Non-parametric three-way mixed ANOVA with aligned rank tests

ARTICLE *in* BRITISH JOURNAL OF MATHEMATICAL AND STATISTICAL PSYCHOLOGY · DECEMBER 2013

Impact Factor: 2.17 · DOI: 10.1111/bmsp.12031 · Source: PubMed

2 AUTHORS, INCLUDING:

X.T. Wang

University of South Dakota

60 PUBLICATIONS **994** CITATIONS

SEE PROFILE

British Journal of Mathematical and Statistical Psychology (2013)

The British Psychological Society

www.wileyonlinelibrary.com

Non-parametric three-way mixed ANOVA with aligned rank tests

Juan C. Oliver-Rodríguez[1]* and X. T. Wang[2]

[1]Universitat Jaume I, Castellón, Spain

[2]University of South Dakota, Vermillion, South Dakota, USA

Research problems that require a non-parametric analysis of multifactor designs with repeated measures arise in the behavioural sciences. There is, however, a lack of available procedures in commonly used statistical packages. In the present study, a generalization of the aligned rank test for the two-way interaction is proposed for the analysis of the typical sources of variation in a three-way analysis of variance (ANOVA) with repeated measures. It can be implemented in the usual statistical packages. Its statistical properties are tested by using simulation methods with two sample sizes (n = 30 and n = 10) and three distributions (normal, exponential and double exponential). Results indicate substantial increases in power for non-normal distributions in comparison with the usual parametric tests. Similar levels of Type I error for both parametric and aligned rank ANOVA were obtained with non-normal distributions and large sample sizes. Degrees-of-freedom adjustments for Type I error control in small samples are proposed. The procedure is applied to a case study with 30 participants per group where it detects gender differences in linguistic abilities in blind children not shown previously by other methods.

1. Introduction

Problems in need of non-parametric tests of variable interactions in the analysis of variance (ANOVA) for research designs arise frequently in basic and applied behaviour research. In an extensive review of several hundred data distributions from both research articles and educational evaluation agencies the normality assumption was violated in a large majority of cases (Micceri, 1989). Fifty per cent of the distributions showed higher tail densities than those found in the normal model. Thirty per cent of the distributions had extreme asymmetries which approached the exponential distribution. In many situations of these kinds, non-parametric tests have better properties than classical parametric tests in terms of power, efficiency, or Type I error biases (Wilcox & Keselman, 2003). However, procedures for the non-parametric testing of variable interactions are not included in the ANOVA modules of standard statistical packages.

One proposed method consists of performing an F parametric contrast on the ranked observations (Conover & Iman, 1981). In one-way designs, this procedure is equivalent to Mann–Whitney and Wilcoxon tests for between- or within-subject comparisons. In two-way designs, however, the presence of main effects has been shown to confound interaction effects, leading to increased Type I errors (Blair, Sawilowsky, & Higgins, 1987;

*Correspondence should be addressed to Juan C. Oliver-Rodríguez, Universitat Jaume I, 12071 Castellón, Spain (email: jcoliver@ya.com).

DOI:10.1111/bmsp.12031

Thompson, 1991a,b). A proposed solution is to treat main effects as confounding variables and to subtract their influence from the observations before ranking and calculating the F statistics. This procedure has been termed the *aligned rank test* (Hodges & Lehmann, 1962) and has been shown to be robust in terms of Type I error rates and statistical power in non-normal distributions (Beasley, 2002; Salter & Fawcett, 1993; Toothaker & Newman, 1994). It can also be implemented in standard statistical packages.

The present study proposes a generalization of this method for the analysis of the sources of variation typically obtained in a three-way ANOVA with repeated measures. Its statistical properties are tested by simulation methods. It is then applied to a case study on gender differences in children with sensory disabilities.

2. Aligned rank tests

2.1. Linear model

The sources of variation for a mixed design with two between-subject fixed effect factors (A and B) and one within-subject fixed effect factor (M) can be specified as follows:

$$\begin{aligned} y_{ijkl} = \mu_{....} &+ \alpha_i + \beta_j + (\alpha\beta)_{ij} + s_{l.(ij)} \\ &+ \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} \\ &+ \varepsilon_{kl(ij)} \end{aligned} \tag{1}$$

$$\begin{aligned} \text{for } i &= 1, 2, \ldots, a \text{ levels of factor A,} \\ j &= 1, 2, \ldots, b \text{ levels of factor B,} \\ k &= 1, 2, \ldots, m \text{ levels of factor M,} \\ \text{and } l &= 1, 2, \ldots, n \text{ participants per experimental condition} \end{aligned}$$

The first line of model (1) contains all between-subject effects: the main effects of factors A (α_i) and B (β_j), their interaction ($\alpha\beta_{ij}$), plus the between-subject error term due to individual differences ($s_{l.(ij)}$). The second and third lines of the model contain all within-subject effects: the main effect of factor M (λ_k), the two-way interactions ($\alpha\gamma_{ik}$), ($\beta\gamma_{jk}$), the three-way interaction ($\alpha\beta\gamma_{ijk}$) and the within-subject error term ($\varepsilon_{kl(ij)}$).

The usual assumptions are that random error effects $s_{l.(ij)}$ and $\varepsilon_{kl(ij)}$ are independently and identically distributed (i.i.d.). The covariance matrices for the within-subject factor are assumed to be equal between groups and have the property of sphericity, it being the case that:

$$\left.\begin{aligned} s_{l.(ij)} &\sim iid(0, \sigma_s^2) \\ \varepsilon_{kl(ij)} &\sim iid(0, \sigma_e^2) \end{aligned}\right\} \text{ and both random effects are independent.} \tag{2}$$

2.2. Aligned rank transformations

Type I error rates for the rank transform method (Conover & Iman, 1981) in a three-way ANOVA design have been shown to increase as a function of the number of other non-null effects in the model (Sawilowsky, Blair, & Higgins, 1989). When all of the remaining

effects were present, maximum values near 1 have been observed, meaning that the null hypothesis in these cases will almost always be rejected when it is true. The aligned rank procedure used here is analogous to the one used for testing the interaction in two-way designs. Here it is generalized by creating a new aligned rank variable for each experimental effect. Each variable is obtained by ranking the observations after removing the confounding sources of variability not contained in the expected mean squares for that effect (Table 1). Transformations for between- and within-subject effects are described in Table 2.

Table 1. Expected mean squares for sources of variation in an analysis of variance of a three-way design with one within-subject factor. A, B, and M are fixed effects

Effect	$E[MS]$
Between subjects	
A	$\sigma_e^2 + m\sigma_{s(ab)}^2 + nbm\sum_{i=1}^{a}\alpha_i^2 \Big/ a-1$
B	$\sigma_e^2 + m\sigma_{s(ab)}^2 + nam\sum_{j=1}^{b}\beta_j^2 \Big/ b-1$
A × B	$\sigma_e^2 + m\sigma_{s(ab)}^2 + nm\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\alpha\beta_{ij}\right)^2 \Big/ (a-1)(b-1)$
S / A × B	$\sigma_e^2 + m\sigma_{s(ab)}^2$
Within subjects	
M	$\sigma_e^2 + \sigma_{ms(ab)}^2 + nab\sum_{k=1}^{m}\gamma_k^2 \Big/ m-1$
A × M	$\sigma_e^2 + \sigma_{ms(ab)}^2 + nb\sum_{i=1}^{a}\sum_{k=1}^{m}\left(\alpha\gamma_{ik}\right)^2 \Big/ (a-1)(m-1)$
B × M	$\sigma_e^2 + \sigma_{ms(ab)}^2 + na\sum_{j=1}^{b}\sum_{k=1}^{m}\left(\beta\gamma_{jk}\right)^2 \Big/ (b-1)(m-1)$
A × B × M	$\sigma_e^2 + \sigma_{ms(ab)}^2 + n\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{m}\left(\alpha\beta\gamma_{ijk}\right)^2 \Big/ (a-1)(b-1)(m-1)$
M × S / A × B	$\sigma_e^2 + \sigma_{ms(ab)}^2$

Note. S/A × B and M × S/A × B are the between-subject and within-subject error terms, respectively.

