplished by computing $\mathbf{Q}^*=\mathbf{QD}$, where $\mathbf{Q}$ is the (NxK) data matrix for the Koch model ranks (14), and then substituting $\mathbf{Y}^*$ with $\mathbf{Q}^*$ in the multivariate model (4).

$V_{(Q)}$ is a permutationally distribution-free test. As sample sizes become large the number of permutations prohibits the computation of an exact test; however, the permutation distribution is χ^2 with df = (J−1)(K−1) asymptotically. As an alternative approach to this statistic proposed by Koch

(1969), the Hotelling's trace could be used, thus calculating $H_{(Q)}$, the statistic in (11), by replacing A_{ijk} (9) with Q_{ijk} (13). As before, $H_{(Q)}$ could be transformed to an F approximation test by (12) or critical values from the multivariate referent distribution (e.g., Hotelling's trace; Pillai's trace) could be obtained in order to assess statistical significance.

Assumptions and Hypotheses for Interaction Tests Performed on Aligned Ranks

It is important to reiterate that statistically significant values of these tests performed on aligned ranks (e.g., $H_{(A)}$, $V_{(Q)}$) do not necessarily imply that the interaction is due to differences in location parameters unless additional assumptions are made. Strictly, statistical tests performed on aligned ranks involve inferences concerning the distribution of the original data. This is because the aligned ranks can be considered placeholders for the percentiles of the original raw score distribution (Y_{ijk}) with the nuisance location parameters removed (M. R. Harwell, personal communication, April 24, 2001). To elaborate, the univariate F-ratio performed on A_{ijk} or Q_{ijk} in a repeated measures design actually evaluates a null hypothesis of exchangeability or permutational equivalence:

$H_0(J\text{x}K)$:

$$\mathbf{G}_1(\mathbf{Y}_1) = \mathbf{G}_2(\mathbf{Y}_2) = \ldots = \mathbf{G}_j(\mathbf{Y}_j) = \ldots = \mathbf{G}_J(\mathbf{Y}_J), \tag{16}$$

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 not only are all J groups expected have identical distribution functions, the K repeated measures are also expected to have identical distribution functions (i.e., IID$[0, \sigma_\varepsilon^2]$ for all j and k).

The multivariate procedures (11 or 15) test a broader null hypothesis of between-group marginal homogeneity:

$H_0(J\text{x}K)$:

$$G_1(Y_{1k}) = G_2(Y_{2k}) = \ldots = G_j(Y_{jk}) = \ldots = G_J(Y_{Jk}), \text{ for } k = 1, \ldots K, \tag{17}$$

where $G_j(Y_{jk})$ is the one-dimensional distribution function of the k^{th} repeated measure for the j^{th} group (Y_{ijk}). Strictly, this is a null hypothesis of distributional equivalence across the J groups for each of the K repeated measures. That is, each of the K repeated measures may have different distribution functions, but as long as there are no distributional differences across the J groups, (17) is true. Thus, to obtain the asymptotic null distributions of the test statistics (11 or 15), it is only necessary to assume the null hypothesis (17) of between-group distributional equivalence (i.e., IID$[0, \sigma_\varepsilon^2]$ for all j for each k separately or IID$[\mathbf{0}_{(K-1)}, \mathbf{D'\Sigma D}]$) rather than to make stronger assumptions concerning joint (or permutational) distributions (i.e., common correlations between pairs of measures).

To illustrate, suppose that on the first and second measures in a $J = 2$ by $K = 3$ split-plot design, both groups are sampled from symmetric distributions with common variances (σ_1^2 and σ_2^2); however, both groups are sampled from identically skewed distributions with a common variance (σ_3^2) for the third repeated measurement. This situation would not violate the multivariate IID$[\mathbf{0}_{(K-1)}, \mathbf{D'\Sigma D}]$ assumption; however, it would violate the univariate IID$[0, \sigma_\varepsilon^2]$ assumption.

Shift Model for Aligned Ranks in Split-Plot Designs

The major purpose of the alignment process is to remove the nuisance effects (i.e., main effects) so that test statistics will be sensitive to the effect of interest (i.e., interaction). The alignment processes (9) and (13) remove the mean values for the nuisance main effects, thus involving linear transformations of the data. However, both A_{ijk} and Q_{ijk} are monotone transformations of the aligned data. As a result, these aligned rank procedures do not guarantee that test statistics

performed on A_{ijk} or Q_{ijk} will reflect shifts in location parameters. Therefore in order to make a credible inference about a single parameter, assumptions about other parameters are necessary (Serlin & Harwell, 2001).

Assuming that all JK cells have identically shaped distributions with a common variance (i.e., IID$[0,\sigma_\varepsilon^2]$ for all j and k), then rejection of the null hypothesis (16) 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 (16) as:

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

where $\mathbf{1}$ is an n_jx1 vector of ones and $\Delta_j = [\delta_{j1}\ \delta_{j2}\ \ldots\ \delta_{jk}\ \ldots\ \delta_{jK}]$ is a 1xK vector of location parameters for the j^{th} group. To illustrate the shift model for the multivariate approach to the split-plot design, define the null hypothesis in (17) as:

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

where δ_{jk} is a scalar location parameter for the jk^{th} cell. It is important to note that if (18) is true so is (19); however, if (19) is true, it does not imply that (18) is true. Likewise, a false (18) does not imply a false (19). These distinctions are important because in order to test a null hypothesis of shifts in location parameters analogous to the null hypotheses in (2) or (6), the univariate null model for ranks (18) requires an assumption that the data for all JK cells are sampled from identically shaped distributions with a common variance. By contrast, the multivariate null model for ranks (19) only requires an assumption that the distribution for each of the K repeated measures is identical for each of the J groups; however, there is no assumption that the 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 (4) and the univariate approach (1) that requires the sphericity assumption (Agresti & Pendergast, 1986). Therefore, just as the null hypotheses for the univariate (2) and multivariate (6) parametric models are equivalent, differing only in the sphericity condition required by the univariate test, the same holds for the univariate (18) and multivariate (19) shift models for aligned ranks. Furthermore, note that the null hypotheses (18) and (19) are equivalent in terms of location parameters. Thus under either the univariate IID$[0,\sigma_\varepsilon^2]$ assumption or the multivariate IID$[\mathbf{0}_{(K-1)},\mathbf{D'\Sigma D}]$ assumption, the null hypotheses in (18) or (19), respectively, reduce to an interaction null hypothesis expressed in terms of location parameters:

$$H_{0(J\text{x}K)}: (\delta_{1k}-\delta_{1k'}) = (\delta_{2k}-\delta_{2k'}) = \ldots = (\delta_{jk}-\delta_{jk'}) = \ldots = (\delta_{Jk}-\delta_{Jk'}) \text{ for } k \neq k';\ k = 1, \ldots K, \quad (20)$$

which is conceptually similar to a rejection of the parametric null hypothesis in (6). The difference between these null hypotheses is that the parametric models (1) and (4) require normally distributed error components, and thus, a rejection of (2) or (6) implies the effect must be attributed to differences among means. The shift models require identical, not necessarily normal, error distributions, and thus, a rejection of (20) implies that the effect can be attributed to differences among location parameters but not necessarily means (e.g., medians). It is important to note, however, that if (20) is false, then (18) and (19) are also false. However, a false (18) or (19) does not imply that (20) 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 these additional distributional assumptions are met.

Friedman Model of Ranks

For data from a repeated measures design, a researcher could employ the Friedman (1937) model and rank the data from 1 to K across the K levels of the repeated measures factor for each subject. The Friedman model of ranks has been applied to related samples data as well as to data originating from repeated measures designs (Zimmerman & Zumbo, 1993). The Friedman model has also been suggested when the assumptions of the split-plot ANOVA are violated (e.g., Beasley, 2000; Rasmussen, 1989). After applying the Friedman model of ranking to 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 & Wolfe, 1973, p. 143).

The Friedman model rank method does not remove the repeated measures main effect (τ_k) from model (1). Beasley (2000) demonstrated that test statistics for the Friedman model 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) can demonstrate low statistical power when a strong repeated measures main effect is present in each group. Aligning the data before applying Friedman ranks results in Type I error rates that are more consistent with the nominal alpha and a gain in statistical power, especially for a univariate approach (Beasley & Zumbo, in press).

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 after alignment (8). 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 ranks (see Table 3).

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 Scheffé's (1959) theorem (see Marascuilo, 1966), the Friedman model for $J \geq 2$ independent samples can be generalized as:

$$F(R) = \frac{\sum_{j=1}^{J}\sum_{k=1}^{K} n_j (\bar{R}_{jk} - \bar{R}_{*k})^2}{K(K+1)/12}. \qquad (21)$$

This test approximates a χ^2 distribution with $df = (J-1)(K-1)$, asymptotically (Beasley, 2000). However, with smaller samples sizes computing an F-ratio on R_{ijk} may be more appropriate if the covariance structure is spherical. Otherwise epsilon-adjusted tests or multivariate procedures are more appropriate (Beasley & Zumbo, in press).

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 $\mathbf{m}_j = [(\bar{R}_{j1}-\bar{R}_{*1}), \ldots, (\bar{R}_{jk}-\bar{R}_{*k}), \ldots, (\bar{R}_{jK}-\bar{R}_{*K})]'$, for $j = 1, \ldots, J$, be a K-dimensional column vector of deviations for the k^{th} measure for each group j. Let $\mathbf{S}_R$ be the total sample covariance matrix of the ranks computed with ordinary least squares. Also, define $\mathbf{S}_R^*$ as the Kronecker product of a diagonal matrix $\mathbf{n} = \text{diag}\{1/n_1, \ldots, 1/n_J\}$ and $\mathbf{S}_R$, $\mathbf{S}_R^* = \mathbf{n} \otimes \mathbf{S}_R$. Then, the following statistic takes the general quadratic form:

$$V(R) = \mathbf{M}' \mathbf{S}_R^{*-} \mathbf{M} \quad (22)$$

where $\mathbf{M} = [\mathbf{m}_1', \ldots, \mathbf{m}_j', \ldots, \mathbf{m}_J']'$ 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 $\mathbf{S}_R^{*-}$. For computational purposes, it should be noted that $V(R)$ is the Pillai's trace (V) scaled by $(N-1)$. That is, a

multivariate profile analysis performed on the Friedman ranks (R_{ijk}) yields a Pillai's trace such that $V_{(R)} = V(N-1)$, which approximates a χ^2 distribution with $df = (J-1)(K-1)$, asymptotically (Beasley, 2000). Thus, this procedure could also be accomplished by computing $\mathbf{R}^*=\mathbf{RD}$, where $\mathbf{R}$ is the (NxK) data matrix for the Friedman model ranks, and then substitute $\mathbf{Y}^*$ with $\mathbf{R}^*$ in the multivariate model (4). As an alternative approach to this statistic proposed by Beasley (2000), the Hotelling's trace could be used, thus calculating $H_{(R)}$, the statistic in (11), by replacing A_{ijk} with R_{ijk}. As shown previously, $H_{(Q)}$ could be transformed to an F approximation test by (12) or critical values from the multivariate referent distribution (e.g., Hotelling's trace; Pillai's trace) could be obtained in order to assess statistical significance.

