Psychology Science, Volume 49, 2007 (3), p. 209-222

Some common features and some differences between the parametric ANOVA for repeated measures and the Friedman ANOVA for ranked data

WILLI HAGER[1]

Abstract

Some relationships between the parametric *ANOVA* of repeated measures and its nonparametric counterpart, the *ANOVA* for ranks after Friedman, are discussed. The main reason for some marked differences between both procedures arises from the fact that the mean correlation between the experimental conditions, i.e. r_B, can vary between the limits $-1 \leq r_B \leq +1$ for the parametric *ANOVA* and usually is greater than zero - only if this is the case, precision is enhanced. In contrast, this correlation always is negative for the Friedman *ANOVA* and only depends on the number K of experimental conditions: $r_{R,B} = -1/(K - 1)$. - In addition, it is stressed that the nonparametric rank tests can be replaced by their parametric counterparts without risking divergent decisions about the statistical hypotheses being tested. The necessary formulae and the respective effect sizes are presented.

Key words: Parametric *ANOVA* for repeated measures and *ANOVA* for ranks after Friedman, parametric tests and effect sizes for ranked data

[1] Prof. Dr. Willi Hager, Georg-Elias-Müller-Institut für Psychologie, Goßlerstr. 14, 37073 Göttingen, Germany; email: whager@uni-goettingen.de

The present article deals with the univariate analysis of variance (*ANOVA*) for repeated measures and interval scaled data and compares it to the Friedman *ANOVA* for ranks. The similarities and differences discussed can be expected to be known by statisticians, but may be new to empirical psychological researchers, so they will be examined here in some detail as the main focus of this article. I shall mostly refer to some textbooks well-known to psychological researchers and to some articles cited therein. - In the second part, the analysis of ranked data by parametric tests is addressed.

Designs with repeated measures are characterized by the fact that the same subjects (Ss) usually, although not necessarily, are observed under the K treatment conditions of a factor called B here, resulting in K measures per subject, y_{ik}. In view of this fact, the formal precondition for a multivariate analysis is given - partly in order to circumvent the circularity assumption associated with the univariate F test, which refers to a certain structure of the variance-covariance matrices (see below). So, some authors categorically urge researchers to perform multivariate analyses (e.g., Erdfelder, Buchner, Faul & Brandt, 2004, p. 161), but others maintain, "that the general use of … [multivariate; WH] statistics cannot be recommended …" (Rouanet & Lépine, 1970, p. 149; see also O'Brien & Kaiser, 1985, p. 329). The relative power of uni- and multivariate tests, however, depends on various factors, which are discussed comprehensively by Kirk (1995, p. 281-282). Tanguma (1999, p. 250) comes to the conclusion that multivariate tests should be preferred, if the number of Ss is only slightly larger than the number K of treatment conditions. In all other cases the univariate analysis is at least as powerful as the multivariate analysis. But this argument is of restricted value since power analysis for univariate and for multivariate analyses can be performed. It may happen, however, that more Ss are needed for the multivariate than for the univariate analysis. In addition, Girden (1992, pp. 25-26, p. 39) addresses the multivariate procedures and concludes, that there is no clear-cut rule, which enables to prefer one type of analysis over the other. - Moreover, in modern textbooks on experimental design, the univariate analysis seems to be preferred (cf. Bortz, 2005; Keppel & Wickens, 2004; Kirk, 1995; Winer, Brown & Michels, 1991, to name but a few). As far as I am concerned, I, too, prefer the univariate analysis of repeated measures data, although not because of the reasons given above, but it would lead too far to deal with these reasons here. - Keselman, Algina, Boik and Wilcox (1999) present robust alternative procedures.

The parametric one-way *ANOVA* for repeated measures

Let us begin with a psychological hypothesis referring to abstract variables such as imagery or frustration, which after choosing empirical variables for the theoretical ones ("operationalizations") leads to predicting the statistical hypothesis to hold, that with $K = 4$ experimental conditions not all means μ_k are equal; this is the alternative hypothesis (H_1) of a one-way *ANOVA*. - After execution of the study the interval scaled raw scores y_{ik} in Table 1 have resulted. The restriction to five Ss only serves to cut short the computations.

The following strictly additive model without interaction parameter is chosen as underlying the data (Winer et al., 1991, p. 261, p. 264):

$$y_{ik} = \mu + \beta_k + \underline{\pi_i} + \varepsilon_{ik}, \quad (1)$$

with: μ: grand mean, β_k: effect of factor B, usually conceived of as a fixed effect; π_i: effect of subject i, usually interpreted as random effect, ε_{ik}: individual error term, here containing the interaction term, which is interpreted as error and not as a systematic source of variation. This model generally is enhanced by several side conditions necessary for reparametrization and distribution assumptions, hold necessary for the validity of the statistical test addressed below (the details are of no relevance for the aim of this paper and can be found in Winer et al., 1991, pp. 261-262, and in Rasch & Kubinger, 2006, pp. 332-333, amongst others).