Table 2. Aligned rank variable transformations for experimental effects in the ANOVA table

Effect	Parameters	Estimates
Between subjects		
A	$R\left(y^{A}_{ijkl}\right) = \text{Rank}(\mu_{....} + \alpha_i + s_{l.(ij)})$	$R\left(\hat{y}^{A}_{ijkl}\right) = \text{Rank}(\overline{y}_{i...} + \hat{s}_{l.(ij)})$
B	$R\left(y^{B}_{ijkl}\right) = \text{Rank}\left(\mu_{....} + \beta_j + s_{l.(ij)}\right)$	$R\left(\hat{y}^{B}_{ijkl}\right) = \text{Rank}(\overline{y}_{.j.} + \hat{s}_{l.(ij)})$
A × B	$R\left(y^{AB}_{ijkl}\right) = \text{Rank}(\mu_{....} + \alpha\beta_{ij} + s_{l.(ij)})$	$R\left(\hat{y}^{AB}_{ijkl}\right) = \text{Rank}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...} + \hat{s}_{l.(ij)})$
Within subjects		
M	$R\left(y^{M}_{ijkl}\right) = \text{Rank}\left(\mu_{....} + \gamma_k + \varepsilon_{kl(ij)}\right)$	$R\left(\hat{y}^{M}_{ijkl}\right) = \text{Rank}(\overline{y}_{..k} + \hat{\varepsilon}_{kl(ij)})$
A × M	$R\left(y^{AM}_{ijkl}\right) = \text{Rank}\left(\mu_{....} + \alpha\gamma_{ik} + \varepsilon_{kl(ij)}\right)$	$R\left(\hat{y}^{AM}_{ijkl}\right) = \text{Rank}(\overline{y}_{i.k} - \overline{y}_{i..} - \overline{y}_{..k} + \overline{y}_{...} + \hat{\varepsilon}_{kl(ij)})$
B × M	$R\left(y^{BM}_{ijkl}\right) = \text{Rank}\left(\mu_{....} + \beta\gamma_{jk} + \varepsilon_{kl(ij)}\right)$	$R\left(\hat{y}^{BM}_{ijkl}\right) = \text{Rank}(\overline{y}_{.jk} - \overline{y}_{.j.} - \overline{y}_{..k} + \overline{y}_{...} + \hat{\varepsilon}_{kl(ij)})$
A × B × M	$R\left(y^{ABM}_{ijkl}\right) = \text{Rank}\left(\mu_{....} + \alpha\beta\gamma_{ijk} + \varepsilon_{kl(ij)}\right)$	$R\left(\hat{y}^{ABM}_{ijkl}\right) = \text{Rank}(\overline{y}_{ijk} - \overline{y}_{ij.} - \overline{y}_{i.k} - \overline{y}_{.jk} + \overline{y}_{i..} + \overline{y}_{.j.} + \overline{y}_{..k} - \overline{y}_{...} + \hat{\varepsilon}_{kl(ij)})$

2.3. Hypotheses

In the usual analysis of factorial designs with normal distributions, main and interaction location hypotheses are expressed in terms of equality of means. These are summary row or column averages for the levels or level combinations of a specific source of variation. In a non-parametric context, however, null location hypotheses are most clearly and flexibly expressed as equalities of cumulative distribution functions. These are summary sets of row or column averages for each of the ordered values obtained at the different levels or level combinations of a source of variation (Shah & Madden, 2004). Hypotheses are listed in Table 3. Equality of complete distributions is therefore being tested rather than single numerical parameters.

2.4. Test statistics

In normal distribution analyses for mixed models and balanced data, F statistics are built by dividing a particular effect mean square (MS_{effect}) by its error mean square (MS_{e}). The same expected error variance term (either between or within subjects) is contained in both of these, but the term for each experimental effect being tested appears only in the numerator (Hocking, 1996). An analogous procedure will be used here, but the standard F test will be applied to each aligned rank variable rather than to the original raw variable. The corresponding test statistic for each effect will be called an *aligned rank F* (F_{AR}; Table 3). Only when the assumption of identical

Table 3. Null hypotheses and aligned rank test statistics for experimental effects in the ANOVA table

Effect	Hypotheses	Test statistic
Between subjects		
A	$F(y_{i\cdot\cdot}) - F(y_{i'\cdot\cdot}) = 0$ for all y and any i, i' levels of factor A	$F_{ar}^{A} = \frac{MS_{y^A}}{MS_{s(ab)}}$
B	$F(y_{\cdot j\cdot}) - F(y_{\cdot j'\cdot}) = 0$ for all y and any j, j' levels of factor B	$F_{ar}^{B} = \frac{MS_{y^B}}{MS_{s(ab)}}$
A × B	$F(y_{ij\cdot}) - F(y_{i\cdot\cdot}) - F(y_{\cdot j\cdot}) + F(y_{\cdots}) = 0$ for all y and any i, j levels of factors A and B	$F_{ar}^{AB} = \frac{MS_{y^{AB}}}{MS_{s(ab)}}$
Within subjects		
M	$F(y_{\cdot\cdot k}) - F(y_{\cdot\cdot k'}) = 0$ for all y and any k, k' levels of factor M	$F_{ar}^{M} = \frac{MS_{y^M}}{MS_e}$
A × M	$F(y_{i\cdot k}) - F(y_{i\cdot\cdot}) - F(y_{\cdot\cdot k}) + F(y_{\cdots}) = 0$ for all y and any i, k levels of factors A and M	$F_{ar}^{AM} = \frac{MS_{y^{AM}}}{MS_e}$
B × M	$F(y_{\cdot jk}) - F(y_{\cdot j\cdot}) - F(y_{\cdot\cdot k}) + F(y_{\cdots}) = 0$ for all y and any j, k levels of factors B and M	$F_{ar}^{BM} = \frac{MS_{y^{BM}}}{MS_e}$
A × B × M	$F(y_{ijk}) - F(y_{i\cdot\cdot}) - F(y_{\cdot j\cdot}) - F(y_{\cdot\cdot k}) + F(y_{ij\cdot}) + F(y_{i\cdot k}) + F(y_{\cdot jk}) - F(y_{\cdots}) = 0$ for all y and any i, j, k levels of factors A and B and M	$F_{ar}^{ABM} = \frac{MS_{y^{ABM}}}{MS_e}$

distributions above holds with equal shapes and dispersion matrices can rejection of a null hypothesis be interpreted as a mean increase or decrease in the variable of interest between experimental conditions (Fay & Proschan, 2010; Vargha & Delaney, 1998).