Assumptions and Hypotheses for Interaction Tests Performed on Friedman Ranks

By using the shift model (18) and requiring the univariate model assumptions of IID[$0,\sigma_\varepsilon^2$] for all j and k, a rejection of (18) using the univariate $F_{(R)}$ test (21) implies that (20) is false (i.e., the interaction is due to differences in location parameters). Likewise, requiring the multivariate model assumption that the random error vectors ($\boldsymbol{\varepsilon}_{jk}$) are independent and identically distributed across the J groups for each of the K repeated measures separately (i.e., IID[$\mathbf{0}_{(K\text{-}1)}$,$\mathbf{D'\Sigma D}$]), a rejection of (19) using $V_{(R)}$ implies that (20) is false. However, if these distributional assumptions are not tenable, inferences concerning shifts in location parameters are not credible. Therefore in the strictest sense, the null hypothesis in (20) applied to the Friedman model ranks implies the equality of ranking patterns across groups, which would involve a Chi-square test of homogeneity of ranking distributions in a JxK! contingency table. Analogous to the null hypotheses for aligned ranks, (20) does not imply that the probabilities of occurrence for each permutation of the ranks are equal in value across groups.

To elaborate, the univariate model null hypothesis of permutational equivalence (16) and the multivariate model null hypothesis of distributional equivalence (17) can be formulated in terms of the probability of ranking patterns for R_{ijk}. Let ϕ_r be the r^{th} permutation of the K Friedman ranks (r = 1, … K!). Let π_{rj} be the probability of the r^{th} permutation for subjects in the j^{th} group. Because the average rank for each individual equals (K+1)/2, the null hypothesis in (20) can be expressed in a form similar to (5):

$$H_0(J\text{x}K): \Delta_1 = \ldots = \Delta_j = \ldots = \Delta_J, \qquad (23)$$

where,

$$\Delta_j = \sum_{r=1}^{K!} \pi_{rj}\,\phi_r.$$

Thus, consistent with the null hypothesis in (16), the univariate $F_{(R)}$ statistic approximates a chi-square distribution with $df = (J-1)(K-1)$ under the null hypothesis:

$H_0(J\text{x}K)$:

$$\pi_{rj} = 1/K!, \text{ for } r = 1, \ldots K! \text{ and } j = 1, \ldots J. \qquad (24)$$

Therefore, $F_{(R)}$ (21) does not necessarily provide a test of (20) because a false (24) does not imply a false (20). It is also important to recognize that if (24) is true so are (16), (17), and (20), but (20) does not imply (24). That is, it is possible to have identical mean ranks without each permutation of ranks occurring with the same frequency. Therefore, using $F_{(R)}$ as an approximate test may occasionally reject (20) incorrectly because (24) is false.

Likewise, $V_{(R)}$ does not necessarily test the null hypothesis (20). The null hypothesis actually tested by $V_{(R)}$ is:

$H_0(J\text{x}K)$:

$$\pi_{r1} = \ldots = \pi_{rj} = \ldots = \pi_{rJ} \text{ for } r = 1, \ldots K! \qquad (25)$$

The asymptotic distribution of $V_{(R)}$ is χ^2 with $df = (J-1)(K-1)$ under (25) but not necessarily under (20). As with the univariate $F_{(R)}$ test, it is

important to recognize that if (25) is true so is (20), but (20) does not imply (25). That is, it is possible for two groups to have identical mean ranks but different permutational distributions. Therefore, using $V_{(R)}$ as an approximate test may occasionally reject (20) incorrectly because (25) is false.

It should be noted that if the univariate null hypothesis (24) is true so is the multivariate null hypothesis (25). However, if (25) is true, it does not imply that (24) is true. Likewise, a false (24) does not imply a false (25). Thus, the univariate $F_{(R)}$ and the multivariate $V_{(R)}$ statistics test two distinctly different, although conceptually related, hypotheses concerning the similarity of ranking patterns among multiple groups. Table 4 shows various scenarios in which these null hypotheses are true or false in a (J=2)x(K=3) split-plot design.

The multivariate model null hypothesis (25) is less restrictive than the univariate model null hypothesis (24) because $F_{(R)}$ uses a fixed covariance structure (i.e., $K(K+1)/12$) in the denominator (Marascuilo & McSweeney, 1967), thus implying compound symmetry of the covariance matrix. Thus, the null hypothesis in (24) implies sphericity because it translates to the assumption that the errors are IID[0, σ_ε^2] for all j and k from the univariate model null hypothesis in (16).

Similarly, the null hypothesis in (25) translates into relaxing the assumption that all K repeated measures have identical distributions. This is analogous to the multivariate model null hypothesis in (17), which only assumes the random error components are independent and identically distributed across the J groups for each of the k measures separately (i.e., IID[$\mathbf{0}_{(K\text{-}1)}$, $\mathbf{D'\Sigma D}$]; Hollander & Wolfe, 1973, p. 145). Thus, $V_{(R)}$ as a multivariate test of the null hypothesis in (25) does not assume sphericity of the covariance matrix. This is because under the null hypothesis in (25) each group is not required to have π_{rj} = 1/K!, which implies a fixed covariance structure and thus sphericity.

If it is tenable to assume that the errors are IID[$0,\sigma_\varepsilon^2$] for all j and k, then rejections of (24) using the univariate $F_{(R)}$ imply an interaction due to location parameters (i.e., a false 20). Likewise, rejections of (25) using the multivariate $V_{(R)}$ imply a false (20) if the errors are assumed to be IID[$\mathbf{0}_{(K\text{-}1)}$, $\mathbf{D'\Sigma D}$].

Although the univariate (24) and multivariate (25) null hypotheses for Friedman ranks can be expressed by different formulations than the univariate (18) and multivariate (19) null hypotheses for the shift model for aligned ranks, the concept of stochastic homogeneity applies to the Friedman ranks (Randles & Wolfe, 1979; Vargha & Delaney, 1998). However, if the additional distributional assumptions are not met, these statistics based on Friedman model ranks should strictly be considered test of stochastic homogeneity (Beasley, 2000; Serlin & Harwell, 2001; Vargha & Delaney, 1998).

Computational Example One

Table 1 shows hypothetical data and sample moments for a J=2 groups by K=3 repeated measures design. An educational psychology research application of this design could be a comparison of the forgetting rates over a three week period (e.g., recall measured at 7, 14, and 21 days) for children classified as slow (j=1) or fast (j=2) learners (e.g., Gentile, Voelkl, Mt. Pleasant, & Monaco, 1995). A medical psychology application would be a comparison of the addiction severity scores of opioid-dependent patients in a Day Treatment program (j=1) versus patients in an Enhanced Standard Methadone program (j=2) at three time points: Pre-treatment, Post-treatment, and Follow-up (e.g., Avants, Margolin, Sindelar, & Rounsaville, 1999).

Analyses of these data using the univariate model (1) show that the between-subjects effect was statistically significant, $F_{(Y)(1,16)}$ = 6.27, p = .023. The covariance structure was non-spherical with a Greenhouse-Geisser epsilon estimate of .681. The Huynh-Feldt correction results in an epsilon estimate of .769. After a Huynh-Feldt correction to the dfs, both the repeated measures main effect [$F_{(Y)(1.54,24.61)}$ = 194.22, p < .001] and the interaction effect [$F_{(Y)(1.54,\ 24.61)}$ = 12.20, p = .001] were statistically significant. A multivariate profile analysis yielded similar findings. Both the Pillai's trace ($V_{(Y)}$ = 0.936)

and Hotelling's trace ($H_{(Y)}$ = 14.706) for the repeated measures main effect were statistically significant (p <.001). For the interaction effect, both the Pillai's trace ($V_{(Y)}$ = 0.494) and Hotelling's trace ($H_{(Y)}$ = 0.977) were statistically significant (p = .006) also.

Examining the moments for each of the JK=6 cells in Table 1, it is apparent that the data are skewed for many cells, thus potentially violating the normality assumptions of both the univariate (1) and multivariate (4) models. This provides a reason for employing rank-based tests. However, given that both the repeated measures and between-subjects main effects were statistically significant, it is necessary to align the data before ranking and subsequent analysis.

Table 1 also shows the aligned data (8) and the aligned ranks (9). Analysis of the aligned ranks showed a statistically significant interaction using the univariate model [$F_{(A)(2,32)}$ = 16.33, p < .001]. The Greenhouse-Geisser epsilon estimate was .839 and the Huynh-Feldt correction was .984. Thus, any correction to the dfs would not affect statistical significance. The multivariate approach yielded a statistically significant Hotelling's trace [$H_{(A)}$ = 1.426 from (11)], which multiplied by (N-1)=17 yields a chi-square approximate statistic of $\chi^2_{(A)(df=2)}$ = 24.242, p < .001. Converting $H_{(A)}$ to an F approximate using (12) yields $F_{H(A)(2,15)}$ = 10.697, p = .001.

Table 2 shows the Koch (1969) model of alignment and ranking. As was the case with the aligned ranks, the results show a statistically significant interaction with a Pillai's trace of $V_{(Q)}$ = 0.574 from (15), which multiplied by (N−1) = 17 yields a Chi-square approximate statistic of $\chi^2_{(Q)(df=2)}$ = 9.758, p < .01. The Hotelling's trace for the Koch model ranks was $H_{(Q)}$ = 1.345 with an F approximate (12) of $F_{H(Q)(2,15)}$ = 10.091, p = .002.

Table 3 shows the aligned data and the Friedman (1937) model of ranking applied to the aligned data. As was the case with the aligned ranks and the Koch ranks, the results show a statistically significant interaction. Analyzing a univariate model and calculating the multiple group extension of the Friedman (1937) statistic (21) yields [$F_{(R)(df=2)}$ = 15.239, p < .001]. The Huynh-Feldt correction of the Greenhouse-Geisser estimate of epsilon was 1.0. Thus, there are no corrections to the dfs. The multivariate approach yielded a statistically significant Pillai's trace of $V_{(R)}$ = 0.624 from (22), which multiplied by (N−1) = 17 yields a Chi-square approximate statistic of 10.608, p < .005. The Hotelling's trace for the Friedman model aligned ranks was $H_{(R)}$ = 1.657 with an F approximate (12) of $F_{H(R)(2, 15)}$ = 12.426, p = .001.

By further examination of the six cells in Table 1, the data at time k = 1 are positively skewed with similar means, variances, and kurtosis values for both groups. At time k = 2, the data for both groups are symmetric with similar variances, but group j = 2 has a higher mean. At time k = 3, there are still location differences, but the data for both groups are negatively skewed with similar variances and kurtosis.