The sums of squares (*SS*) can be found in the following *ANOVA* table (for their computation see, e.g., Winer et al., 1991, pp. 222-227).

Table 1:

Raw scores y_{ik} for a one-way *ANOVA* with repeated measures.

	Independent variable/Factor B				
Ss	**B_1**	**B_2**	**B_3**	**B_4**	**$P_i = \sum_k y_{ik}$**
1	16	28	30	34	108.00
2	10	18	14	22	64.00
3	21	20	24	30	95.00
4	20	34	38	36	128.00
5	14	28	26	30	98.00
	$\sum y_{i1} = 81$; $\sum y^2_{i1} = 1393$; $M_1 = 16.20$; $MSe_1 = 20.20$	$\sum y_{i2} = 128$; $\sum y^2_{i2} = 3448$; $M_2 = 25.60$; $MSe_2 = 42.80$	$\sum y_{i3} = 132$; $\sum y^2_{i3} = 3792$ $M_3 = 26.40$: $MSe_3 = 76.80$	$\sum y_{i4} = 152$; $\sum y^2_{i4} = 4736$ $M_4 = 30.40$; $MSe_4 = 28.80$	$\sum P_i = \sum y_{ik} = 493$ $\sum\sum y^2_{ik} = 13369$ $M = 24.65$ $MSe = 42.15$

Annotations. y_{ik}: raw score of subject in condition k; P_i: sum of the raw scores per row; MSe_k: variance within condition B_k: M_k: means in B_k.

Table 2:

Table of the *ANOVA* with repeated measures on factor B

Sources of variation	*SS*	*df*	*MS*	*F*	f^2_B
between Ss	540.80	4	135.20	----	----
within Ss	675.75	15	45.01	----	----
- factor B	542.15	3	180.7167	24.7585	----
-- BxSs/residual	133.60	12	11.1333	----	4.0580
-- wthin cells	674.40	16	42.15	----	.8039
Total	1216.55	19	----	----	----

Annotations. The critical value of F is $F_{crit(.05;3;12)} = 3.890$; thus the alternative hypothesis (H_1) is accepted, saying that not all means are equal. If one chooses the conservative F test according to Greenhouse and Geisser (1959) with $df_{num} = 1$ and $df_{den} = N - 1$, the critical value is $F_{crit(.05;1;4)} = 7.7086$; this value also leads to accepting the H_1.

The following relationships hold: $SST = SSbetwSs + SSwithinSs = 540.80 + 675.75 = 1216.55$ [$df_T = K*N - 1 = (N - 1) + K*(N - 1)$]; $SSBxSs = SSRes = SSwithinSs - SSB = 675.75 - 542.15 = 133.60$ [$df_{den} = (K - 1)*(N - 1) = N*(K - 1) - (K - 1)$; $SSwithin,B = SST - SSB = SSBxSs + SSbetwSs = 1216.55 - 542.15 = 674.40 = 133.60 + 540.80$ [$df_{den} = K*N - 1 - (K - 1) = (K - 1)*(N - 1) + (N - 1)$]. - The *ANOVA F* test then is performed by using:

$$f_{,rep\text{-}B} = \frac{SSB/(K-1)}{SSBxSs/[(K-1)*(N-1)]} = \frac{SSB}{SSBxSs/(N-1)} \tag{2}$$

Since this test does not rest on the usual assumption of homogeneity of the MSe_k, but on the weaker assumption of circularity (see Collier, Baker, Mandeville & Hayes, 1967, p. 342; Greenhouse & Geisser, 1959, and Winer et al., 1991, pp. 241-245), it seems advisable to determine the variance-covariance matrix; with repeated measures, the data are linearly correlated for each pair of treatment conditions, leading to $K*(K - 1)/2$ correlations between the levels of B, $r_{betw,rep\text{-}B}$. For each of these correlation a covariance $COV_{k,k'}$ can be computed by multiplying each correlation by the standard deviations $Se_k*Se_{k'}$ - these standard deviations are computed using N as the denominator -, whereas the covariances $COV_{k,k'}$ rest on $N - 1$ (Winer et al, 1991, p. 237).

$$Cov_{kk'} = \frac{\sum(y_{ik} - \bar{y}_k)*(y_{ik'} - \bar{y}_{k'})}{N-1} = r_{betw,rep\text{-}B,k,k'}*Se_k*Se_{k'}*N/(N-1) \tag{3}$$

The computations lead to the results given in the subsequent matrix:

$$\mathbf{S}_B = \begin{pmatrix} 20.20 & 11.60 & 27.40 & 17.90 \\ 11.60 & 42.80 & 53.20 & 30.20 \\ 24.60 & 53.20 & 76.80 & 45.80 \\ 17.90 & 30.20 & 45.80 & 28.80 \end{pmatrix} \tag{4}$$

From this matrix the mean covariance $\overline{COV}$ can be computed:

$$\overline{COV} = \sum COV_{kk'}/[K*(K-1)/2] \tag{5}$$

That is: $\overline{COV} = 186.10/6 = 31.0167$.