Violation of the sphericity assumption in normal distribution analyses generates a positive bias in the *F* statistic (Box, 1954; Huynh & Feldt, 1976). As a consequence an increase in Type I error rate occurs and the proportion of incorrect rejections of the null hypothesis will be larger than α. The bias correction commonly used is adjustment of the numerator and denominator of the *F* test degrees of freedom by multiplying them by an estimate of an epsilon (ε) parameter, which is a function of the degree to which the real covariance matrix departs from sphericity. It is routinely

performed in repeated measures procedures of standard statistical packages such as SAS or SPSS. However, Lecoutre (1991) detected an error for mixed models that has still not been corrected in these packages and proposed a new estimate ($\tilde{\varepsilon}$) that produces in normal distributions an additional reduction of bias (Beasley, 2002; Chen & Dunlap, 1994):

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

where N is the total number of participants, g is the number of groups (or $a \times b$ in the linear model (1)), m is the number of levels of the within-subject factor and $\hat{\varepsilon}$ is the estimated parameter from the pooled within-group covariance matrix (Winer, Brown, & Michels, 1991, p. 257). This Lecoutre adjusted F test will be denoted L and its aligned rank version L_{AR}. In a two-factor mixed model design this correction has been shown to produce satisfactory results when applied to aligned rank variables in exponential (asymmetric) and double exponential (heavy-tailed symmetric) distributions (Beasley, 2002). For this reason its generalization to the three-factor case will also be assessed.

Another alternative under violation of the sphericity assumption is to use a multivariate ANOVA, which is also customarily included in SPSS or SAS output (Vallejo & Lozano, 2006). In comparison with the univariate ε adjusted procedure its statistical power depends on N, m, and ε. The multivariate alternative is recommended when $N \geq m + 30$, $\hat{\varepsilon} \leq .85$ and $m \leq 8$ (Algina & Keselman, 1997). For large samples the Hotelling test has also been shown to have statistical power advantages over the univariate adjusted degrees-of-freedom test when it is applied to aligned rank variables in two-way mixed models with exponential or double exponential distributions (Beasley, 2002). The Hotelling test will be denoted H and its aligned rank version H_{AR}. Its generalizability to the three-factor situation will also be assessed.

3. Method

3.1. Simulation procedures

A simulated experiment was conducted for each of the 96 conditions defined by all possible combinations of presence or absence of main effects, two- and three-way interactions, two covariance matrices which either met or violated the sphericity condition, two sample sizes (10 and 30 participants per group) and three distributions (normal, exponential and double exponential). Two levels of each between-subject factor and four levels of the within-subject factors were used. A thousand replications per condition were run.

The presence or absence of effects was respectively defined by adding or subtracting a constant $c = 0.125$ or $c = 0$ to or from two or more different levels of each experimental effect. They were therefore specified in the following manner: for the main effects,

$$\begin{aligned}
&\alpha_1 = \beta_1 = \gamma_1 = c,\\
&\alpha_2 = \beta_2 = \gamma_2 = -c,\\
&\gamma_3 = \gamma_4 = 0;
\end{aligned} \tag{4}$$

for the two-way interactions,

$$\begin{aligned}
&\alpha\beta_{11} = \alpha\beta_{22} = \alpha\gamma_{11} = \alpha\gamma_{24} = \beta\gamma_{11} = \beta\gamma_{24} = c,\\
&\alpha\beta_{12} = \alpha\beta_{21} = \alpha\gamma_{14} = \alpha\gamma_{21} = \beta\gamma_{14} = \beta\gamma_{21} = -c,\\
&\alpha\gamma_{12} = \alpha\gamma_{13} = \alpha\gamma_{22} = \alpha\gamma_{23} = 0,\\
&\beta\gamma_{12} = \beta\gamma_{13} = \beta\gamma_{22} = \beta\gamma_{23} = 0;
\end{aligned} \tag{5}$$

and for the three-way interaction,

$$\begin{aligned}
&\alpha\beta\gamma_{111} = \alpha\beta\gamma_{124} = \alpha\beta\gamma_{214} = \alpha\beta\gamma_{221} = c,\\
&\alpha\beta\gamma_{114} = \alpha\beta\gamma_{121} = \alpha\beta\gamma_{211} = \alpha\beta\gamma_{214} = -c,\\
&\alpha\beta\gamma_{112} = \alpha\beta\gamma_{113} = \alpha\beta\gamma_{122} = \alpha\beta\gamma_{123} = 0,\\
&\alpha\beta\gamma_{212} = \alpha\beta\gamma_{213} = \alpha\beta\gamma_{222} = \alpha\beta\gamma_{223} = 0.
\end{aligned} \tag{6}$$

Random variability from the normal, exponential or double exponential distributions was generated with a mean of 0, a standard deviation of 1, and the following two covariance matrices with Greenhouse–Geisser correction either $\varepsilon = 1$,

$$\sum = \begin{bmatrix} 1 & .6 & .6 & .6 \\ & 1 & .6 & .6 \\ & & 1 & .6 \\ & & & 1 \end{bmatrix},$$

or $\varepsilon = 0.69$,

$$\sum = \begin{bmatrix} 1 & .3 & .7 & .3 \\ & 1 & .3 & .7 \\ & & 1 & .3 \\ & & & 1 \end{bmatrix}.$$

The algorithm used was an extension of the Fleishman power method running on SAS IML software (Headrick & Sawilowsky, 1999).

3.2. Computation of aligned rank tests

At each one of the thousand replications per simulation condition, the data table was put in a univariate format with five columns: one for each factor in the ANOVA model (A, B, and M), one for the subject number and one for the dependent variable. There were as many rows as there were data points. The aligned rank transformations were then calculated as follows:

(1) Eight additional columns were obtained in the data table containing the marginal means for the levels and level combinations of each source of variation: $\bar{y}_{i..}$, $\bar{y}_{.j.}$, $\bar{y}_{ij.}$, $\bar{y}_{..k}$, $\bar{y}_{i.k}$, $\bar{y}_{.jk}$, $\bar{y}_{ijk}$, plus the average value for all within-subject measures of each simulated participant, $\bar{s}_{l.(ij)}$.
(2) The error terms were obtained on two additional columns. The between-subject error term was calculated by using the formula $s_{l.(ij)} = \bar{s}_{l.(ij)} - \bar{y}_{ij.}$ The within-subject error term was obtained as residuals of a three-way mixed ANOVA run on the original data by the SAS general linear models (GLM) procedure.
(3) The linear combinations of marginal means and error terms for the experimental effects in Table 2 were calculated in seven new columns: $\hat{y}^{A}_{ijkl}$, $\hat{y}^{B}_{ijkl}$, $\hat{y}^{AB}_{ijkl}$, $\hat{y}^{M}_{ijkl}$, $\hat{y}^{AM}_{ijkl}$, $\hat{y}^{BM}_{ijkl}$, $\hat{y}^{ABM}_{ijkl}$.
(4) A rank transformation procedure was applied to the above linear combinations to obtain the seven new aligned rank variables for the experimental effects in Table 2: $R(\hat{y}^{A}_{ijkl})$, $R(\hat{y}^{B}_{ijkl})$, $R(\hat{y}^{AB}_{ijkl})$, $R(\hat{y}^{M}_{ijkl})$, $R(\hat{y}^{AM}_{ijkl})$, $R(\hat{y}^{BM}_{ijkl})$, $R(\hat{y}^{ABM}_{ijkl})$.

In obtaining the test statistics, the new aligned rank variables were transposed to a multivariate format so that the GLM repeated measures procedure could be applied. Thus, sphericity diagnostics and remedial statistics could be obtained. The resulting table had one variable for each between-subject experimental effect (A and B), and a set of four (m) within-subject variables for each one of the aligned rank transformations above. A separate repeated measures analysis was then conducted on each set and only results for the experimental effect that corresponded to each aligned rank variable were in turn recorded from the output table.