In analyzing real data, it is difficult to trust sample statistics for skew and kurtosis, especially for small sample sizes. Therefore, judging whether the IID assumptions are tenable presents a conundrum. Although such practice is not advised, for the sake of illustration, suppose that these sample moments are valid estimates of population parameters. This data pattern then illustrates a situation in which there is a violation of the univariate shift model (18) distributional assumptions (i.e., IID[$0,\sigma^2_\varepsilon$] for all j and k); however, the multivariate shift model (19) assumption (i.e., IID[$\mathbf{0}_{(K\text{-}1)}$,$\mathbf{D'\Sigma D}$]) seems tenable. That is, the univariate model requires that all six cells have identical distribution functions; whereas, the multivariate model only requires the two groups to have identical distribution functions for each of the K = 3 measures separately. Given that all three multivariate aligned rank tests led to rejections of the interaction null hypothesis in (17), the interaction can be attributed to shifts in location parameters (i.e., a false 20). Furthermore, one may conclude that the stochastic dominance of one group over the other was not constant across the K = 3 repeated measures.

Computational Example Two

Table 5 shows the sample moments and the univariate and multivariate test statistics for the Original Data, Aligned Ranks, Koch Model Ranks, and Friedman Model Ranks for hypothetical data from $J = 3$ groups by $K = 4$ repeated measures design (see Appendix for data). A medical psychology research application of this design could be a comparison of the number of errors in recall over $K = 4$ trials for men with treated blood pressure elevation ($j = 3$), men with untreated elevated blood pressure ($j = 2$), and a group of normotensive males ($j = 1$) (e.g., Waldstein, et al., 1991). A genetic association research application would be an alcohol sensitivity study in which motor coordination of humans with $J = 3$ different genotypes (e.g., aa, AA, Aa) was measured once before ($k = 1$) and three times after ingesting a standard dose of alcohol (e.g., Boomsma, Martin, & Molenaar, 1989).

Suppose that these sample moments are valid estimates of population parameters, then upon examination of the Original Data, it can be seen that Group One has positively skewed data with minor changes in spread (variance) and location (mean and median) across the four measures. Similarly, Group Three also has positively skewed data with minor changes in variance over time. However, Group Three also exhibits significant increases in location over the four time periods. Thus, if this example only included Groups One and Three, even the more restrictive distributional assumptions of the univariate shift model (18) would be tenable. That is, the eight cells for Groups One and Three have similar variance and shape (i.e., IID$[0,\sigma_\varepsilon^2]$ for all j and k) and differ only in location.

By contrast, Group Two has data that is positively skewed initially ($k = 1$). Subsequently, Group Two increases in location, fluctuates in spread, and changes from a positively skewed shape at $k = 1$ to a symmetric shape at $k = 2$ and then to a negatively skewed shape at the third and fourth measures. In comparing Group Two to the other groups, neither the univariate (18) nor the multivariate shift model (19) distributional assumptions are met. Therefore, the significant test statistics that result in rejections of the null hypotheses (16) or (17) cannot be attributed to a single parameter. Thus, the rejection must be interpreted as the groups demonstrating stochastic heterogeneity in trends (growth curves). Namely, Group Two appears to be stochastically dominant over the other two groups at time points $k = 2$ and 3 and stochastically dominant over Group One at $k = 4$; however, contrast procedures are necessary to test this interpretation.

Multiple Comparison Procedures for Aligned Rank Procedures

Given that the three rank-based procedures are viable approaches to analyzing repeated measures data, then contrast procedures based on these methods should hold quite generally (Agresti & Pendergast, 1986; Beasley, 2000, 2002; Koch, 1969). The most typical form is a product interaction contrast (Hochberg & Tamhane, 1987, pp. 294-303; Marascuilo & Levin, 1970) defined as:

$$\hat{\psi} = a_1(b_1\bar{U}_{11} + b_2\bar{U}_{12} + \ldots + b_k\bar{U}_{1k} + \ldots + b_K\bar{U}_{1K}) + a_2(b_1\bar{U}_{21} + b_2\bar{U}_{22} + \ldots + b_k\bar{U}_{2k} + \ldots + b_K\bar{U}_{2K}) + a_j(b_1\bar{U}_{j1} + b_2\bar{U}_{j2} + \ldots + b_k\bar{U}_{jk} + \ldots + b_K\bar{U}_{jK}) + a_J(b_1\bar{U}_{J1} + b_2\bar{U}_{J2} + \ldots + b_k\bar{U}_{Jk} + \ldots + b_K\bar{U}_{JK}); \quad (26)$$

where $\bar{U}_{jk}$ is a general term for the mean rank of the j^{th} group on the k^{th} repeated measure.

Define $\mathbf{a} = (a_1 + a_2 + \ldots + a_j \ldots + a_J)'$ as a vector of contrast coefficients that compares the J independent samples and $\mathbf{b} = (b_1+b_2+\ldots+b_k+\ldots+b_K)'$ as a vector of contrast coefficients 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). In some cases, it may be desirable to normalize the trend coefficients, **b**, so that the metric of the repeated measures variable will not

change, thus making confidence intervals more interpretable.

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:

$$SE^2_{\hat{\psi}} = \sum_{j=1}^{J} \left(\frac{a_j^2}{n_j}\right) \frac{(\mathbf{b}' \mathbf{E} \mathbf{b})}{(N - J)}, \qquad (27)$$

where **E** is the error matrix (4) computed for U_{ijk} (i.e., any of the three ranking procedures). This approach assumes homogeneity of variance of the transformed scores:

$$U^*_{ij} = \sum_{k=1}^{K} b_k U_{ijk}. \qquad (28)$$

This requirement of homogeneity of variance for transformed scores implies the sphericity of the pooled covariance matrix (4). Thus from the perspective of rank-based tests, this approach requires that the error components are IID$[0, \sigma^2_\varepsilon]$ for all j and k.

From a multivariate perspective, a standard error that does not require homogeneity of variance of the transformed scores (i.e., sphericity) can be calculated by defining J separate Sums of Squares (SS):

$$\mathrm{SS}_{U^*_j} = \sum_{i=1}^{n_j} (U^*_{ij} - \bar{U}^*_j)^2, \qquad (29)$$

where $\bar{U}^*_j$ is the mean for the j^{th} group for the transformed scores U^*_{ij} in (29). The standard error is calculated as:

$$SE^2_{\hat{\psi}} = \sum_{j=1}^{J} \left(\frac{a_j^2}{n_j}\right) \frac{\mathrm{SS}_{U^*_j}}{(n_j - 1)}, \qquad (30)$$

A $(1-\alpha)$% confidence interval for the contrast of aligned ranks can be formed by:

$$\hat{\psi} \pm S\,(SE_{\hat{\psi}}). \qquad (31)$$

The null hypothesis H_0: $\psi = 0$ is rejected if the confidence interval in (31) does not cover zero. If the univariate IID$[0, \sigma^2_\varepsilon]$ assumption is tenable, $SE_{\hat{\psi}}$ can be defined as the square root of (26). However, $SE_{\hat{\psi}}$ should be defined as the square root of (30) if the transformed scores have heterogeneous variances (i.e., the sphericity condition does not hold).

The definition of S depends on the type of contrast conducted. For example, in the $J = 3$ by $K = 4$ design from Example Two, suppose that after rejecting the null hypothesis (17) the interest was in assessing whether the linear trend, $\mathbf{b}'_L = \{-3\ -1\ +1\ +3\}/\sqrt{20}$, of Group One is stochastically different from the linear trend of the other two groups combined, $\mathbf{a}'_1 = \{+1\ -0.5\ -0.5\}$, and whether the linear trends for Groups Two and Three are stochastically different, $\mathbf{a}'_2 = \{0\ +1\ -1\}$. In this case, the trend coefficients, $\mathbf{b}_L$, were normalized so that the metric of the repeated measures variable was not changed, thus making subsequent confidence intervals more interpretable.

Also, consider the same group comparisons for the Initial Change from Time $k = 1$ to Time $k = 2$, $\mathbf{b}'_C = \{-1\ +1\ \ 0\ \ 0\}$. Thus, $c = 4$ post hoc tests would be conducted. To construct a post hoc confidence interval, S could be defined as a critical value from Student's t distribution using the Dunn-Sidák correction, $\alpha_{DS} = [1-(1-\alpha)^{1/c}]/2$:

$$S = t_{(1-\alpha_{DS}),df_e}. \qquad (32)$$

For $c = 4$ contrasts, $\alpha_{DS} = .00637$; however, df_e for (32) differs for the univariate (27) and multivariate approaches (30). For the univariate pooled standard error (27), $df_e = (N-J)$; however, if the standard error in (30) is used then a Welch (1947) correction must be applied to df_e. For defining S in terms of the sampling distribution of the Hotelling's trace or other multivariate referent distribution, refer to Gabriel (1968) and Sheehan-Holt (1998).

For computational convenience, the interaction contrasts can be calculated by

transforming the data into a single variable: **Ub**, where **U** is the *N*x*K* data matrix and **b** is the *K*x1 vector of trend coefficients. Then, the group contrasts, **a**, can be performed on the transformed data. The univariate pooled standard error (27) can be computed from methods that assume equal variances, such as Fisher's LSD. The multivariate standard error (30) can be computed from methods that do not assume equal variances, such as Tamhane's (1979) *T*2.

It is debatable whether the multivariate (30) or univariate (27) 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 is non-spherical (Boik, 1981).

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).

Defining Confidence Intervals for Interpretable Parameters

Reasons for rejecting an interaction null hypothesis are of more interest than the simple conclusion that it is false; therefore, the contrast testing procedures detailed in the previous section are of great utility. Furthermore, there is a trend toward interpreting confidence intervals instead simply reporting *p*-values in a variety of research disciplines (Campbell & Gardner, 1988; Gardener & Altman, 1986; Serlin, 1993). Moreover, it is important to construct confidence intervals around interpretable parameters when possible. Thompson (2002) discusses a bootstrap methodology to compute confidence intervals for effect sizes from parametric analyses. For location parameters less sensitive to skewness, confidence intervals for medians have been proposed (Bonett & Price, 2002; Campbell & Gardner, 1988; Hodges & Lehmann, 1963).

Unfortunately, aligned ranks have no inherent meaning except that they serve as placeholders for the percentiles of the original raw score distribution with the nuisance location parameters removed. Thus, the rank statistics previously discussed are useful for assessing the statistical significance of the interaction, but they do not provide direct information about the nature or magnitude of the effect. For this reason, Koch, et al. (1980) suggested that results from nonparametric omnibus tests should be accompanied by appropriate descriptive statistics (e.g., frequency distributions or percentiles) and nonparametric estimates for confidence intervals. Newson (2002) reviewed methods for computing confidence intervals for rank-based statistics, which convey estimates and boundaries for informative parameters such as Cliff's (1996) *d* and Somers' (1962) *D*.