For the computation of the mean correlation between the pairwise treatment conditions, i.e. $r_{betw,rep\text{-}B}$, the following equation is used (Kirk, 1995, p. 274; Winer et al., 1991, p. 226, pp. 237-238, p. 264) (index rep-B: the Factor B is a repeated factor)

$$r_{betw,rep\text{-}B} = \overline{COV}/MSe = 31.0167/42.15 = .7359 \tag{6}$$

$$MSBxSs = MSe*(1 - r_{betw,rep\text{-}B}) = 42.15*(1 - .7359) = 11.1333 \tag{7}$$

This relationship also holds on the theoretical level:

$$\sigma^2_{BxSs} = \sigma^2_e * (1 - \rho_{betw,dep\text{-}B}) \tag{8}$$

Accordingly, an alternative computation of *MSBxSs* is possible:

$$MSBxSs = MSe - \overline{COV} = 42.15 - 31.0167 = 11.1333 \tag{9}$$

The variation between the Ss is related to the mean covariance between (Winer et al., 1991, p. 238):

$$MSbetwSs = MSw + (K - 1) * \overline{COV} = 42.15 + (3) * 31.0167 = 135.20 \tag{10}$$

Using the data given in Table 2, it becomes possible to compute the value of the measure $\hat{\varepsilon}$, which shows to which extent the assumption of circularity is violated and which ranges from $1/(K - 1)$ (maximum violation) through 1 (see Kirk, 1995, S. 280-281, or Winer et al., 1991, pp. 246-255). For the data chosen $\hat{\varepsilon} \approx 0{,}35$. Since the lower bound of $\hat{\varepsilon}$ is .3333 here, the circularity assumption does not hold. This fact, however, will not be regarded here, since under the favored interpretation of the assumptions connected with parametric tests as auxiliary hypotheses without empirical content the observed deviation from circularity is no matter of concern (see Westermann, 2000, pp. 337-338, for the details).

The effect size f^2_B was determined in two ways as shown in the subsequent formulas (Cohen, 1988, p. 281):

$$f^2_B = \frac{SSB}{SSe} \tag{11}$$

With this effect size, the fact that repeated measures are underlying, is disregarded of. This leads to better comparability with respect to the effect size based on a one-way *ANOVA* without repeated measures. - Alternatively the effect size can be determined by acknowledging the fact of repeated measures (rep-B):

$$f^2_{rep\text{-}B} = \frac{SSB}{SSBxSs} \tag{12}$$

Overall, there are very precise functional relationships with the univariate *ANOVA* for repeated measures, concerning the mean covariance or the mean correlation, respectively, and the sums of squares for error. The question then arises, whether relationships of the same kind also hold for the Friedman test for repeated measures.

The Friedman *ANOVA* for ranked data

The application of the Friedman *ANOVA* for ranks (Friedman, 1937) is often recommended for cases, where one or more of the parametric assumptions of the univariate *ANOVA* for repeated measures are violated. Under the interpretation of statistical tests as auxiliary hypotheses without empirical content (see above), this, however, should be no reason to renounce parametric tests such as the *F* test. Another justification for the use of parametric tests lies in the fact that nearly all of them can be interpreted as approximate randomization tests, for which the necessary (discrete) sampling distributions are derived by mere combinatorical manipulations (see Edgington, 1995, for the details). For these reasons, the application of the Friedman *ANOVA* should be restricted to cases where the data have ordinal scale level. In order to assure, that hypotheses on location are tested with the non-parametric *ANOVA*, it should be assumed, that the underlying distributions, resulting when the experiment is repeated under the same side conditions, are of (nearly) equal shape. The exact functions of these distributions, however, remain unspecified (cf. Westermann, 2000). This assumption must also be met, if rank hypotheses are subjected to the conventional parametric analyses and if an interpretation concerning equality or differences of location is intended. This assumption, then, is not restricted to non-parametric analyses, but also is necessary for parametric analyses. This weak assumption, by the way, is not empirically testable, as is always the case with auxiliary hypotheses without empirical content (see Westermann, 2000, pp. 337-338). Because of this weak assumption the Friedman and other rank tests continue to be non-parametric, but they are no longer distribution-free.

As compared to the parametric *ANOVA* the Friedman test has an asymptotic relative efficiency of $ARE_{FR} = (3/\pi)*[K/(K+1)] = .955*[K/(K+1)]$.

If one interprets the data considered hitherto as representing measures on ordinal level, the values y_{ik} must be rank-transformed into ranks R_{ik} before being able to perform a Friedman *ANOVA*. This is done by assigning ranks from 1 through *K* for each of the *N* Ss. Table 3 shows the results of this transformation.

Table 3:
Rank transformation of the raw scores in Table 1.