Lecoutre's $\tilde{\varepsilon}$ was calculated from the Greenhouse–Geisser $\hat{\varepsilon}$ statistic (equation (3)). The adjusted numerator and denominator degrees of freedom for L_{AR} were then computed by multiplying the standard degrees of freedom by $\tilde{\varepsilon}$. The observed L_{AR} probability values were then obtained from the empirical F_{AR} value and the adjusted degrees of freedom by using the SAS F probability function. F_{AR} and H_{AR} statistics were directly recorded from the output tables.

3.3. Data analysis procedure

Performance comparisons were made between F and F_{AR} for between-subject tests and for within-subject tests when the sphericity assumption held. Comparisons were made between L and L_{AR}, H and H_{AR} for within-subject tests when the sphericity assumption was violated. A decision criterion of $\alpha = .05$ was used. Results of the adjusted Huynh–Feldt F procedure were also described for the raw and aligned rank scales (HF and HF_{AR}). Comparative results of F and F_{AR} statistics were also included under lack of sphericity for replication purposes.

Null hypothesis rejection rates were tabulated for each of the sources of variation (A, B, M, A × B, A × M, B × M, A × B × M) and experimental condition, yielding a table with 7 × 98 rows and a column for each of the above statistics. In conditions where an experimental effect was absent a two-tailed binomial test was used to detect Type I error deviations from the nominal $\alpha = .05$ rate. For a thousand replications, rates that were either larger than .0635 or smaller than .0366 were detected as deviant, and were respectively considered as liberal or conservative. In conditions where an experimental effect was present a two-tailed McNemar test was used to test differences in power rates between raw and aligned rank statistics since the data fed to both were the same and their corresponding results were therefore

correlated. Sphericity and non-sphericity conditions for between-subject effects were pooled.

Frequencies of deviance from the nominal $\alpha = .05$ Type I error rate in either the liberal or conservative directions were then summarized for each effect category (main effects, two-way interaction and three-way interaction), comparison type (between or within subject), distribution, sample size, and within-subject sphericity condition. Frequencies of detection of power differences between the raw and aligned rank tests were also summarized for each effect category under the same above conditions. Each particular detection frequency was obtained from results of the sources of variation contained in an effect category over the four conditions defined by presence or absence of effects in the two remaining categories. As an example, the detection frequencies for the between-subject main effect category were obtained, for each distribution and sample size, from the number of detections observed on binomial or McNemar tests for the two between-subject main effects (A and B) over the four conditions defined by presence and absence of two-way and three-way interactions. Tables were made containing both the inferential detection frequencies and descriptive statistics of Type I error and power rates for their corresponding results. Graphical representations of mean Type I error and mean power rates were also displayed.

4. Results

4.1. Large sample (n = 30)

We begin with the comparison between F and F_{AR}. When the sphericity assumption held, Type I error rates were similar with normal distributions (Table 4). For exponential or double exponential distributions, performance of the F_{AR} test was similar or slightly closer to the nominal $\alpha = .05$ rate in terms of the number of times deviance was detected by the binomial test (Tables 5 and 6). The F test had a slight power advantage with normal distributions (Table 4). When the data followed the other two distributions the power advantage favoured the F_{AR} test (Tables 5 and 6) and was especially large for the exponential with an overall average increase of .19 across effect categories (Figure 1).

Turning now to the comparison between L and L_{AR}, when the sphericity assumption did not hold and the distribution was normal the L statistic showed similar performance or was closer to the nominal Type I error rate in terms of liberal detections (Table 4). No detectable differences were observed for exponential and double exponential distributions (Tables 5 and 6). In terms of power, the L test had a slight advantage for normal distributions (Table 4). When the data followed the other two distributions the L_{AR} test was at an advantage (Tables 5 and 6), especially so for the exponential with an overall average increase of .22 across effect categories (Figures 2 and 3).

Finally, we compare H and $H_{AR.}$ When the sphericity assumption was violated and the distribution was normal the H_{AR} test showed similar performance in terms of Type I error (Table 4). A similar or less conservative and closer performance to the nominal $\alpha = .05$ rate was observed for the H_{AR} test with the exponential distribution (Table 5). No detectable differences were obtained for the double exponential (Table 6). In terms of power, the H test had a slight advantage for normal distributions (Table 4). When the data followed the other two distributions the H_{AR} test was favoured (Tables 5 and 6), especially so for the exponential with an overall average increase of .185 across effect categories (Figures 2 and 3).

Table 4. Performance comparisons between raw and aligned rank statistics with a normal distribution: Estimates, frequencies of inferential detection of deviations from the nominal $\alpha = .05$ Type I error rate and of differences in power

		Type I error rates[a]						Power rates[a]					
		Between-subject tests			Within-subjects tests			Between-subject tests			Within-subject tests		
Effect	Statistic	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}
$n = 30$													
Main	F, F_{AR}	53^{1}, 53^{1}	6.7, 7.2	4.2	57^{0}, 56^{0}	5.9, 3.7	4.3	400^{7}, 384^{0}	35, 35	6.2	733^{4}, 714^{0}	16, 17	1.5
	L, L_{AR}				58^{0}, 55^{0}	6.7, 5.1	2.8				465.5^{1}, 453.5^{0}	10, 10	4.4
	H, H_{AR}				48^{0}, 50.5^{0}	6.6, 7.1	2.1				639^{2}, 615.5^{0}	13, 9.6	11.4
Two-way interaction	F, F_{AR}	49^{0}, 51^{0}	8.1, 6.4	5.4	49^{0}, 51.5^{1}	5.7, 7.3	6.9	392^{3}, 376.5^{0}	46, 44	8.7	728^{5}, 715^{0}	11, 6.8	9.2
	L, L_{AR}				48^{0}, 49^{1}	7.4, 7.3	2.7				445^{2}, 434^{0}	18, 17	5.7
	H, H_{AR}				50^{0}, 50^{0}	6.1, 6.7	2.7				631^{6}, 606^{0}	18, 17	10.2
Three-way interaction	F, F_{AR}				55^{2}, 59^{1}	14, 11.5	3.9				730^{3}, 713.5^{0}	22, 22	2.6
	L, L_{AR}				49^{0}, 51^{0}	8.3, 9.4	1.3				451^{2}, 436^{0}	14, 13	4.7
	H, H_{AR}				43^{0}, 47^{0}	3.4, 7.5	5.3				628.5^{3}, 609^{0}	8.7, 14	9.5
Total		51^{1}, 52^{1}			50.7^{2}, 52.1^{3}			396^{10}, 380.25^{0}			605.7^{28}, 588.5^{0}		
$n = 10$													
Main	F, F_{AR}	54^{1}, 52^{1}	7.3, 7.1	2.8	48^{0}, 49.5^{0}	7.2, 5.8	2.9	161^{0}, 159^{0}	16, 14	8.0	272.5^{0}, 263^{0}	13, 12	3.6
	L, L_{AR}				51^{0}, 52^{0}	2.1, 2.2	2.2				161.5^{0}, 160.5^{0}	20, 16	6.6
	H, H_{AR}				48^{0}, 54^{0}	7.8, 4.7	3.4				225^{1}, 220^{0}	19, 13	12
Two-way interaction	F, F_{AR}	49^{0}, 55^{1}	7.4, 6.9	6.8	$^{1}53^{1}$, 56.5^{2}	10, 8.4	4.9	166^{0}, 164^{0}	22, 19	9.7	273^{2}, 265^{0}	13, 14	9.5
	L, L_{AR}				52^{2}, 54^{1}	12, 13	5.8				162^{0}, 164^{0}	13, 10	6.5
	H, H_{AR}				55^{1}, 56^{1}	6.3, 13	8.0				229^{0}, 224^{0}	16, 14	3.2