Confidence Intervals for Aligned Ranks

The cell means for the aligned ranks provide descriptions of the degree to which the *JK* cells have different locations due to discrepancies from the marginal distributions (i.e., due to interaction). Thus, these cell means give information about interaction trends relative to main effects and which cells contribute more to the omnibus interaction effect. For repeated measures designs, Agresti and Pendergast (1986) suggested dividing ranks by (*NK*+1). These values, $U_{ijk} = A_{ijk}/(NK+1)$, have a grand mean, $\bar{U}_{**}=0.5$, that is equivalent to the median of the aligned scores. The cell means, $\bar{U}_{jk}$, provide the probability that a randomly selected observation from cell *jk* is larger than an independent observation selected at random from another cell after removing the main effects. This approach suggested by Agresti and Pendergast (1986) is consistent, though not identical, to Cliff's (1996) notion of dominance[1] and the computation of relative effects[2] (Brunner, et al., 2002). It is also similar to the Hodges and Lehmann (1963) median difference,

which estimates the typical difference between individual observations from different cells.

As noted, the interaction contrasts can be accomplished by transforming the data, **Ub**, and performing the group contrasts, **a**, on the transformed data. The upper panel of Table 6 shows the means and standard deviations for **U** = **A**/(*NK*+1), the data transformed by the linear trend contrast, $\mathbf{Ub}_{\mathrm{L}}$, and the data differenced by the Initial Change contrast, $\mathbf{Ub}_{\mathrm{C}}$. The upper panel of Table 7 shows the univariate-based (27) and multivariate-based (30) 95% confidence intervals for the four contrasts previously discussed performed on the adjusted aligned ranks.

The cell mean for Group 1 at time $k = 1$ had the highest mean of 0.9476. This indicates that, after removal of the main effects, this cell had higher scores relative to the other cells and that a randomly selected observation from this cell has a very high probability (0.9476) of being larger than an independent observation selected at random from any other cell. Likewise, the cell mean for Group 1 at time $k = 4$ had the lowest mean of 0.1012, and thus, a randomly selected observation from this cell has a very low probability of being larger than an independent observation selected at random from any other cell.

Similar to Cliff's (1996) *d*-statistic, the difference in these probabilities can be used to judge the stochastic dominance of one cell over another. Thus, the aligned ranks for Group 1 have a descending trend in that relative to the main effects the observations in Group 1 tend to get stochastically smaller over time. By examining the original data in Table 5, Group 1 had a slight increase in means across the $K = 4$ time points. Therefore, the aligned ranks provide information about which cells have stochastically larger scores relative to the main effects. In other words, given that there was a repeated measures main effect with increasing means for all three groups combined, the trend for Group 1 was descending in a relative manner. This can be seen in the data transformed by the linear contrast coefficient, $\mathbf{Ub}_{\mathrm{L}}$, in which the probability of larger scores (i.e., stochastic dominance) for observations in Groups 1 tends to decrease at a rate of -.620 on average.

For Group 2, the probability of larger scores tends to increase at average rate of .336 relative to the main effects. For Group 3, the stochastic dominance of scores relative to the main effects increases at a slight lower rate (.202) as compared to Group 2. For comparing Group 1 to Groups 2 and 3 combined, the results show a value of $\widehat{\psi}_{\mathbf{a}_1\mathbf{b}_{\mathrm{L}}} = -0.8891$. This indicates that Groups 2 and 3 combined, as compared to Group 1, have a very high probability of having stochastic larger scores at time $k = 4$ and smaller scores at $k = 1$. To elaborate, suppose Case A is a randomly selected case from Group 2 or 3 and Case B is a randomly selected case from Group 1. The probability that Case A will have a steeper ascending (positive monotonic) trend across the $K = 4$ time points than Case B from Group 1 is 0.8891.

The univariate 95% simultaneous confidence interval indicates that plausible values range between -1.1276 and -0.6506. The multivariate 95% simultaneous confidence interval gives a tighter band of plausible values that range between -1.0474 and -0.7308. Note that the sign of the contrast value only indicates the direction of the stochastic dominance; it does not indicate a negative probability. Also, this approach can yield a bound on the confidence interval that exceeds 1 (-1 in this case), thus, an asymmetrical confidence interval with 1 (or -1) as the upper (or lower) bound may be constructed. Other methods create this bound and asymmetrical confidence interval by computing the standard errors in a different manner (see Endnotes 1 and 2; Brunner, et al., 2002; Cliff, 1996; Newson, 2001). The difference between Groups 2 and 3 is not statistically significant: both the univariate and multivariate 95% confidence intervals contained zero as a plausible value (see Table 7).

By examining the data transformed by the initial change contrast coefficient, $\mathbf{Ub}_{\mathrm{C}}$, it is observed that observations from time $k = 1$ tend to be stochastically larger than observations taken at $k = 2$, for Groups 1 and 3. For Group 2, the measures taken at $k = 2$ are stochastically larger than the scores from $k = 1$ and the probability of randomly selecting a larger score at $k = 2$ increases by 0.3657 relative to the main

effects. Thus, Group 2 has a tendency for scores to become stochastically larger from $k = 1$ to $k = 2$; whereas, Groups 1 and 3 have a tendency for scores to decrease relative to the main effects.

As compared to Group 1, Groups 2 and 3 combined have a higher probability of scores becoming stochastically larger from time point $k = 1$ to $k = 2$, $\hat{\psi}_{\mathbf{a}]\mathbf{b}C} = -0.4824$. The univariate 95% simultaneous confidence interval indicates that plausible values range between -0.7236 and -0.2385. The multivariate 95% simultaneous confidence interval gives a tighter band of plausible values that range between -0.6531 and -0.3117. The contrast of Group 2 with Group 3 is statistically significant; thus, the probability that Group 2 has stochastically larger scores at $k = 2$ relative to $k = 1$ as compared to Group 3 is 0.8552.

To elaborate, suppose a randomly selected case from Group 2 and a randomly selected case from Group 3. The probability that the case selected from Group 2 will have a stochastically larger gain from time $k = 1$ to $k = 2$ as compared to the latter case from Group 3 is 0.8552. The univariate 95% simultaneous confidence interval indicates that plausible values range between .5833 and 1.1271. The multivariate 95% simultaneous confidence interval gives a wider band of plausible values that range between 0.4779 and 1.2326. As with previous analyses, a researcher may choose to construct an asymmetrical confidence interval with 1 as the upper bound or use other methods that compute standard errors in a different manner (Brunner, et al., 2002; Cliff, 1996; Newson, 2001).

Confidence Intervals for Koch Model Ranks

Using the logic of Agresti and Pendergast (1986), the Koch ranks can be transformed by:

$$U_{ijk} = [Q_{ijk} - [((N+1)/2)]]/[(K-1)(N+1)].$$

These values have a grand mean of 0.5. The cell means provide descriptions of the degree to which the JK cells have different locations due to discrepancies from the marginal distributions. As shown in Table 6, the cell mean values for the Koch ranks (middle panel) are similar to the aligned rank cell means (upper panel). Thus, it would seem that the Koch ranks could be interpreted in a similar manner, but whether they represent probabilities in the same sense that the aligned ranks is debatable.

In Table 7, note that the Koch model tends to give lower estimates of the contrast effects with smaller standard errors, thus, one ay question the statistical power of the Koch model relative to the aligned rank procedure. For identically skewed (i.e., multivariate exponential) error distributions, Tandon and Moeschberger (1989) found the Koch model to have similar power as parametric procedures, whereas, Beasley (2002) found the aligned rank procedure to have more statistical power than parametric tests for interactions. It is debatable whether these differences are due to estimation bias, violations of assumptions, or differences in statistical power.

Confidence Intervals for Friedman Ranks

A different logic is used to standardize the Friedman Ranks:

$$U_{ijk} = [R_{ijk} - [((K+1)/2)]]/[(K^2-1)/12].$$

For each subject, U_{ijk} has a mean of 0 and unit variance, which is similar in concept to Hettmansperger's (1984) standardization of ranks. As previously noted, the interaction contrasts can be accomplished by transforming the data, **Ub**, and then performing the group contrasts, **a**, on the transformed data. The lower panel of Table 6 shows the means and standard deviations for $\mathbf{U}=[R_{ijk}-[((K+1)/2)]]/[(K^2-1)/12]$. To transform the data by the linear trend contrast, $\mathbf{b}_L$ is standardized, rather than normalized, so that it also has a variance of one, rather than a sum of squares of one, $\mathbf{b}'_L=\{-1.3416-0.4472+0.4472+1.3416\}$. The values of $\mathbf{U}\mathbf{b}_L/K$ are a linear transformation of Page's (1963) L statistic and represent each individual's rank correlation with the linear trend coefficients (Lyerly, 1952). Thus, the mean values of $\mathbf{U}\mathbf{b}_L/K$ for each group represent the group's average concordance with the ordered alternative, in this case linear trend. The contrasts, **a**, applied to these values will estimate how the groups differ

Table 1: Hypothetical Data and the Aligned Ranking Procedure for the $J = 2$ by $K = 3$ Split-Plot Design in Example One.

	Original Data				Aligned Data			Aligned Ranks		
	$k = 1$	$k = 2$	$k = 3$	$\bar{Y}_{ij*}$	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
Group One $j = 1$ Slow Learners or Day Treatment	1.1	6.2	7.2	4.83	.56	-.03	-.53	39	28	18
	2.2	4.8	6.1	4.37	2.13	-.96	-1.16	53	10	8
	2.3	7.1	8.0	5.80	.79	-.10	-.70	44	24	15
	2.4	8.1	9.4	6.63	.06	.07	-.13	30	31	23
	3.2	7.3	10.4	6.97	.53	-1.06	.54	37	9	38
	3.4	9.3	10.5	7.73	-.04	.17	-.13	26	34	22
	4.1	8.1	9.3	7.17	1.23	-.46	-.76	48	20	14
	10.1	10.4	10.2	10.23	4.16	-1.23	-.2.93	54	7	1
Mean	3.60	7.66	8.89	6.72	1.18	-.45	-.73	41.38	20.38	17.38
Median	2.80	7.70	9.35	6.80	.68	-.28	-.61	41.5	22.00	16.50
SD	2.78	1.75	1.62	1.84	1.39	.56	1.03	10.25	10.60	11.03
Variance	7.72	3.05	2.64	3.37	1.93	.32	1.06	105.13	112.27	121.70
Skew	2.24	-.06	-.76	.75	1.69	-.36	-1.47	-.23	-.12	.54
Kurtosis	5.65	-.09	-.75	1.08	2.88	-1.97	3.22	-1.22	-1.88	1.17
Group Two $j = 2$ Fast Learners or Enhanced Standard Methadone	1.0	7.9	8.8	5.90	-.61	.60	0	16	41	29
	2.4	9.2	10.1	7.23	-.54	.57	-.03	17	40	27
	2.2	10.1	11.8	8.03	-1.54	.67	.87	3	42	45
	2.3	10.9	11.1	8.10	-1.51	1.40	.10	4	50	33
	3.1	10.1	13.2	8.80	-1.41	-.10	1.50	5	25	51
	3.3	9.9	12.1	8.43	-.84	.07	.77	12	32	43
	3.2	11.2	14.4	9.60	-2.11	.20	1.90	2	35	52
	4.4	12.3	13.1	9.93	-1.24	.97	.27	6	46	36
	4.9	11.2	14.2	10.10	-.91	-.30	1.20	11	21	47
	9.2	13.1	14.3	12.20	1.29	-.50	-.80	49	19	13
Mean	3.60	10.59	12.31	8.83	-.94	.36	.58	12.50	35.10	37.60
Median	3.15	10.50	12.60	8.62	-1.07	.39	.52	8.50	37.50	39.50
SD	2.26	1.50	1.89	1.74	.92	.59	.82	13.90	10.63	12.36
Variance	5.11	2.24	3.59	3.03	.85	.35	.67	193.17	112.99	152.71
Skew	1.85	-.06	-0.63	.31	1.63	.26	.05	2.35	-.32	-.74
Kurtosis	4.35	.21	-0.50	.76	3.90	-.56	-.53	6.22	-1.18	.075
Epsilon*	.769				.769			.984		

Note: * Based on the Huynh-Feldt adjustment of the Greenhouse-Geisser estimate of epsilon from the pooled within-group covariance matrix.