	Independent variable/Factor B				
Ss	B_1	B_2	B_3	B_4	$RP_i = \sum_k y_{ik}$
1	1	2	3	4	10
2	1	3	2	4	10
3	2	1	3	4	10
4	1	2	4	3	10
5	1	3	2	4	10
	$\sum R_{i1} = 6$; $\sum R^2_{i1} = 8$; $\overline{R}_1 = 1.20$; $MSe,R_1 = .20$	$\sum R_{i2} = 11$; $\sum R^2_{i2} = 27$; $\overline{R}_2 = 2.20$; $MSe,R_2 = .70$	$\sum R_{i3} = 14$; $\sum R^2_{i3} = 42$; $\overline{R}_3 = 2.80$; $MSe,R_3 = .70$	$\sum R_{i4} = 19$; $\sum R^2_{i4} = 73$; $\overline{R}_4 = 3.80$; $MSe,R_4 = .20$	$\sum R_{ik} = \sum RP_i = 50$ $\sum R^2_{ik} = 150$ $\overline{\overline{R}} = 2.50$ $MSe,R = .45$

In the next step, the variance-covariance matrix, based on rank correlations (Spearman's $r_{R,S}$ as a special case of the Pearson correlation $r_{X,Y}$) is constructed. The corresponding variance-covariance matrix $\boldsymbol{S_{R,B}}$ for the data in Table 3 takes on the following form:

$$\boldsymbol{S_{R,B}} = \begin{pmatrix} +.20 & -.30 & +.05 & +.05 \\ -.35 & +.70 & -.45 & +.05 \\ +.05 & -.45 & +.70 & -.30 \\ +.05 & +.05 & -.30 & +.20 \end{pmatrix} \tag{13}$$

The test statistic proposed by Friedman (1937) and mostly used is $\chi^2_{R,FR}$ (cf. Conover, 1999, p. 370; Marascuilo & McSweeney, 1977, p. 360). But a better approximation to the exact sampling distributions results, when using the parametric test statistic $F_{R,FR}$ (Conover, 1999, p. 370), which can be computed either by using the $\chi^2_{R,FR}$ statistic or directly in the same way as in the parametric case with the exception, that ranked data are used (see also Rasch & Kubinger, 2006, pp. 335-336). Generally, it is of no importance whether one uses $\chi^2_{R,FR}$ or $F_{R,FR}$ as the test statistic, since both usually result in the same decisions on the statistical hypotheses being tested (Zimmerman & Zumbo, 1993, p. 488); these facts are well-known to statisticians, but not necessarily to psychologists. Differences may arise, though, since the analysis using $\chi^2_{R,FR}$ incorporates no correction for ties, whereas the *ANOVA* via $F_{R,FR}$ does. (Of course, the $\chi^2_{R,FR}$ statistic can be modified according to the presence of tied ranks.)

The $F_{R,FR}$ test takes on the following form (Conover & Iman, 1981, p. 126):

$$F_{R,FR} = \frac{SSB,R/(K-1)}{SSBxSs,R/[(K-1)*(N-1)]} = \frac{SSB,R}{SSBxSs,R/(N-1)}. \tag{14}$$

Table 4:
ANOVA by ranks according to Friedman.

Sources of variation	*SSR*	*df*	*MSR*	$F_{R,FR}$	$f^2_{R,,B,FR}$
between Ss	0	4	----	----	----
within Ss	25.00	15	1.6667	----	----
- factor B	17.80	3	5.9333	9.8889	----
-- BxSs/Residual	7.20	12	.60	----	2.4722
-- within cells	7.20	16	.45	----	----
Total	25.00	19	----	----	----

Annotations. $F_{crit(.05;3;12)} = 3.890$; which means that once again the H_1 is accepted.

What about the several relationships between the mean correlation and the *SSs* shown above for the parametric case? To answer this question, let us consider the several sums of squares for ranks, *SSR,* whose computation is done as in the parametric case, but for the ranks R_{ik} instead of the original values y_{ik}. In addition, the *df*s are the same and will not be resumed here. - *SSbetwSs,R* = 0, since the row sums of the *K* individual ranks is equal for all Ss. Moreover, the sum of these individual rank sums is: $\sum R_{ik} = \sum RP_{ik} = N*K*(K + 1)/2$. - *SST,R* = *SSbetwSs,R* + *SSwithinSs,R* = 0 + 25.00 = 25.00; *SSBxSs,R* = *SSwithinSs,R* – *SSB,R* = 25.00 – 17.80 = 7.20; *SSe,R* = *SST,R* – *SSB,R* = *SSBxSs,R* + *SSbetwSs,R* =150.00 – 142.80 = 7.20. The equality *SSBxSs,R* = *SSe,R* always holds, since *SSbetwSs,R* = 0 in any case, and the equality *SSwithinSs,R* = *SST,R* is always valid for the same reason. - These relationships do not seem to be well-known - at least not in the psychological literature - as far as I know.