Continued

Table 4. (*Continued*)

		Type I error rates[a]						Power rates[a]					
		Between-subject tests			Within-subjects tests			Between-subject tests			Within-subject tests		
Effect	Statistic	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}
Three-way	F, F_{AR}				52^{0}, 52^{0}	7.7, 2.2	8.9				281^{0}, 280.5^{0}	6.2, 9.2	5.9
interaction	L, L_{AR}				52.5^{0}, 56^{0}	10, 10	4.3				168^{0}, 163^{0}	13, 10	7.0
	H, H_{AR}				54^{1}, 56^{2}	8.4, 10	3.8				221^{0}, 220.5^{1}	5.6, 9	14
Total		51.5^{1}, 53.5^{2}			${}^{1}51.7^{5}$, ${}^{0}54^{6}$			163.5^{0}, 161.5^{0}			221.4^{3}, 217.8^{1}		

Note. F and F_{AR} tests are reported only in conditions where assumptions hold. L, L_{AR}, H, and H_{AR} are reported only on within-subject tests in conditions where the sphericity assumption was violated. Means, standard deviations (SDs), standard deviations of differences (SD_{dif}) and inferential detection frequencies represented as superscripts were obtained from results for the different sources of variation of one effect category (main effects, two-way or three-way interaction) summarized over conditions for the two remaining effect categories. Right and left superscripts on mean Type I error values represent the number of liberal and conservative deviations detected by the binomial test, respectively. Superscripts on mean power rates indicate the number of comparative advantages detected by the McNemar test. Maximum possible detection counts for either test were 8 and 4 for between-subject main and two-way interaction categories, and 4, 8, and 4 for within-subject main effect, two, and three-way interaction categories. Absence of a superscript indicates a 0.

[a]Values have been multiplied by 10^{3}.

Table 5. Performance comparisons between raw and aligned rank statistics with a exponential distribution: Estimates, frequencies of inferential detection of deviations from the nominal $\alpha = .05$ Type I error rate and of differences in power

		Type I error rates[a]						Power rates[a]					
		Between-subject tests			Within-subject tests			Between-subject tests			Within-subject tests		
Effect	Statistic	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}
$n = 30$													
Main	F, F_{AR}	$47^{0}, 50^{0}$	5.6, 6.7	7.2	$47^{0}, 44^{0}$	7.3, 4.3	4.3	$412^{0}, 594^{8}$	33, 20	28	$724.5^{0}, 920^{4}$	24, 12	16
	L, L_{AR}				$44^{0}, 49^{0}$	5.2, 2.1	3.3				$465^{0}, 680^{4}$	15, 18	22
	H, H_{AR}				$53^{0}, 50^{0}$	4.6, 5.6	4.0				$646.5^{0}, 833.5^{4}$	22, 8.2	19
Two-way	F, F_{AR}	$^{1}47, 50^{0}$	5.7, 9.4	10	$48^{0}, 47^{0}$	3.8, 7.5	7.2	$398^{0}, 587^{4}$	41, 26	20	$734^{0}, 924^{s}$	21, 12.5	17
interaction	L, L_{AR}				$47^{0}, 49^{0}$	4.8, 4.1	2.4				$463^{0}, 688^{8}$	10, 16	13
	H, H_{AR}				$^{2}42, 48.5^{0}$	6.2, 7.6	4.5				$651^{0}, 839^{8}$	10, 11	9.3
Three-way	F, F_{AR}				$52^{1}, 56^{0}$	10, 7.2	4.1				$728^{0}, 918^{4}$	9.3, 6.9	9.4
interaction	L, L_{AR}				$49^{0}, 52^{0}$	9.1, 6.4	5.2				$453^{0}, 678.5^{4}$	10, 25	19
	H, H_{AR}				$49^{0}, 53.5^{0}$	4.3, 8.7	10				$656^{0}, 836^{4}$	13, 18	15
Total		$^{1}47^{0}, {}^{0}50^{0}$			$^{2}47.9^{1}, 49.9^{0}$			$405^{0}, 590.5^{12}$			$613.4^{0}, 813^{48}$		
$n = 10$													
Main	F, F_{AR}	$52^{0}, 56^{2}$	7.7, 7.2	4.9	$42^{0}, 46.5^{0}$	5.0, 7.2	5.5	$181^{0}, 242^{8}$	14, 13	12	$286^{0}, 406^{4}$	6.1, 8.4	6.4
	L, L_{AR}				$49^{0}, 58^{0}$	5.0, 7.3	5.1				$169^{0}, 235^{4}$	14, 20	8.5
	H, H_{AR}				$51^{1}, 52^{0}$	10, 7.4	5.6				$259^{0}, 330^{4}$	10, 14	8.3
Two-way	F, F_{AR}	$49^{0}, 50.5^{0}$	5.1, 9.5	7.4	$49^{0}, 52^{2}$	8.1, 9.4	6.3	$170^{0}, 232^{4}$	19, 18	11	$288^{0}, 420^{8}$	15, 22	11
interaction	L, L_{AR}				$46^{1}, 55^{2}$	9.1, 11	3.8				$165^{0}, 235^{8}$	15, 10	6.2
	H, H_{AR}				$45^{0}, 53^{0}$	5.3, 5.4	3.5				$255^{0}, 344^{8}$	16, 16	8.9
Three-way	F, F_{AR}				$54^{1}, 61^{1}$	8.2, 10	11				$292^{0}, 418.5^{4}$	13, 15	15
interaction	L, L_{AR}				$49^{0}, 50^{0}$	6.9, 5.9	1.2				$175.5^{0}, 247^{4}$	7.9, 8.7	11
	H, H_{AR}				$^{2}39.5, {}^{0}54.$	8.1, 2.4	5.9				$248^{0}, 348^{4}$	4.6, 11	15
Total		$50.5^{0}, 53.25^{2}$			$^{2}47.2^{3}, 53.5^{5}$			$175.5^{0}, 237^{12}$			$237.5^{0}, 331.5^{48}$		

Note. See notes to Table 4.

[a]Values have been multiplied by 10^{3}.