Table 2: Hypothetical Example of Koch's Model of Ranking for Interactions for Hypothetical Data in Table 1.

	Koch's Model for Analyzing Interaction Effects								
	$T_{ij}(1,1)$	$T_{ij}(1,2)$	$T_{ij}(1,3)$	$T_{ij}(2,1)$	$T_{ij}(2,2)$	$T_{ij}(2,3)$	$T_{ij}(3,1)$	$T_{ij}(3,2)$	$T_{ij}(3,3)$
Group One $j = 1$ Slow Learners or Day Treatment	9.5	12	13	7	9.5	12	6	7	9.5
	9.5	17	17	2	9.5	8	2	11	9.5
	9.5	13	14	6	9.5	14	5	5	9.5
	9.5	11	12	8	9.5	7	7	12	9.5
	9.5	14.	10	5	9.5	2	9	17	9.5
	9.5	10	11	9	9.5	11	8	8	9.5
	9.5	15	15	4	9.5	9.5	4	9.5	9.5
	9.5	18	18	1	9.5	18	1	1	9.5
Group Two $j = 2$ Fast Learners or Enhanced Standard Methadone	9.5	6	8	13	9.5	14	11	5	9.5
	9.5	7	9	12	9.5	14	10	5	9.5
	9.5	4	3	15	9.5	6	16	13	9.5
	9.5	1	5.5	18	9.5	17	13.5	2	9.5
	9.5	5	2	14	9.5	3	17	16	9.5
	9.5	8	5.5	11	9.5	5	13.5	14	9.5
	9.5	2	1	17	9.5	1	18	18	9.5
	9.5	3	7	16	9.5	16	12	3	9.5
	9.5	9	4	10	9.5	4	15	15	9.5
	9.5	16	16	3	9.5	9.5	3	9.5	9.5
	$Qij1$			$Qij2$			$Qij3$		
Group One $j = 1$ Slow Learners or Day Treatment	34.5			28.5			22.5		
	43.5			19.5			22.5		
	36.5			29.5			19.5		
	32.5			24.5			28.5		
	33.5			16.5			35.5		
	30.5	$\bar{Q}_{11}$	37.00	29.5	$\bar{Q}_{12}$	24.94	25.5	$\bar{Q}_{13}$	23/56
	39.5	$SD(Q_{11})$	5.37	23.0	$SD(Q_{12})$	4.95	23.0	$SD(Q_{13})$	6.92
	45.5	$\text{Var}(Q_{11})$	28.86	28.5	$\text{Var}(Q_{12})$	24.53	11.5	$\text{Var}(Q_{13})$	47.89
Group Two $j = 2$ Fast Learners or Enhanced Standard Methadone	23.5			36.5			25.5		
	25.5			35.5			24.5		
	16.5			30.5			38.5		
	16.0			44.5			25.0		
	16.5			26.5			42.5		
	23.0			25.5			37.0		
	12.5			27.5			45.5		
	19.5	$\bar{Q}_{21}$	21.70	41.5	$\bar{Q}_{22}$	31.35	24.5	$\bar{Q}_{23}$	32.45
	22.5	$SD(Q_{21})$	8.08	23.5	$SD(Q_{22})$	7.76	39.5	$SD(Q_{23})$	8.93
	41.5	$\text{Var}(Q_{21})$	65.34	22.0	$\text{Var}(Q_{22})$	60.23	22.0	$\text{Var}(Q_{23})$	79.75

Table 3: Friedman Model of Aligned Ranks for Hypothetical Data in Table 1.

	Aligned Data			Friedman Aligned Ranks		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
Group One $j = 1$ Slow Learners or Day Treatment	.56	-.03	-.53	3	2	1
	2.13	-.96	-1.16	3	2	1
	.79	-.10	-.70	3	2	1
	.06	.07	-.13	2	3	1
	.53	-1.06	.54	2	1	3
	-.04	.17	-.13	2	3	1
	1.23	-.46	-.76	3	2	1
	4.16	-1.23	-.2.93	3	2	1
Mean	1.18	-.45	-.73	2.625	2.125	1.250
Median	.68	-.28	-.61	3.000	2.000	1.000
SD	1.39	.56	1.03	.518	.641	.707
Variance	1.93	.32	1.06	.268	.411	.500
Skew	1.69	-.36	-1.47	-.644	-.068	2.828
Kurtosis	2.88	-1.97	3.22	-2.240	.741	8.000
Group Two $j = 2$ Fast Learners or Enhanced Standard Methadone	-.61	.60	0	1	3	2
	-.54	.57	-.03	1	3	2
	-1.54	.67	.87	1	2	3
	-1.51	1.40	.10	1	3	2
	-1.41	-.10	1.50	1	2	3
	-.84	.07	.77	1	2	3
	-2.11	.20	1.90	1	2	3
	-1.24	.97	.27	1	3	2
	-.91	-.30	1.20	1	2	3
	1.29	-.50	-.80	3	2	1
Mean	-.94	.36	.58	1.200	2.400	2.400
Median	-1.07	.39	.52	1.000	2.000	2.500
SD	.92	.59	.82	.633	.516	.699
Variance	.85	.35	.67	.400	.267	.489
Skew	1.63	.26	.05	3.162	.484	-.780
Kurtosis	3.90	-.56	-.53	10.000	-2.277	-.146
Epsilon*	.769			1.000		

Note: * Based on the Huynh-Feldt adjustment of the Greenhouse-Geisser estimate of epsilon from the pooled within-group covariance matrix.

Table 4: Hypothetical Population Distribution of Probabilities (π_r*) for Friedman Model Ranks with Descriptive Statistics for Each Element (R_{ijk}). in a $J = 2$ by $K = 3$ Split-Plot Design.

Permutation			Probability of r^{th} Permutation					Group Configuration		Status of Null Hypotheses		
R_1	R_2	R_3	π_r1	π_r2	π_r3	π_r4	π_r5	$j = 1$	$j = 2$	H_0 (23)	H_0 (24)	H_0 (25)
1	2	3	$1/6$	$1/12$	$10/24$	$1/12$	$1/6$	π_r1	π_r1	True	True	True
1	3	2	$1/6$	$1/6$	$1/24$	$1/12$	$1/8$	π_r2	π_r2	True	**False**	True
2	1	3	$1/6$	$1/4$	$1/24$	$1/6$	$1/24$	π_r2	π_r3	True	**False**	**False**
2	3	1	$1/6$	$1/4$	$1/24$	$1/6$	$1/24$	π_r4	π_r4	True	**False**	True
3	1	2	$1/6$	$1/6$	$1/24$	$1/4$	$7/24$	π_r4	π_r5	True	**False**	**False**
3	2	1	$1/6$	$1/12$	$10/24$	$1/4$	$1/3$	π_r3	π_r4	**False**	**False**	**False**
		$\overline{R}_1$	2.000	2.000	2.000	2.333	2.333					
		σ^2_{R1}	0.667	0.500	0.917	0.556	0.806					
		$\overline{R}_2$	2.000	2.000	2.000	1.833	1.833					
		σ^2_{R2}	0.667	0.833	0.167	0.639	0.472					
		$\overline{R}_3$	2.000	2.000	2.000	1.833	1.833					
		σ^2_{R3}	0.667	0.667	0.917	0.639	0.556					
		ε	1.000	0.923	0.640	0.992	0.903					

Table 5: Sample Moment and Tests Statistics for Hypothetical Data from the J=3 by K=4 Split-Plot Design in Example Two.

Original Data	Group j = 1 (n_1 = 8) (e.g., Normotensive; aa)				Group j = 2 (n_2 = 10) (e.g., Untreated EBP; AA)				Group j = 3 (n_3 = 8) (e.g., Treated EBP, Aa)			
	k = 1	k = 2	k = 3	k = 4	k = 1	k = 2	k = 3	k = 4	k = 1	k = 2	k = 3	k = 4
Mean	5.80	6.11	6.88	7.21	5.80	11.79	13.29	15.84	5.88	7.05	9.59	13.70
Median	5.10	5.20	5.95	6.55	5.10	11.75	13.70	17.10	5.20	6.20	8.70	13.05
SD	2.63	2.93	2.73	2.86	2.82	1.55	1.77	2.72	2.85	2.91	2.80	2.78
Variance	6.90	8.56	7.43	8.19	7.98	2.40	3.12	7.39	8.15	8.46	7.82	7.70
Skew	2.34	2.45	2.33	2.37	2.09	0.10	-0.75	-0.78	2.13	2.59	2.36	2.40
Kurtosis	5.94	6.44	5.85	6.15	5.19	0.08	-0.65	-1.21	5.45	7.06	5.99	6.28
H-F = .795, $F_{(Y)(4.77,\ 54.86)} = 53.42$, $p < .001$; $H_{(Y)} = 10.06$, $p < .001$; $V_{(Y)} = 1.66$, $p < .001$												
Aligned Ranks												
Mean	99.50	58.25	33.69	10.63	15.20	69.50	69.90	67.65	57.75	22.25	38.75	83.88
Median	99.50	56.50	34.00	10.50	6.50	66.00	72.50	71.00	57.00	21.50	39.00	83.50
SD	2.45	6.88	5.81	3.32	24.53	16.08	18.44	26.67	12.14	4.30	8.89	6.22
Variance	6.00	47.36	33.78	11.05	602.0	258.5	340.0	711.2	147.4	18.50	79.00	38.70
Skew	0	1.09	0.21	-0.36	2.84	0.03	-1.75	0.00	0.01	1.19	-0.09	-0.06
Kurtosis	-1.20	0.37	-0.93	-0.53	8.46	-1.61	4.75	-1.95	-0.86	1.93	-2.20	-1.66
H-F = .893, $F_{(A)(5.36,\ 61.61)} = 43.10$, $p < .001$; $H_{(A)} = 8.50$, $p < .001$; $V_{(A)} = 1.61$, $p < .001$												
Koch Ranks												
Mean	80.25	60.75	47.38	27.63	32.35	62.25	59.55	61.85	54.81	36.94	53.69	70.56
Median	81.00	57.25	49.25	27.50	28.50	63.25	57.50	59.50	55.75	34.50	52.75	70.50
SD	7.16	8.22	8.27	7.66	12.56	10.15	12.37	12.89	7.20	8.17	7.28	7.81
Variance	51.29	67.57	68.41	58.70	157.7	103.0	153.1	166.1	51.78	66.82	53.00	61.03
Skew	-1.02	0.84	-0.35	0.70	1.72	0.08	0.37	0.86	-0.06	0.79	0.14	-0.05
Kurtosis	1.49	0.42	-0.28	0.68	3.39	-0.79	1.34	-0.04	0.01	-0.22	-1.87	-0.51
H-F = 1.00, $F_{(Q)(6,\ 69)} = 30.35$, $p < .001$; $H_{(Q)} = 7.52$, $p < .001$; $V_{(Q)} = 1.55$, $p < .001$												
Friedman Ranks												
Mean	4.00	3.00	2.00	1.00	1.30	2.80	2.90	3.00	2.75	1.00	2.25	4.00
Median	4.00	3.00	2.00	1.00	1.00	2.50	3.00	3.00	3.00	1.00	2.00	4.00
SD	0	0	0	0	0.95	0.92	0.88	0.94	0.46	0	0.46	0
Variance	0	0	0	0	0.90	0.84	0.77	0.89	0.21	0	0.21	0
Skew					3.16	0.47	-1.02	0.00	-1.44		1.44	
Kurtosis					10.00	-1.81	1.83	-2.13	0.00		0.00	
H-F = .931, $F_{(R)(df = 5.59)} = 56.50$, $p < .001$; $H_{(R)} = 8.80$, $p < .001$; $V_{(R)} = 1.60$, $p < .001$												