From the matrix in formula (13) the mean covariance can be computed as $\overline{COV}_R$ = –.90/6 = –.15 and the mean rank correlation as $r_{R,betw,rep--B}$ = –.15/.45 = –.3333. According to Marascuilo and McSweeney (1977, p. 359) the mean rank correlation with any Friedman test depends only on the number *K* of treatment conditions: $r_{R,betw,rep-B} = -1/(K - 1)$. In contrast, with the parametric one-way *ANOVA* with repeated measures the mean correlation $r_{\text{betw,rep-B}}$ theoretically (also most probably not empirically) can fall within the whole range of –1.00 through +1.00, so that for most cases, where $r_{\text{betw,rep-B}} > 0$, follows, that also *MSBxSs* < *MSe*. In the rare cases, where $r_{\text{bet,rep-B}} < 0$, *MSBxSs* will be greater than *MSe*. This relationship also holds for the Friedman test, but because of $r_{R,betw,rep-B} = -1/(K - 1)$ in a modified form: $MSBxSs,R = MSe,R*(1 + |r_{R,betw,rep-B}|)$ and therefore: *MSBxSs,R* > *MSe,R* throughout, that means opposite to the usual parametric case with $r_{\text{betw,rep-B}} > 0$. Thus, for the example results: *MSBxSs,R* = *MSe,R**(1.3333) = .45*1.3333 = .60. In spite of *MSBxSs,R* > *MS,e,R* the correct mean square for testing the hypotheses is *MSBxSs,R*, both in the parametric and in the rank case.

The effect sizes are computed according to the subsequent formula:

$$f^2_{R,B,FR} = \frac{SSB,R}{SSe,R} = \frac{SSB,R}{SSBxSs,R} \tag{15}$$

For the data in Table 4 the effect size takes on the value $f^2_{R,B,FR}$ = 2.4722.

In addition, various simulations have been performed with different values of *K* and of *N*, without leading to results differing in any way from the ones resting on only *N* = 5 and *K* = 4 as an illustrative example. The correlation always turned out to be $r_{R,betw,rep-B} = -1/(K - 1)$ and the relationships $MSBxSs,R = MSe,R*[1 + 1/(K - 1)]$ and *SSBxSs,R* = *SSe,R* also showed up in any simulation.

The well-known Wilcoxon test for two experimental conditions can be considered a special case of the Friedman test. If one assigns ranks R_{ik} to the original scores y_{ik}, the rank correlations between these ranks always take on the value –1, as follows from $r_{\text{R,betw,rep-B}} = -1/(2 - 1) = -1$, as a couple of simulations with different *N's* and different allocations of the ranks showed.

Parametric testing of further hypotheses about ranked data

As it appears, it is not well-known in psychology that the usual rank tests, such as the *U*, the Wilcoxon- and the *H* test can be replaced by their parametric counterparts including *z* tests for ranked data in the same way as for interval scaled data. In order to make this change easier, the most important parametric counterparts of the rank tests will be presented, since the relevant formulas are not mentioned in wide-spread statistics textbooks. As these formulas are scattered about statistical literature, no claim is made concerning originality.

1) *U* test: The $n_l + n_2 = N$ data are ranked from 1 through *N* and afterwards the ranked data are allocated to the two treatment conditions ($K = 2$).

$z_{R,U}$-Test: The rank sum, which is larger according to prediction (index "l"), is used as the test statistic, that is T_l; T_l must not be the larger rank sum empirically, in which case the *z* test takes on a negative sign leading to accepting the H_0, provided the Type-II error probability β is controlled. The expectations under the H_0 being tested are given by $E(T_l)$ and $E(T_s)$ (smaller rank sum according to prediction) (Conover, 1999, p. 281; Marascuilo & McSweeney, 1977, p. 296):

$$E(\mathrm{T_l}) = E(\Sigma \mathrm{R_{i,l}}) = n_l*(n_l + n_s + 1)/2 = n_l*(N + 1)/2, \tag{16}$$

$$E(T_s) = E(\Sigma R_{i,s}) = n_s*(n_l + n_s + 1)/2 = n_s*(N + 1)/2. \tag{17}$$

The standard deviation of T_l and of T_s, i.e. $s_{R,T,,U}$, is given by (Marascuilo & McSweeney, 1977, p. 296):

$$s_{R,T,U} = \sqrt{\frac{n_l * n_s * (n_l + n_s + 1)}{12}} = \sqrt{\frac{n_l * n_s * (N + 1)}{12}}\ . \tag{18}$$

The hypotheses tested take on the following form, given for the case most often encountered, that is for directional hypotheses:

$$H_{0,\mathrm{U}}\colon T_l \le E(T_l) \text{ vs. } H_{1,\mathrm{U}}\colon T_l > E(T_l) \text{ or } H_{0,\mathrm{U}}\colon E(\overline{R}_1) \le E(\overline{R}_2) \text{ vs. } H_{1,\mathrm{U}}\colon E(\overline{R}_1) > E(\overline{R}_2). \tag{19}$$