Table 6. Performance comparisons between raw and aligned rank statistics with a double exponential distribution: Estimates, frequencies of inferential detection of deviations from the nominal $\alpha = .05$ Type I error rate and of differences in power

		Type I error rates[a]						Power rates[a]					
		Between-subject tests			Within-subject tests			Between-subject tests			Within-subject tests		
Effect	Statistic	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}	Means	*SD*s	SD_{dif}
$n = 30$													
Main	F, F_{AR}	$52.5^{0}, 52^{0}$	6.5, 5.9	6.6	$50^{0}, 52^{0}$	3.6, 5.1	4.1	$403^{0}, 441^{8}$	33, 28	11	$724^{0}, 761^{4}$	2.3, 3.2	3.2
	L, L_{AR}				$^{1}38.5, ^{1}40$	5.0, 6.2	3.1				$447^{0}, 495^{4}$	7.0, 9.0	8.7
	H, H_{AR}				$^{1}45, ^{1}45$	7.4, 9.6	3.6				$641^{0}, 679^{4}$	16, 20	9.7
Two-way interaction	F, F_{AR}	$51.5^{0}, 55^{0}$	5, 4.1	5.0	$^{1}48^{1}, ^{1}49^{0}$	10, 8.5	5.7	$416^{0}, 451^{4}$	39, 37	12	$722.5^{0},760.5^{8}$	11, 8.9	8.4
	L, L_{AR}				$49^{0}, 50^{0}$	7.3, 6.5	5.9				$452^{0}, 500^{8}$	15, 19	12
	H, H_{AR}				$48^{0}, 49^{0}$	4.6, 7.5	5.7				$645^{0}, 680^{7}$	19, 14	9.0
Three-way interaction	F, F_{AR}				$50^{0}, 56^{0}$	0.8, 2.2	1.5				$724.5^{0}, 764^{4}$	2.1, 5.2	4.4
	L, L_{AR}				$49.5^{0}, 49^{0}$	7.3, 5.7	2.9				$446^{0}, 491^{4}$	8.5, 7.9	6.2
	H, H_{AR}				$47^{0}, 50^{0}$	7.4, 6.3	6.1				$643^{0}, 673^{3}$	12, 8.0	13
Total		$52^{0}, 53.5^{0}$			$^{3}47.2^{1}, ^{3}48.9^{0}$			$409.5^{0}, 446^{12}$			$605^{0}, 644.8^{46}$		
$n = 10$													
Main	F, F_{AR}	$56^{1}, 57^{2}$	8.3, 7.3	6.0	$50.5^{0}, 50.\ 5^{0}$	6.6, 5.4	7.3	$166^{0}, 179^{6}$	18, 15	7.1	$278^{0}, 303^{3}$	13, 19	12
	L, L_{AR}				$54.5^{0}, 52^{0}$	5.8, 5.5	4.7				$161^{0}, 182^{3}$	4.1, 4.9	5.3
	H, H_{AR}				$^{1}44, ^{0}50$	5.7, 4.0	4.0				$229^{0}, 249^{3}$	3.0, 5.3	5.8
Two-way interaction	F, F_{AR}	$52^{0}, 55^{1}$	6.5, 7.9	5.8	$^{2}45, ^{1}46$	9.8, 8.1	3.9	$174^{0}, 184.5^{2}$	15, 16	8.5	$286^{0}, 307^{5}$	15, 11	6.1
	L, L_{AR}				$^{1}50, ^{1}50^{2}$	9.4, 10	4.8				$165^{0}, 180^{6}$	15, 12	5.9
	H, H_{AR}				$48^{0}, 55^{2}$	3.2, 6.3	4.0				$244.5^{0}, 257^{1}$	18, 19	6.6
Three-way interaction	F, F_{AR}				$52^{0}, 52^{0}$	4.5, 6.0	5.0				$292^{0}, 312^{3}$	11, 16	7.8
	L, L_{AR}				$54^{0}, 59^{2}$	7.7, 7.1	1.0				$171^{0}, 182.5^{1}$	14, 15	7.3
	H, H_{AR}				$50^{0}, 53.5^{0}$	2.2, 4.8	2.6				$230^{0}, 255^{3}$	18, 7.2	14
Total		$54^{1}, 56^{3}$			$^{4}49.7^{0}, ^{2}52^{6}$:			$170^{0}, 181.75^{8}$			$228.5^{0}, 247.5^{28}$		

Note. See notes to Table 4.

[a]Values have been multiplied by 10^{3}.

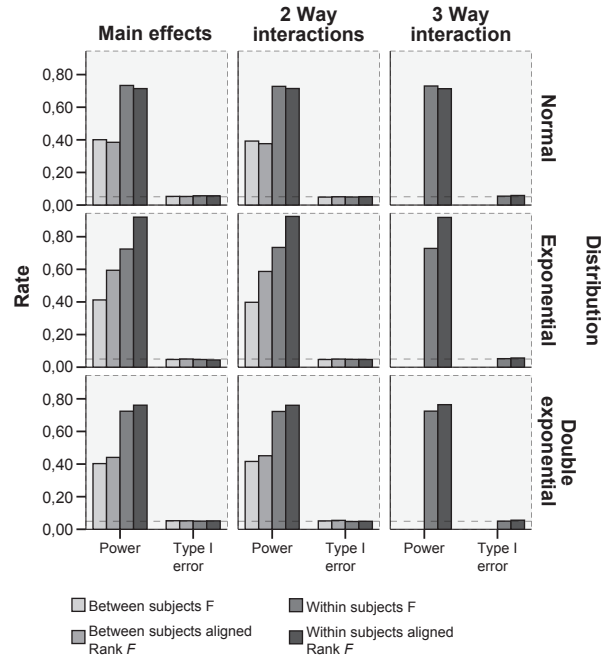

Figure 1. Type I error and power rates for the *F* tests run on the raw and aligned rank scales when assumptions hold and $n = 30$.

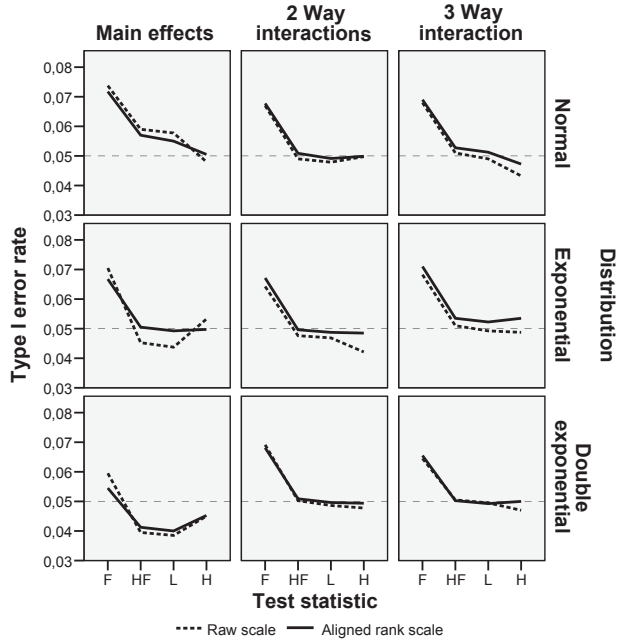

Figure 2. Type I error rates for the uncorrected *F*, the univariate adjusted Huynh–Feldt (*HF*), Lecoutre (*L*), and the multivariate Hotelling (*H*) tests run on the raw and aligned rank scales under conditions of no covariance sphericity and $n = 30$.

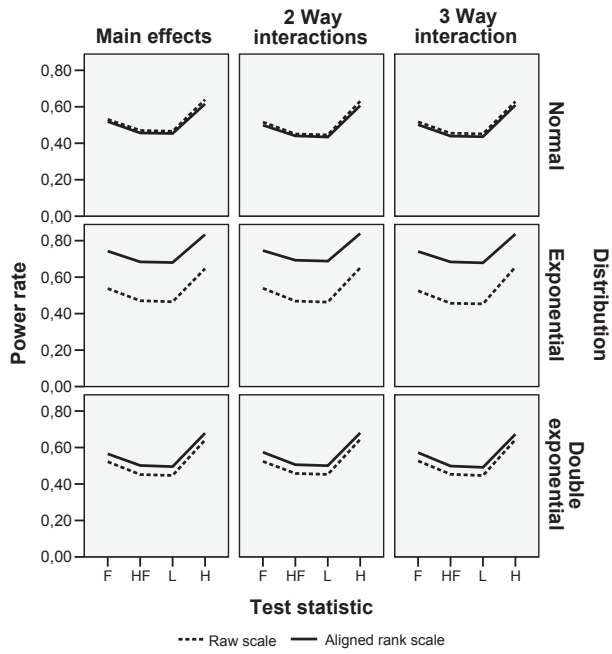

Figure 3. Power rates for the uncorrected *F*, the univariate adjusted Huynh–Feldt (*HF*), Lecoutre (*L*), and the multivariate Hotelling (*H*) tests run on the raw and aligned rank scales under conditions of no covariance sphericity and $n = 30$.