Note: H-F = Huynh-Feldt adjustment of the Greenhouse-Geisser estimate of epsilon from the pooled within-group covariance matrix.

Table 6: Sample Moment and Tests Statistics for Hypothetical Data from the J=3 by K=4 Split-Plot Design in Example Two.

	Group j = 1 (n_1 = 8) (e.g., Normotensive; aa)				Group j = 2 (n_2 = 10) (e.g., Untreated EBP; AA)				Group j = 3 (n_3 = 8) (e.g., Treated EBP; Aa)			
Aligned Ranks $\mathbf{U}=\mathbf{A}/(NK+1)$												
	k = 1	k = 2	k = 3	k = 4	k = 1	k = 2	k = 3	k = 4	k = 1	k = 2	k = 3	k = 4
Mean	.9476	.5548	.3208	.1012	.1448	.6619	.6657	.6443	.5500	.2119	.3690	.7988
SD	.0233	.0655	.0554	.0317	.2337	.1531	.1756	.2540	.1156	.0410	.0847	.0592
Linear $\mathbf{Ub}_L$	-.6201 *SD* = .0194				.3359 *SD* = .3157				.2020 *SD* = .1132			
Change $\mathbf{Ub}_C$	-.2778 *SD* = .0592				.3657 *SD* = .2199				-.2391 *SD* = .0910			
Koch Ranks $\mathbf{U} = Q_{ijk} - [((N+1)/2)]/[(K-1)(N+1)]$												
	k = 1	k = 2	k = 3	k = 4	k = 1	k = 2	k = 3	k = 4	k = 1	k = 2	k = 3	k = 4
Mean	.8241	.5833	.4182	.1744	.2327	.6019	.5685	.5969	.5100	.2894	.4961	.7045
SD	.0884	.1015	.1021	.0946	.1550	.1253	.1528	.1591	.0888	.1009	.0899	.0965
Linear $\mathbf{Ub}_L$	-.4727 *SD* = .0666				.2369 *SD* = .2184				.1767 *SD* = .1094			
Change $\mathbf{Ub}_C$	-.2407 *SD* = .1848				.3691 *SD* = .1759				-.2207 *SD* = .1504			
Friedman Ranks $\mathbf{U} = Q_{ijk} - [((N+1)/2)]/[(K-1)(N+1)]$												
	k = 1	k = 2	k = 3	k = 4	k = 1	k = 2	k = 3	k = 4	k = 1	k = 2	k = 3	k = 4
Mean	1.3416	.4472	-.4472	-1.3416	-1.0733	.2683	.3578	.4472	.2236	-1.3416	-.2236	1.3416
SD	0	0	0	0	.8485	.8219	.7832	.8433	.41404	0	.41404	0
Linear $\mathbf{Ub}_L/K$	-1.0000 *SD* = 0				.5200 *SD* = .5750				.5000 *SD* = .1852			
Change $\mathbf{Ub}_C/K$ SD	-.3162 *SD* = 0				.4742 *SD* = .4104				-.5535 *SD* = .1464			

Table 7: Results of Contrast Procedures for the $J = 2$ by $K = 3$ Split-Plot Design in Example Two.

Rank (Contrast)		Univariate Approach $df = 23$; $S = 2.701$			Multivariate Approach				
Aligned Ranks	$\widehat{\psi}$	*SE* (27)	Lower Bound	Upper Bound	df^*	*S*	*SE* (30)	Lower Bound	Upper Bound
$\mathbf{a}_1\mathbf{b}_L$	-0.8891	0.0883	-1.1276	-0.6506	12.11	2.920	0.0542	-1.0474	-0.7308
$\mathbf{a}_2\mathbf{b}_L$	0.1339	0.0984	-0.1319	0.3997	11.73	2.936	0.1076	-0.2461	0.5139
$\mathbf{a}_1\mathbf{b}_C$	-0.4824	0.0903	-0.7263	-0.2385	18.22	2.762	0.0618	-0.6531	-0.3117
$\mathbf{a}_2\mathbf{b}_C$	0.8552	0.1007	0.5833	1.1271	12.53	2.903	0.1084	0.4779	1.2326
Koch Ranks	$\widehat{\psi}$	*SE* (27)	Lower Bound	Upper Bound	*df*	*S*	*SE* (30)	Lower Bound	Upper Bound
$\mathbf{a}_1\mathbf{b}_L$	-0.6795	0.0655	-0.8564	-0.5026	20.27	2.732	0.0461	-0.8054	-0.5536
$\mathbf{a}_2\mathbf{b}_L$	0.0602	0.0730	-0.1369	0.2572	13.79	2.861	0.0792	-0.2102	0.3305
$\mathbf{a}_1\mathbf{b}_C$	-0.3150	0.0730	-0.5122	-0.1178	12.06	2.922	0.0758	-0.5365	-0.0935
$\mathbf{a}_2\mathbf{b}_C$	0.5898	0.0813	0.3702	0.8094	15.90	2.806	0.0770	0.3335	0.8461
Friedman Ranks	$\widehat{\psi}$	*SE* (27)	Lower Bound	Upper Bound	*df*	*S*	*SE* (30)	Lower Bound	Upper Bound
$\mathbf{a}_1\mathbf{b}_L$	-1.5100	0.1592	-1.9400	-1.0800	11.24	2.958	0.0966	-1.7957	-1.2243
$\mathbf{a}_2\mathbf{b}_L$	0.0200	0.1774	-0.4591	0.4991	11.24	2.958	0.1933	-0.6696	0.7096
$\mathbf{a}_1\mathbf{b}_C$	-0.2767	0.1123	-0.5800	0.0266	11.82	2.932	0.0685	-0.4775	-0.0759
$\mathbf{a}_2\mathbf{b}_C$	1.0277	0.1251	0.6898	1.3657	11.82	2.932	0.1371	0.5443	1.5111

Notes: From (32) $\alpha_{DS} = .00637$. $\mathbf{a}_1$={+2 -1 -1} is a comparison of Group One to a combination of Groups Two and Three. $\mathbf{a}_2$={0 +1 -1} is a comparison of Groups Two and Three. $\mathbf{b}_L$={-3 -1 +1 +3} is a linear polynomial contrast. $\mathbf{b}_Q$={+1 -1 -1 +1} is a quadratic polynomial contrast. *The *df*s for the Multivariate Approach were computed from the Welch (1947) correction.

in their concordance with the ordered alternative (i.e., linear trend) on average. As shown in the lower panel of Table 6, Group 1 has a perfect negative rank correlation with the linear trend with no variance, which means that relative to the main effects each person in Group 1 had a descending trend or was discordant with the ordered alternative. Groups 2 and 3 had rank correlations with the linear trend (concordance) of approximately 0.50. Comparing Group 1 to Groups 2 and 3 combined, it is apparent that there are strong differences in their average rank correlation, $\hat{\psi}_{\mathbf{a}_1\mathbf{b}_L}$ = -1.510. The univariate 95% simultaneous confidence interval indicates that plausible values range between -1.9400 and -1.0800. The multivariate 95% simultaneous confidence interval gives a tighter band of plausible values that range between -1.7957 and -1.2243.

This type of interpretation can be used for any trend contrast that involves a linear combination of all K repeated measures by thinking of the trend in terms of ordered alternatives. These results can also be couched in terms of stochastic heterogeneity (Beasley, 2000; Vargha & Delaney, 1998) in that Groups 2 and 3 combined, as compared to Group 1, have a very high probability of yielding stochastic larger scores at time k = 4 and smaller scores at k = 1 (i.e., very high probability of having stochastically larger or steeper slopes). Group 2 did not significantly differ from Group 3 in terms of linear trend (i.e., the confidence interval contains zero).

To transform the data by the initial trend contrast is standardized, $\mathbf{b}'_C = \{-\sqrt{2}\ +\sqrt{2}\ \ 0\ \ 0\}$. The values of $\mathbf{U}\mathbf{b}_C/K$ represent each individuals rank correlation with this ordered alternative. The results in the bottom panel of Table 7 show that Group 1 does not significantly differ from Groups 2 and 3 combined (i.e., the confidence interval contains zero). However, the change from time k = 1 to k = 2 was positive for Group 2 and negative for Group 3 (see Table 6, lower panel). The difference in these rank correlations was -1.0277. The univariate 95% simultaneous confidence interval indicates that plausible values for the difference in rank correlation range between 0.6898 and 1.3657. The multivariate 95% simultaneous confidence interval gives a wider band of plausible values that range between 0.5433 and 1.5111.

For analyses such as Initial Change contrast, $\mathbf{U}\mathbf{b}_C$, only two of the K repeated measures are used and thus interpretations reduce to the interpretations similar to the sign test. However, this approach includes information from the other time points; thus, these effects are relative to the other time points. If a more direct interpretation is desired, then the signs or signed ranks for the differences for the two measures could be computed and statistical analyses conducted to compare the groups. This is a methodology proposed by Cliff (1996) and is not detailed here.

Conclusion

Rank-based methods could be applied to the data in a multiple group repeated measures experiment because the normality assumptions of the split-plot ANOVA model in (1) are violated. In such a case, testing against the shift model null hypothesis (20) would be of interest because it seems conceptually similar to the differences among means in the parametric model hypotheses in (2) or (6). However, if aligned rank procedures are employed and tests of interactions are conducted, then (20) may be rejected incorrectly because some other hypothesis (i.e., 16, 17, 24 or 25) is false. That is, a statistically significant test statistic may be attributable to differences in other distributional characteristics (i.e., variance or shape) rather than reflecting solely differences in location, unless additional distributional assumptions are made (Serlin & Harwell, 2001).