The test statistic $z_{R,T,U}$ is computed as follows (Marascuilo & McSweeney, 1977, p. 274, p. 296), with $N = n_l + n_s = n_1 + n_2$:

$$z_{R,T,U} = \frac{T_l - E(T_l)}{s_{R,T,U}} = \frac{T_l - n_l*(N + 1)/2}{\sqrt{n_l * n_s * (N + 1)/12}}\ . \tag{20}$$

The approximation to the unit normal distribution can be called satisfactory when $n_l > 10$ and $n_s > 10$ (Marascuilo & McSweeney, 1977, p. 274). - The formulas, though, do not take into account tied or equal ranks. But their numbers may be relatively large (up to 60%) without exerting a substantive influence on the test statistics - usually, tied ranks only lead to differences in the second or third decimal place of the test statistic. The presence of tied ranks always leads to lowering the standard error of the test statistic used, so that not apply-

ing the usually complicated corrections for ties only will lead to a slightly conservative decision, that is, the H_0 is longer retained than with the correction for tied ranks.

The *ES* appropriate for this test is:

$$\delta_{R,B,U,z} = \frac{E(T_l) - E(T_s)}{\sqrt{\frac{n_l * n_s * (N+1)}{12}}} = \frac{[n_l * (N+1)/2] - [n_s * (N+1)/2]}{\sqrt{\frac{n_l * n_s * (N+1)}{12}}}. \quad (21)$$

As is well-known, the *t* test can be applied here, too, which almost always leads to the same decisions as the *z* test:

$$t_{\mathrm{R,B,U}} = \frac{\overline{R}_l - \overline{R}_s}{s_{e,B,R} * \sqrt{\frac{2}{n}}} \quad (22)$$

The effect size is:

$$\delta_{R,B,U,t} = \frac{E(\overline{R}_l) - E(\overline{R}_s)}{\sigma_{e,B,R}} \quad (23)$$

2) Wilcoxon test for dependent samples: Here, the differences between the raw scores are computed for each subject, and these differences then are rank transformed from 1 through *N*. Afterwards, the *N* differences are enhanced by the sign of the original differences. Then, the sum of ranks with a positive sign is computed and the rank sum with a negative sign: $\sum R(+)_i$ und $\sum R(-)_i$. - If the prediction is directed, the following hypotheses are tested: $H_{0,\mathrm{W}}$: $E(\sum R_i) \leq 0$ *vs.* H_1: $E(\sum R_i) > 0$.

The distribution of the test statistic $z_{R,W}$ is approximately normal, if $N > 30$ (Bortz, Lienert & Boehnke, 2000, p. 262; Conover, 1999, p. 353, gives $N > 50$ for a satisfactory approximation).

$$z_{R,W} = \frac{\sum R_i}{\sqrt{N*(N+1)*(2*N+1)/6}} = \frac{T_l - N*(N+1)/4}{\sqrt{N*(N+1)*(2*N+1)/24}}. \quad (24)$$

(The first part of the foregoing formula follows Conover, 1999, p. 353, the second part can be found in a slightly modified form in Bortz et al., 2000, p. 262).

2b) $t_{R,W}$ test: Using the empirical values for the two rank sums, $R(+)_{rep} = \sum R(+)_i$ and $R(-)_{rep} = \sum |R(-)_i|$ and for $\sum R_i$, a *t* value can be computed, namely $t_{R,W}$ (after Conover & Iman, 1981, p. 126). $d_{i,\mathrm{R}}$ stands for the individual rank differences, $s_{\mathrm{D,R}}$ for the standard error of the *N* signed ranks.

$$t_{R,W} = \frac{\sum R_i}{\sqrt{\frac{N}{N-1} * \sum R_i^2 - \frac{1}{N-1} * (\sum R_i)^2}} = \frac{[\overline{R}(+) - \overline{R}(-)] * \sqrt{N*(N-1)}}{\sqrt{\sum d_{i,R}^2 - N*[\overline{R}(+) - \overline{R}(-)]^2}} = \frac{\overline{R}_i}{s_{D,R} * \sqrt{\frac{1}{N}}} . \quad (25)$$

This $t_{R,W}$ value again takes tied ranks into account, the $z_{R,W}$ expressions in formula (22) do not. - The effect size for the tests can be defined as, with $E[T(+)] = N*(N + 1)/24$:

$$\delta_{R,B,W,z} = \frac{T(+) - \{E[T(+)]\}}{\sqrt{N*(N+1)*(2*N+1)/24}} . \quad (26)$$

Or:

$$\delta_{R,B,W,t} = \frac{E[\overline{R}(+)] - E[\overline{R}(-)]}{\sigma_{e,B,R}} = \frac{E[\overline{R}_i]}{\sigma_{e,B,R}} . \quad (27)$$