4.2. Small sample ($n = 10$)

We begin again with the comparison between *F* and F_{AR}. When the sphericity assumption held, the F_{AR} statistic showed slightly more liberal Type I error rates with all three distributions (Tables 4–6). In terms of power, the *F* test had a small advantage for normal distributions (Table 4). When the data followed the other two distributions the F_{AR} test held the advantage (Tables 5 and 6), especially so for the exponential with an overall average increase of .10 across effect categories (Figure S1 as a web supplement).

Turning now to the comparison between *L* and L_{AR}, when the sphericity assumption did not hold, the L_{AR} statistic showed detectably more liberal levels of Type I error only with exponential and double exponential distributions (Tables 4–6). In terms of power, the L_{AR} test had similar levels for normal distributions (Table 4) and an advantage for the other two distributions (Tables 5 and 6). This was larger when the data followed the exponential with an overall average power increase of .07 across effect categories (Figures S2 and S3 as a web supplement).

It was also interesting to observe less deviant Type I error rates with normal distributions and small samples for both *L* versus *HF* statistics (3 vs. 4 liberal deviation detections) and for their aligned rank versions L_{AR} versus HF_{AR} (1 vs. 4 liberal deviation detections). The L_{AR} test also showed less deviant Type I error rates in comparison with the HF_{AR} test with the exponential distribution (2 vs. 4 liberal deviation detections) (Figure S2 as a web supplement).

Finally, we compare *H* and H_{AR}. When the sphericity assumption was violated, the H_{AR} statistic showed slightly more detections of liberal deviations in Type I error levels with

normal and double exponential distributions (Tables 4 and 6). When the data followed the exponential and double exponential distributions H_{AR} showed fewer detections of conservative deviations from the nominal α rate (Tables 5 and 6). In terms of power, the H_{AR} test performed similarly to the H test with normal distributions (Table 4). H_{AR} had a power advantage with the other two distributions (Tables 5 and 6), which was larger for the exponential with an overall average increase of .09 across effect categories (Figures S2 and S3 as a web supplement).

5. Discussion

For normal distributions the classical ANOVA statistics performed better than the non-parametric ones in most cases. The Lecoutre correction reduced Type I error rates in small samples in comparison with the Huynh–Feldt adjusted F test reported in standard analysis packages, supporting previous analytical work (Lecoutre, 1991). Its use by researchers may benefit accuracy of results in the analysis of three-way repeated measures designs. Twenty years after becoming aware of the mistake, it would be advisable for statistical companies to take steps to correct it.

When the assumption of normality did not hold, aligned rank statistics showed improved performance over the classical ANOVA tests when sample sizes were large. The advantage occurred for both the exponential (asymmetric) and double exponential (symmetric heavy-tailed) distributions with similar levels of Type I error and increases in power that were especially large with the exponential. Like commonly used rank-sum tests such as Kruskal–Wallis, large-sample aligned rank analyses for three-way mixed model ANOVA sources of variation can be implemented by widely used statistical packages.

For non-normal distributions and small sample sizes aligned rank statistics showed smaller increases in power but also some inflation of Type I error. Liberal decision rates in small samples have also been observed in previous studies on the aligned rank test of the two-way interaction (Beasley, 2002; Richter & Payton, 2005). A proposed solution has been to apply a modified Box-type small-sample degrees-of-freedom adjustment to aligned rank statistics (Box, 1954; Brunner, Dette, & Munk, 1997; Richter & Payton, 2005). In the analysis of the two-way interaction for independent samples and at least seven observations per group, this adjusted aligned rank procedure has allowed for controlling Type I error at nominal levels while maintaining a power advantage over the F test. It has also shown improved performance over alternative non-parametric statistics such as rank versions of the Wald test using the same small-sample adjustments (Akritas & Arnold, 1994; Brunner *et al.*, 1997; Richter & Payton, 2005). It would therefore be of interest to compare the performance of adjusted aligned rank tests on the analysis of three-way mixed models with adjusted rank-based Wald tests or with alternative rank methods requiring specialized statistical software (Crimin, Abebe, & McKean, 2008; Erceg-Hurn & Mirosevich, 2008). Results could be valuable in helping to clarify and simplify the use of non-parametric tools by behavioural researchers.

6. A case study

In a study on the use of visual language four samples of 30 children were obtained, stratified by visual ability (blind and sighted) and gender (boys and girls) (Rosel, 1982; Rosel, Caballer, & Jara, 2005). Two linguistic measures were recorded for each boy or girl:

the ability to narrate a story invented by the child, based on cue words such as family, parents and friendship; and the ability to describe one of the story characters. The study therefore follows a mixed design with two between-subject factors and one within-subject factor, all of them with two levels each. The statistical linear model specified in equation (1) can therefore be applied.

A standard analysis showed statistical significance only on the within-subject linguistic measure main effect, $F(1, 116) = 32.76, p = .0001$. However, diagnostic residual analyses revealed a marked skewness of 2.24 on the adjusted Fisher–Pearson standardized moment coefficient. The normality assumption was rejected according to the Shapiro–Wilk test, $W = 0.85, p < .0001$.

Results from the previous simulation reveal an increase in statistical power for the aligned rank test in comparison with the F statistic in cases of pronounced asymmetry, as well as similar Type I error levels. The F_{AR} statistic was for that reason applied to this set of data. The same main effect for linguistic ability was also obtained, F_{AR} (1, 116) = 38.61, $p = .0001$, but also an effect of gender × visual ability, F_{AR} (1,116) = 5.51, $p = .0206$, and a three-way interaction of gender × visual × linguistic ability, F_{AR} (1,116) = 5.14, $p = .0252$ (Figure 4). For comparison, the corresponding F test results for gender × visual ability were $F(1, 116) = 0.77$, $p = .3814$, and for gender × visual ability × linguistic ability, $F(1, 116) = 0.92$, $p = .3392$. The three-way interaction with the aligned rank test indicates that in narration measures positive differences between sighted and blind children were observed in both boys and girls. However, in description measures these positive differences were only observed in boys, since sighted girls performed worse than blind girls.