In order to test against (20) and make inferences in terms of location parameters, distributional assumptions must be made. Credible inferences concerning location parameters (20) require the assumption that the population distributions are of identical shape (Serlin & Harwell, 2001; Vargha & Delaney, 1998). This may seem restrictive, however, because parametric statistical tests, which also require IID[$0,\sigma_\varepsilon^2$] or IID[$\mathbf{0}_{(K\text{-}1)}$,$\mathbf{D}'\mathbf{\Sigma}\mathbf{D}$] with the additional restriction that the error distributions have a normal shape (Bradley, 1968) have been conducted for decades.

Unfortunately, these distributional assumptions present a conundrum for data analysis. Specifically, the sample estimates of skew and kurtosis are unstable, especially with small sample sizes. Therefore, it is difficult to judge the tenability of the IID assumptions. The choices are: (a) accept the assumptions without testing their tenability or (b) test the assumptions based on unstable estimates. Furthermore, estimates of skew and kurtosis are more reliable with larger samples sizes. However, parametric procedures are more likely to be robust with large samples sizes and the advantage of rank-based procedures over parametric methods in terms of statistical power is likely to decrease.

To circumvent this conundrum, 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 statistically significant results are not attributed to location parameters alone but rather to any distributional difference. Vargha and Delaney (1998) and Beasley (2002) have suggested analyses of hypotheses related to stochastic heterogeneity. Similarly, Cliff (1996) has argued that rank-based and other nonparametric methods provide ordinal answers to ordinal questions, which are equivalent to results of stochastic heterogeneity and that these results correspond more closely to the goals of many researchers. These forms of hypotheses 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).

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; therefore, aligned ranks may still inherit the distributional properties of the original data (e.g., heterogeneity of variance). When the distributions have heterogeneous variances or have different shapes, the null hypothesis of equal location parameters (20) and the null hypothesis of identical distributions are no longer equivalent. Therefore, as analogs to parametric procedures, aligned rank tests are likely to be sensitive to variance heterogeneity, especially with unequal sample sizes (Algina & Keselman, 1998; Kowalchuk, Keselman, & Algina, 2003; Lei, Holt, & Beasley, 2004).

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. Thus, it may be conjectured that the aligned rank procedures as tests of location parameters would be somewhat robust to heterogeneous variance and differences in shape when sample sizes are equal; however, Lei, et al. (2004) have shown that tests that correct for unequal variances (e.g., Huynh, 1978) performed on aligned ranks still detect distributional (i.e., variance) differences when location parameters do not reflect an interaction. Furthermore, with increasing disparity among sample sizes, aligned rank procedures become more sensitive to detecting any distributional difference and thus should strictly be considered tests of stochastic homogeneity.

Vargha and Delaney (1998) explicated 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 (20) are equivalent for identical, asymmetric distributions. Therefore, statistically significant values for interaction tests performed on aligned ranks, and the subsequent rejections of the associated null hypotheses, 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 & Langer, 2000; Vargha & 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 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_{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: 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 ranks can generally be attributed to differences in location parameters (Marascuilo & McSweeney, 1977, pp. 304-305), which is fortunate because it is difficult to test the tenability of the IID assumptions associated with the shift models. Newson (2002) reviewed methods for constructing confidence intervals that are robust to between-group differences in parameters other than location (e.g., variance; skew). Technically, however, statistically significant tests performed on aligned ranks cannot be attributed solely to differences in location parameters. Given the difficulty of testing model assumptions especially with small samples, results from these procedures should be interpreted in terms of stochastic heterogeneity (Beasley, 2002; Varga & Delaney, 1998). Newson (2002) and Cliff (1996) suggest that rank-based statistics are based on population parameters, related to Somer's (1962) D, which are extremely informative in terms of stochastic dominance and can be estimated using corresponding sample statistics. Thus, although aligned rank-procedures produce what may be considered a more ambiguous formulation of the underlying null hypothesis that is of interest conceptually, the conclusions are consistent with the ordinal answers that Cliff (1996) has extolled as the effect of actual interest to many researchers.

Notes

1. In a two-group Between-Subjects design, Cliff (1996) has shown that transforming the ranks by $[(2R_{ijk} - 1)/N]$ yields a rank mean difference equal to the d statistic. This transformation will only yield standard errors similar to Cliff's method asymptotically. This is because they are based on different counting procedures. Furthermore, this transformation does not necessarily extend to multiple groups and dependent measures. Thus, the transformation suggested by Agresti and Pendergast (1986) was used.

2. Brunner, et al. (2002) showed a linear transformation of unaligned ranks $[(R_{ijk} - ½)/NK]$, similar to the Agresti and Pendergast (1986) suggestion, will yield cell means that provide estimates of relative treatment effects. Test statistics performed on these values will provide valid tests of fully nonparametric hypotheses. According to Brunner, et al. (2002), however, these values cannot simply be used to compute standard errors, unless the sample size is large. Constructing accurate confidence intervals using the Brunner, et al. method involves a more complicated procedure of computing partial ranks and logit transformations. Whether the Brunner, et al. method can be applied to aligned ranks has yet to be investigated. Thus, for the sake of simplicity the transformation suggested by Agresti and Pendergast (1986) was used.

References

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

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

Akritas, M. G., Arnold, S. F., & Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, *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. *Journal of Educational & Behavioral Statistics*, *23*, 152-169.

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

Allison, D. B., et al. (1999). Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *American Journal of Human Genetics*, *65*, 531-544.

Avants, S. K., Margolin, A., Sindelar, J., & Rounsaville, B. J. (1999). Day treatment versus enhanced standard methadone services for opioid-dependent patients: A comparison of clinical efficacy and cost. *American Journal of Psychiatry*, *156*, 95.

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

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

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.

Beckett, J., & Schucany, W. R. (1979). Concordance among categorized groups of judges. *Journal of Educational Statistics*, *4*, 125-137.

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

Bonett, D. G., & Price, R. M. (2002). Statistical inference for a linear function of medians: Confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods*, *7*, 370-383.

Boomsma, D. I., Martin, N. G., & Molenaar, P. C. M. (1989). Factor and simplex models for repeated measures: Applications to two psychomotor measures of alcohol sensitivity in twins. *Behavior Genetics*, *19*, 79-96.

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. *Annals of Mathematical Statistics*, *25*, 290-302.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.

Brunner, E., Domhof, S., & Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial experiments*. NY: Wiley.

Brunner, E., & Langer, F. (2000). Nonparametric analysis of ordered categorical data in designs with longitudinal observations and small sample sizes. *Biometrical Journal*, *42*, 663-675.

Campbell, M. J., & Gardner, M. J. (1988). Calculating confidence intervals for some non-parametric analyses. *British Medical Journal*, *296*, 1454-1456.

Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Erlbaum.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*, 675-701.

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

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than P values: Estimation rather than hypothesis testing. *British Medical Journal*, *292*, 746-750.

Gentile, J. R., Voelkl, K. E., Mt. Pleasant, J., & Monaco, N. M. (1995). Recall after relearning by fast and slow learners. *Journal of Experimental Education*, *63*, 185-197.

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. *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: Simulation & Computation, 26*, 605-618.

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

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

Hodges, J. L., & Lehmann. E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *Annals of Mathematical Statistics*, *33*, 482-497.

Hodges, J. L., & Lehmann. E. L. (1963). Estimates of location based on ranks. *Annals of Mathematical Statistics*, *34*, 598-611.

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*. NY: Wiley.

Hora, S. C., & Conover, W. J., (1984). The *F*-statistic in the two-way layout with rank-score transformed data. *Journal of the American Statistical Association*, *79*, 668-673.

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, *2*, 360-378.

Hotelling, H. (1951). A generalized T-test and measure of multivariate dispersion. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics & Probability*, *2*, 23-41.

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

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

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

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

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

Keselman, H. J., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350-386.

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

Koch, G. G. (1969). Some aspects of the statistical analysis of "split-plot" experiments in completely randomized layouts. *Journal of the American Statistical Association*, *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. *International Statistical Review*, *48*, 249-265.

Koch, G. G., & Sen, P. K. (1968). Some aspects of the statistical analysis of the mixed model. *Biometrics*, *24*, 27-48.

Kowalchuk, R. K., Keselman, H. J., & Algina, J. (2003). Repeated measures interaction test with aligned ranks. *Multivariate Behavioral Research*, *38*, 433-461.

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.

Lei, X., Holt, J., & Beasley, T. M. (2004). Aligned rank tests as robust alternatives for testing interactions in multiple group repeated measures designs with heterogeneous covariances. *Journal of Modern Applied Statistical Methods*, *3*(*2*), 462-475.

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

Lyerly, S. B. (1952). The average Spearman rank correlation coefficient. *Psychometrika*, *17*, 412-428.

Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sunderland, MA: Sinauer.

Marascuilo, L. A., (1966). Large-sample multiple comparisons. *Psychological Bulletin*, *65*, 280-290

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

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks-Cole.

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

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

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

Newson, R. (2002). Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *Stata Journal*, *2*(*1*), 45-64.

Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, *58*, 216-230.

Pillai, K. C. S. (1960). *Statistical tables for tests of multivariate hypotheses*. Manila: University of the Philippines, Statistical Service Center.

Randles, R.H., & Wolfe, D.A. (1979). *Introduction to the theory of nonparametric statistics*. NY: Wiley.

Rasmussen, J. L. (1989). Parametric and non-parametric analysis of groups by trials design under variance-covariance inhomogeneity. *British Journal of Mathematical & Statistical Psychology*, *42*, 91-102.

Sawilowsky, S., S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the Type I error and power properties of the rank transform in factorial ANOVA. *Communications in Statistics*, *14*, 25-267.

Salter, K. C., & Fawcett, R. F. (1993). The ART test of interaction: A robust and powerful test of interaction in factorial models. *Communications in Statistics: Simulation & Computation*, *22*, 137-153.

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

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education*, *61*, 350-360.

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

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

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, *27*, 799-811.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(*3*), 24-31.

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

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

Waldstein, S. R., et al. (1991). Learning and memory function in men with untreated blood pressure elevation. *Journal of Consulting & Clinical Psychology*, *59*, 513-517.

Welch, B. L. (1947). The generalization of Student's problem when several different population variances are involved. *Biometrika*, *34*, 28-35.

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

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

Zimmerman, D. W. (1996). A note on homogeneity of variance of scores and ranks. *Journal of Experimental Education*, *64*, 351-362.

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*, 139-149.