3) One-way *H* test with $K > 2$, independent samples. The raw scores y_{ik} of the N Ss are transformed into the ranks from 1 through N over all Ss, and afterwards the ranks are re-assigned to the K experimental conditions. - The F distributions lead to a slightly closer approximation to the exact distributions than the usually applied $\chi^2_{R,H}$ distributions (Conover, 1999, p. 297, p. 418; cf. Conover & Iman, 1981, p. 126). The hypotheses tested concern the equality of the mean ranks (H_0) vs. at least two mean ranks are not equal (H_1). - For testing these hypotheses using the $F_{R,H}$ test Conover (1999, p. 257) gives the subsequent formula (cf. Conover & Iman, 1981, p. 125; Silverstein, 1974):

$$F_{R,H} = \frac{n*\sum(\overline{R}_k - \overline{R})^2/(K-1)}{SSe,R/[K*(N-1)]} = \frac{SSB,R/(K-1)}{SSe,R/(N-K)} = \frac{MSB,R}{MSe,R} . \quad (28)$$

Thus, the $F_{R,H}$ value is computed in complete analogy of the parametric F value in the one-way case, but using the ranks: $SSB,R = \Sigma_k(\Sigma_i R_{ik})^2/n - (\Sigma_k\Sigma_i R_{ik})^2/(K*n)$ with $df_{num} = (K - 1)$; $SST,R = \Sigma_k\Sigma_i R^2{}_{ik} - (\Sigma_k\Sigma_i R_{ik})^2/(K*n)$ with $df_T = N - 1$ and $SSe,R = \Sigma_k\Sigma_i R^2{}_{ik} - \Sigma_k(\Sigma_i R_{ik})^2/n$ with $df_{den} = K*(n - 1)$. *SSB,R* denotes the sums of squares between the experimental conditions, *SSe,R* stands for the sums of squares of errors and *SST,R* for the total sums of squares.

The effect size is defined as:

$$f^2_{R,B,H} = \frac{SSB,R}{SSe,R} \quad (29)$$

Parametric testing of hypotheses about contrasts involving ranked data

Using *ANOVA*-like tests always means that only bidirectional statistical hypotheses can be tested, and these hypotheses are very exact, if they are null hypotheses, and very inexact, if they are alternative hypotheses. Considering the fact that most of the psychological hy-

potheses are directional it would be advantageous to have the possibility of testing directional statistical hypotheses as well. The method of choice, which enables testing of directional (and bidirectional) hypotheses is the method of planned contrasts. Each contrast is associated with one *df* in its numerator. If one considers two contrasts simultaneously it does not matter whether these contrasts are orthogonal (linearly independent) to each other or not (non-orthogonal contrasts). If one uses this method, no global test such as *ANOVA* is performed first. The cumulation of error probabilities, occurring when two or more contrasts have to be tested, is adjusted by the versatile Boole-Bonferroni-Methode. - For each contrast, the assignment of ranks has to be done separately, i.e. from 1 through $n_k + n_{k'}$ or from 1 through N with $K' = 2$.

Contrasts concerning mean ranks are of the following form, in which the $c_{k,t}$ are the contrast coefficients per treatment condition and per contrast $D_{R,t}$, which have to sum up to 0 (Marascuilo & McSweeney, 1977, p. 306):

$$D_{R,t} = c_{1,t} * \overline{R}_1 + c_{2,t} * \overline{R}_2 + \ldots + c_{K,t} * \overline{R}_K. \tag{30}$$

In the model of Kruskal and Wallis (*H* test), mainly the $z_{R,H}$ and the $t_{R,H}$ test are appropriate for testing hypotheses about orthogonal and non-orthogonal contrasts. - Equal samples sizes are assumed throughout for simplicity's sake.

$$z_{R,H,t} = \frac{D_{R,t}}{\sqrt{\frac{N*(N+1)}{12} * \frac{\sum c_{k,t}^2}{n}}} = \frac{\sum c_{k,t} * \overline{R}_k}{\sqrt{\frac{N*(N+1)}{12} * \frac{\sum c_{k,t}^2}{n}}} \tag{31}$$

$$t_{R,H,t} = \frac{z_{R,H,t}}{\sqrt{\frac{K*n-1}{K*n-2} - \frac{1}{K*n-2} * z_{R,H,t}^2}} = \frac{\sum c_{k,t} * \overline{R}_k}{\sqrt{s_{e,B,R} * \frac{\sum c_{k,t}^2}{n}}} \tag{32}$$

Formula (31) can be found in Marascuilo and McSweeney (1977, p. 306), the first part of formula (32) in Conover and Iman (1981, p. 125). - It is also possible to test the hypotheses of interest by pairwise *U* tests as given in formula (20). Choosing this way, it must be taken into account, that the ranks have to be allocated separately for each test to be performed.

The effect size is given by:

$$\delta_{R,B,H,z} = \frac{\sum c_{k,t} * E(\overline{R}_k)}{\sqrt{\frac{N*(N+1)}{12}}} \tag{33}$$

$$\delta_{R,B,H,t} = \frac{\sum c_{k,t} * E(\overline{R}_k)}{s_{e,B,R}} \tag{34}$$

The difference between the two standard deviations is the reason, that usually the z values are a little smaller than the t values - without consequences regarding the decisions.