The alternative analysis allowed for detection of gender differences not observed with other statistical methods (Rosel *et al.*, 2005). Whereas sighted boys scored higher on both

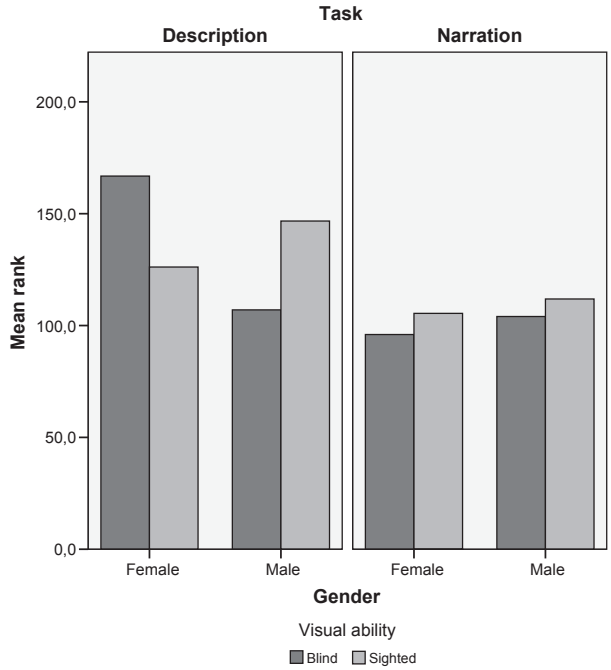

Figure 4. Mean rank score of linguistic abilities as a function of gender and visual skill.

linguistic measures than blind boys, sighted girls only scored higher on the narration task but lower on the character description task in comparison to blind girls. This may be due to compensatory mechanisms of visual impairment that are specific to gender and verbal skill. Gender differences in language are not general but ability-specific (Hyde & Linn, 1988). In blind children they have not been found in tasks such as word definitions (Kemter, 1999). The alternative analysis therefore provides new answers and allows new questions to be raised on the processes and adaptive function of gender-dependent communication skills of children with sensory disabilities.

This example illustrates the usefulness of diagnostic procedures in the selection of test statistics. One recurrent problem in behavioural studies is low statistical power (Maxwell, 2004). One consequence is the generation of inconclusive results in the literature. Analysis strategies that are tailored to the characteristics of the data may provide increased sensitivity to detect phenomena of applied or theoretical interest.

Acknowledgements

We would like to thank Robert Steiner and David W. Smith of New Mexico State University for their support in conducting this study.

References

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. doi:10.1080/01621459.1994.10476475

Algina, J., & Keselman, H. J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods*, *2* (2), 208–218. doi:10.1037/1082-989X.2.2.208

Beasley, T. M. (2002). Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multivariate Behavioral Research*, *37* (2), 197–226. doi:10.1207/S15327906MBR3702_02

Blair, R. C., Sawilowsky, S. S., & Higgins, J. J. (1987). Limitations of the rank transform statistic in tests for interactions. *Communications in Statistics: Simulation and Computation*, *16* (4), 1133–1145. doi:10.1080/03610918708812642

Box, G. E. P. (1954). Some theorems on quadratic forms in the study of analysis of variance problems: Effects of inequality of variances in one-way classifications. *Annals of Mathematical Statistics*, *25* (2), 290–302. doi:10.1214/aoms/1177728786

Brunner, E., Dette, H., & Munk, A. (1997). Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association*, *92*, 1494–1502. doi:10.2307/2965420

Chen, R. S., & Dunlap, W. (1994). A Monte Carlo study on the performance of a corrected formula for $\hat{\varepsilon}$ suggested by Lecoutre. *Journal of Educational and Behavioral Statistics*, *19* (2), 119–126. doi:10.3102/10769986019002119

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician*, *35* (3), 124–129. doi:10.2307/2683975

Crimin, K., Abebe, A., & McKean, J. W. (2008). Robust general linear models and graphics via a user interface. *Journal of Modern Applied Statistical Methods*, *7* (1), 318–330. Retrieved from: http://www.jmasm.com/may-2008-vol-7-no-1/

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods. *American Psychologist*, *63* (7), 591–601. doi:10.1037/0003-066X.63.7.591

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon–Mann–Whitney or *t*-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, *4*, 1–39. doi:10.1214/09-SS051

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate non-normal distributions: Extending the Fleishman power method. *Psychometrika*, *64* (1), 25–35. doi:10.1007/BF02294317

Hocking, R. R. (1996). *Methods and applications of linear models*. New York: Wiley.

Hodges, J. L., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *Annals of Mathematical Statistics*, *33* (2), 482–497. doi:10.1214/aoms/1177704575

Huynh, H., & Feldt, L. (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* (1), 69–82. doi:10.3102/10769986001001069

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta analysis. *Psychological Bulletin*, *104*, 53–69. doi:10.1037/0033-2909.104.1.53

Kemter, P. (1999). Concept formation and space perception in the blind – a summary of behavioral psychological studies. *Die Rehabilitation*, *38* (1), 27–32

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*(4), 371–372. doi: 10.3102/10769986016004371

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences and remedies. *Psychological Methods*, *9*(2), 147–163. doi: 10.1037/1082-989X.9.2.147

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105* (1), 156–166. doi:10.1037/0033-2909.105.1.156

Richter, S. J., & Payton, M. E. (2005). An improvement to the aligned rank statistic for two-factor analysis of variance. *Journal of Applied Statistical Science*, *14* (3–4), 225–235.

Rosel, J. (1982). *Niveles cognitivos y desarrollo léxico-sintáctico. Estudio evolutivo-diferencial en escolares ciegos y videntes* [Cognitive stages and lexical-syntactic development. A developmental-differential study in blind and sighted school children], Unpublished doctoral dissertation, University of Salamanca, Spain.

Rosel, J., Caballer, C., & Jara, P. (2005). Verbalism in the narrative language of children who are blind and sighted. *Journal of Visual Impairment and Blindness*, *99* (7), 413–425.

Salter, K. C., & Fawcett, R. F. (1993). The ART test of interaction: A robust and powerful rank test of interaction in factorial models. *Communications in Statistics: Simulation and Computation*, *22*, 137–153. doi:10.1080/03610919308813085

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the Type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics*, *14*, 255–267. doi:10.3102/10769986014003255

Shah, D. A., & Madden, L. V. (2004). Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology*, *94* (1), 33–43. doi:10.1094/PHYTO.2004.94.1.33

Thompson, G. L. (1991a). A note on the rank transform for interactions. *Biometrika*, *78*, 697–701. doi:10.1093/biomet/78.3.697

Thompson, G. L. (1991b). A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association*, *86*, 410–419. doi:10.2307/2290586

Toothaker, L. E., & Newman, D. A. (1994). Nonparametric competitors to the two-way ANOVA. *Journal of Educational and Behavioral Statistics*, *19*, 237–273. doi:10.3102/10769986019003237

Vallejo, G., & Lozano, L. M. (2006). Modelos de análisis para los diseños multivariados de medidas repetidas [Multivariate repeated measures designs]. *Psicothema*, *18* (2), 293–299.

Vargha, A., & Delaney, H. D. (1998). The Kruskal–Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics*, *23* (2), 170–192. doi:10.3102/10769986023002170

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8* (3), 254–274. doi:10.1037/1082-989X.8.3.254

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

Received 4 January 2013; revised version received 10 August 2013

Supporting Information

The following supporting information may be found in the online edition of the article:

Figure S1. Type I error and power rates for the *F* tests run on the raw and aligned rank scales when assumptions hold and $n = 10$.
Figure S2. Type I error rates for the uncorrected *F*, the univariate adjusted Huyhn Feldt (*HF*), Lecoutre (*L*), and the multivariate Hotelling (*H*) tests run on the raw and aligned rank scales under conditions of no covariance sphericity and $n = 10$.
Figure S3. Power rates for the uncorrected *F*, the univariate adjusted Huyhn Feldt (*HF*), Lecoutre (*L*), and the multivariate Hotelling (*H*) tests run on the raw and aligned rank scales under conditions of no covariance sphericity and $n = 10$.