Appendix I: SAS Code

```
data egtwo;
input k1 k2 k3 k4 group;
cards;
   3.90  4.20  5.10  5.10  1
   4.10  4.00  5.00  5.20  1
   4.30  5.00  5.40  6.10  1
   5.00  5.10  6.00  6.00  1
   5.20  5.30  5.90  7.20  1
   5.80  6.00  7.00  7.00  1
   6.10  6.20  7.30  7.10  1
  12.00 13.10 13.30 14.00  1
   3.00  9.20 10.10 11.30  2
   4.10 10.10 11.00 12.20  2
   4.00 11.20 11.90 13.00  2
   4.20 12.30 13.10 17.20  2
   5.20 11.20 14.30 15.20  2
   5.00 11.30 13.40 18.30  2
   6.00 12.20 14.90 17.00  2
   6.20 12.40 14.00 17.90  2
   7.30 13.50 15.20 18.20  2
  13.00 14.50 15.00 18.10  2
   3.00  4.90  7.70 11.60  3
   4.10  6.10  7.70 11.70  3
   5.10  5.90  8.10 13.20  3
   5.00  5.80  8.70 12.80  3
   5.30  6.30  9.70 13.90  3
   5.90  6.30  8.70 13.00  3
   6.10  7.00  9.90 13.10  3
  12.50 14.10 16.20 20.30  3
;proc sort out=two;by group;
data three;options ls=120;
proc iml; use two;
```

```
read all var{k1 k2 k3 k4} into
Y;read all var{group} into
Group;
JJ=max(Group);K=ncol(Y);N=nrow(Y
);NV=j(JJ,1,0);
Q=j(N,K,0);FR=j(N,K,0); CK=j((k-
1),K,0);CJ=j((jj-1),JJ,0);
dfjk=(JJ-1)*(K-1);dfeu=N-JJ)*K-
1);
smv=min(JJ,K);smv=smv-1;
mmv=((ABS(K-JJ))-1)/2;
nmv=(N-JJ-K)/2;
dfem=2#((smv#nmv)+1);
do hh=1 to JJ;
  do ii=1 to N;
    if    group[ii,1]=hh    then
NV[hh,1]=NV[hh,1]+1;
end;end;
RMMEAN=Y[:,];RMMEAN=(j(N,1,1))*R
MMEAN;
PMEAN=Y[,:];GMEAN=PMEAN[:,];PMEA
N=PMEAN*(j(1,K,1));
AD=(Y-PMEAN-
RMMEAN)+GMEAN;AR=RANKTIE(AD);AR=
AR/((N*K)+1);
do hh=1 to K;
  do ii=1 to K;
   DX=Y[,hh]-
Y[,ii];RDX=RANKTIE(DX);Q[,hh]=Q[
,hh]+RDX;
end; end;
Q=(Q-((N+1)/2))/((K-1)*(N+1));
do ii=1 to N;
  FR[ii,]=RANKTIE(AD[ii,]);
end;
FR=(FR-((K+1)/2))/(((K##2)-
1)/12);
do ii=1 to (K-1);
  CK[ii,ii]=1; CK[ii,(ii+1)]=-1;
end;
do ii= 1 to (JJ-1);
  CJ[ii,ii]=1;CJ[ii,(ii+1)]=-1;
end;
CJK=CJ@CK;
AMEANK=AR[:,];QMEANK=Q[:,];RMEAN
K=FR[:,];
do ii=1 to JJ;
  if    ii=1    then    zz=1;else
zz=zz+NV[(ii-1),1];
  if ii=1 then zzz=NV[ii,1];else
zzz=zzz+NV[ii,1];
 do hh=zz to zzz;
  if hh=zz then AJ=AR[hh,]; else
AJ=AJ//AR[hh,];
  if hh=zz then QJ=Q[hh,]; else
QJ=QJ//Q[hh,];
  if hh=zz then RJ=FR[hh,]; else
RJ=RJ//FR[hh,];
end;
MAJ=AJ[:,];DMAJ=MAJ-
AMEANK;DMAJ=DMAJ#(NV[ii,1]);
EAJ=AJ-
(((j((NV[ii,1]),1,1))*MAJ));
if ii = 1 then AMEAN=MAJ; else
AMEAN=AMEAN//MAJ;
if ii = 1 then DEVA=MAJ; else
DEVA=DEVA||MAJ;
if ii = 1 then EA=EAJ; else
EA=EA//EAJ;
MQJ=QJ[:,];DMQJ=MQJ-QMEANK;
DMQJ=DMQJ#(NV[ii,1]);
EQJ=QJ-
(((j((NV[ii,1]),1,1))*MQJ));
if ii = 1 then QMEAN=MQJ; else
QMEAN=QMEAN//MQJ;
if ii = 1 then DEVQ=MQJ; else
DEVQ=DEVQ||MQJ;
if ii = 1 then EQ=EQJ; else
EQ=EQ//EQJ;
MRJ=RJ[:,];DMRJ=MRJ-
RMEANK;DMRJ=DMRJ#(NV[ii,1]);
ERJ=RJ-
(((j((NV[ii,1]),1,1))*MRJ));
if ii = 1 then RMEAN=MRJ; else
RMEAN=RMEAN//MRJ;
if ii = 1 then DEVR=MRJ; else
DEVR=DEVR||MRJ;
if ii = 1 then ER=ERJ; else
ER=ER//ERJ;
end;
EA=EA`*EA;TA=AR-
((j(N,1,1))*AMEANK);TA=TA`*TA;
EQ=EQ`*EQ;TQ=Q-
((j(N,1,1))*QMEANK);TQ=TQ`*TQ;
ER=ER`*ER;TR=FR-
((j(N,1,1))*RMEANK);TR=TR`*TR;
HTA=((CJK*(DEVA`))`)*(ginv((CJK*
((diag((1/nv)))@EA)*((CJK`)))))*
(CJK*(DEVA`));
VA=
((CJK*(DEVA`))`)*(ginv((CJK*((di
ag((1/nv)))@TA)*((CJK`)))))*(CJK
*(DEVA`));
```

```
HTQ=((CJK*(DEVQ`))`)*(ginv((CJK*
((diag((1/nv)))@EQ)*((CJK`)))))*
(CJK*(DEVQ`));
VQ=
((CJK*(DEVQ`))`)*(ginv((CJK*((di
ag((1/nv)))@TQ)*((CJK`)))))*(CJK
*(DEVQ`));
HTR=((CJK*(DEVR`))`)*(ginv((CJK*
((diag((1/nv)))@ER)*((CJK`)))))*
(CJK*(DEVR`));
VR=
((CJK*(DEVR`))`)*(ginv((CJK*((di
ag((1/nv)))@TR)*((CJK`)))))*(CJK
*(DEVR`));
FHA=HTA#(dfem/(smv#dfjk));pvalam
=1-(probf(FHA,dfjk,dfem));
FHQ=HTQ#(dfem/(smv#dfjk));pvalqm
=1-(probf(FHQ,dfjk,dfem));
FHR=HTR#(dfem/(smv#dfjk));pvalrm
=1-(probf(FHR,dfjk,dfem));
FA=(((CJK*(DEVA`))`)*(ginv((CJK*
((diag((1/nv)))@I(K))*((CJK`))))
)*(CJK*(DEVA`))/(TRACE(EA)))*(df
eu/dfjk);
pvalau=1-(probf(FA,dfjk,dfeu));
FQ=(((CJK*(DEVQ`))`)*(ginv((CJK*
((diag((1/nv)))@I(K))*((CJK`))))
)*(CJK*(DEVQ`))/(TRACE(EQ)))*(df
eu/dfjk);
pvalqu=1-(probf(FQ,dfjk,dfeu));
FRC=((CJK*(DEVR`))`)*(ginv((CJK*
((diag((1/nv)))@I(K))*((CJK`))))
)*(CJK*(DEVR`))/((K#(K+1))/12);
pvalru=1-(probchi(FRC,dfjk));

Print 'Univariate Tests';
Rowun={"Aligned   Ranks   F(A)",
"Koch Ranks F(Q)",
"Chi-Square   -   Friedman   Ranks
F(R)"};
ColUN={"TEST"   "DFh"   "DFe"   "p-
value"};
Uprt=(FA//FQ//FRC)||(dfjk//dfjk/
/dfjk)||(dfeu//dfeu//0)||
(pvalau//pvalqu//pvalru);
print          UPrt[rowname=rowun
colname=colun];

Print 'Multivariate Tests';Print
'DFh =' dfjk;Print 'DFe =' dfem;
Rowmn={"Aligned    Ranks    (A)",
"Koch   Ranks   (Q)",   "Friedman
Ranks (R)"};
ColmN={"Pillia    Trace    V(*)"
"Hotelling   Trace   H(*)"   "F-
approx" "p-value"};
Mprt=(VA//VQ//VR)||(HTA//HTQ//HT
R)||(FHA//FHQ//FHR)||(pvalam//pv
alqm//pvalrm);
print          MPrt[rowname=rowmn
colname=colmn];

DLINE={-3          -1          1
3};DLINE=DLINE/(20##.5);
DCHNG={-1 1 0 0};
YL=Y*(DLINE`);YC=Y*(DCHNG`);
AL=AR*(DLINE`);AC=AR*(DCHNG`);
QL=Q*(DLINE`);QC=Q*(DCHNG`);
FL=(FR*((DLINE`)#2))/4;FC=(FR*((
DCHNG`)#(2##.5)))/4;

outx=Y||AR||Q||FR||YL||YC||AL||A
C||QL||QC||FL||FC||Group;
create xxx from outx[colname={k1
k2 k3 k4 ak1 ak2 ak3 ak4 qk1 qk2
qk3 qk4 fk1 fk2 fk3 fk4 yl yc al
ac ql qc fl fc group}];
append from outx;
data last;set xxx;
proc glm;class group;
model k1 k2 k3 k4=group/nouni;
contrast  'Group  1  vs  2  +  3'
group 1 -.5 -.5;
contrast 'Group 2 vs 3' group 0
1 -1;
repeated  time  4  (1  2  3  4)
polynomial/summary;run;
proc glm;class group;
model      ak1      ak2      ak3
ak4=group/nouni;
contrast  'Group  1  vs  2  +  3'
group 1 -.5 -.5;
contrast 'Group 2 vs 3' group 0
1 -1;
repeated  time  4  (1  2  3  4)
polynomial/ summary;run;
proc glm;class group;
model qk1 qk2 qk3 qk4 = group /
nouni;
contrast  'Group  1  vs  2  +  3'
group 1 -.5 -.5;
```

```
contrast 'Group 2 vs 3' group 0
1 -1;
repeated time 4 (1 2 3 4)
polynomial/ summary;run;
proc glm;class group;
model fk1 fk2 fk3 fk4 = group /
nouni;
contrast 'Group 1 vs 2 + 3'
group 1 -.5 -.5;
contrast 'Group 2 vs 3' group 0
1 -1;
repeated time 4 (1 2 3 4)
polynomial/ summary;run;
proc glm;class group;
model yl yc al ac ql qc fl fc =
group / nouni;
contrast 'Group 1 vs 2 + 3'
group 1 -.5 -.5;
contrast 'Group 2 vs 3' group 0
1 -1;
repeated time 4 (1 2 3 4)
polynomial/ summary;run;
```