On the basis of the Friedman model (repeated measures) the same tests can be applied, but using different standard deviations:

$$z_{\mathrm{R,FR,t}} = \frac{D_{R,t}}{\sqrt{\frac{K*(K+1)}{12} * \frac{\sum c_{k,t}^2}{N}}} = \frac{\sum c_{k,t} * \overline{R}_k}{\sqrt{\frac{K*(K+1)}{12} * \frac{\sum c_{k,t}^2}{N}}} \tag{35}$$

The $t_{\mathrm{R,FR,t}}$ value can be computed using the first part of formula (32), bearing in mind, however, that the number of Ss, N, has to be inserted (instead of $K*n$) (Conover & Iman, 1981, p. 126). - It is also possible to employ pairwise Wilcoxon tests as in formula (22) or (23), remembering that the two rank sums or the one rank sum have to be computed separately for each test.

The effect size is defined as:

$$\delta_{\mathrm{R,B,FR,z}} = \frac{\sum c_{k,t} * E(\overline{R}_k)}{\sqrt{\frac{K*(K+1)}{12}}} \tag{36}$$

$$\delta_{\mathrm{R,B,FR,t}} = \frac{\sum c_{k,t} * E(\overline{R}_k)}{s_{e,B,R}} \tag{37}$$

The z_R-, t_R- and F_R tests should be considered approximate tests, that is, their sampling distributions are - in the same way as the generally used χ^2_{R} distributions - only approximations for the exact sampling distributions (Conover & Iman, 1982, p. 126), but the goodness of approximation can be considered as satisfactory or better in general, if the total sample size is not too small, that is, if $N \geq 30$. This sample size will easily be approached or even exceeded, if the study is preceded by a power analysis, as it should be (Cohen, 1988; Rasch, 2003).

Whenever one of the tests proposed preceedingly, it must be remembered, that the allocation of ranks is different for non-repeated and repeated measurements.

References

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler* (7th ed.). Berlin: Springer. (Statistics for the human and the social sciences).

Bortz, J., Lienert, G.A. & Boehnke, K. (2000). *Verteilungsfreie Methoden der Biostatistik* (2. Aufl.). Berlin: Springer. (Distribution-free methods in bio-statistics).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collier, R.O. jr., Baker, F.B., Mandeville, K. & Hayes, T. (1967). Estimation of test size for several test procedures based on variance ratios in the repeated measures design. *Psychometrika, 32,* 339-353.

Conover, W.J. (1999). *Practical nonparametric statistics* (3rd ed.). New York: Wiley.

Conover, W.J. & Iman, R.L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124-129.

Erdfelder, E., Buchner, A., Faul, F. & Brandt, M. (2004). GPOWER: Teststärkeanalysen leicht gemacht. In E. Erdfelder & J. Funke (Hrsg.), *Allgemeine Psychologie und deduktivistische Methodologie* (pp. 148-166). Göttingen: Vandenhoeck & Ruprecht. (GPOWER: Power analysis made easy)

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.

Girden, E.R. (1992). ANOVA: *Repeated measures.* Newbury Park, CA: Sage.

Greenhouse, S.W. & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 23,* 95-112.

Hays, W.L. (1988). *Statistics* (4th ed.). London, UK/Orlando, FL: Holt, Rinehart & Winston.

Hager, W. (2004). *Testplanung zur statistischen Prüfung psychologischer Hypothesen*. Göttingen: Hogrefe. (Power analysis for examining psychological hypotheses)

Keppel, G. & Wickens, T.D. (2004). *Design and analysis. A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Keselman, H.J., Algina, J., Boik, R.J. & Wilcox, R.R. (1999). New approaches to the analysis of repeated measurements. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 251-268). Stamford, CT: Jai Press.

Kirk, R.E. (1995). *Experimental design* (3rd ed.). Belmont, CA: Wadsworth.

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

O'Brien, R.G. & Kaiser, M.K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin, 97,* 316-33.

Rasch, D. (2003). Determining the optimal sample size of experiments and surveys in empirical research. *Psychology Science, 45,* 3-48.

Rasch, D. & Kubinger, K.D. (2006). *Statistik für das Psychologiestudium.* Heidelberg: Spektrum - Elsevier. (Statistics for psychologists)

Rouanet, H. & Lépine, D. (1970). Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology, 23,* 147-163.

Silverstein, A.B. (1974). Relations between analysis of variance and its nonparametric analogues. *Psychological Reports, 34,* 331-333.

Tanguma, J. (1999). Analyzing repeated measures designs using univariate and multivariate methods. A primer. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 233-250). Stanford, CT: JAI Press.

Westermann, R. (2000). *Wissenschaftstheorie und Experimentalmethodik.* Göttingen: Hogrefe. (Philosophy of science and experimental methods)

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

Zimmerman, D.W. & Zumbo, B.D. (1993). The relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 481-517). Hillsdale, NJ: Erlbaum.