Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 6

5-2016

New Nonparametric Rank Tests for Interactions in Factorial Designs with Repeated Measures

Jos Feys
KU Leuven, Belgium, jos.feys@faber.kuleuven.be

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 78-99.

ISSN 1538 − 9472

New Nonparametric Rank Tests for Interactions in Factorial Designs with Repeated Measures

Jos Feys
University of Leuven
Leuven, Belgium

New rank tests for interactions in factorial designs are presented and applied to some common factorial designs with repeated measures. The resulting *p*-values of these tests are compared, along with those obtained by parametric and randomization tests.

Keywords: Nonparametric interaction tests, aligned rank test, new rank-based methods, Friedman ranks, randomization, gain scores

Introduction

Techniques have been proposed for the nonparametric analysis of interactions in factorial designs. They rank the observations and then perform parametric tests on ranks. The aligned rank tests belong to one class of these. Aligning implies that some estimate of a location (e.g., for the effect on a certain level of a given factor), such as the mean or median of the observation, is subtracted from each observation. These data, thus aligned according to the desired main and/or interaction effects, are then ranked and parametric tests are performed on these aligned ranks. A second class of such tests, named the new rank based methods, first rank the (not aligned) observations, and then the relative treatment effects are defined in reference to the distribution of the variables measured and estimated through elaborate calculations on these ranks.

The alignment methodology was introduced by Hodges and Lehmann (1962) and extended to two-way layouts by Sen (1968). McSweeney (1967) developed a test (M test) for interaction using the aligned ranks in the two-way layout. The aligned rank tests were publicized by: Hettmansperger (1984), Puri and Sen (1985), Sawilowsky (1990), and Higgins and Tashtoush (1994). More recently, Beasley

Jos Feys is an Honorary Senior Research Fellow in the department of Kinesiology. Email at: jos.feys@faber.kuleuven.be.

and Zumbo (2003; 2009) added the aligned Friedman rank test for interactions in split-plot or repeated measures designs. Reviews of the aligned rank tests have been provided by Sawilowsky (1990), Higgins and Tashtoush (1994), Toothaker and Newman (1994), Kelley and Sawilowsky (1997), Richter and Payton (1999), Peterson (2002), and Rodriguez, Álvarez, and Remirez (2009). Salazar-Álvarez, Tercero-Gómez, Temblador-Pérez, and Conover (2014) recently reviewed nonparametric test for interactions. They overlooked the – in my opinion – important contributions by Beasley and Zumbo (2003; 2009) on designs with repeated measures. The general conclusion of these reviews was that the aligned rank tests are valid nonparametric alternatives for the parametric tests for the interaction, especially when sample sizes are small (Sawilowsky, 1990) or the departure from normality of the distribution of the observations is extreme (e.g., heavy tailed; Kelley & Sawilowsky, 1997).

Pioneers on the new rank-based methods were Akritas (1990), Akritas and Arnold (1994), Akritas, Arnold, and Brunner (1997), and Akritas and Brunner (1997). Brunner and Puri (2001) reviewed these methods.

New Rank Tests

The Aligned Rank Transform Formulae

The rank transform procedure, as proposed by Conover and Iman (1976), in essence replaces original observations with their ranks and then computes parametric tests on these ranks. Higgins and Tashtoush (1994) showed that this method is flawed when applied to tests for interaction in factorial designs. The underlying reason is that, when nonlinear transformations (such as the rank transform) are made on a set of data, interaction structures that exist may or may not exist in the transformed data, and vice versa. Therefore, the rank transform procedure cannot be applied to test interactions. They advocated the use of the alignment of the data before ranking, the aligned rank transform, thus combining the notion of alignment of data and the rank transform. This procedure removes the effect of nuisance – as they called it – parameters when testing for effects of parameters of interest.

Two-way between designs

For two-way designs (A × B, 2 between factors), the mathematical linear model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \ , \qquad (1)$$

where $i = 1,\ldots, r$, $j = 1,\ldots, c$, $k = 1,\ldots, n$, and the ϵ_{ijk}s are independent and identically distributed (i.i.d.) random variables with mean 0 and common standard deviation σ is (Higgins & Tashtoush, 1994, p. 203). The α_is and β_js represent the row (A) and column (B) effect, respectively, and $(\alpha\beta)_{ij}$ is the interaction. The adjustment factors proposed are based upon the usual estimates of the parameters. These estimates, with their respective means, are

$$\hat{\mu} = \bar{Y}, \quad \hat{\alpha}_i = \bar{Y}_i - \bar{Y}, \quad \hat{\beta}_j = \bar{Y}_j - \bar{Y}, \quad \alpha\beta_{ij} = \bar{Y}_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y} \ .$$

The aligned data for testing A × B interactions have the form

$$AB_{ijk} = Y_{ijk} - \bar{Y}_i - \bar{Y}_j + \bar{Y} \ . \tag{2}$$

The row (A) and column (B) effects (means) are subtracted from the individual observations and the overall mean is added (to compensate for the 2 subtractions). To apply the aligned rank transform to test for interactions, the AB_{ijk}s are ranked, and the ranked data are analyzed with a full model parametric procedure which includes all effects (i.e. effect A, effect B, and effect A × B), for example an ANOVA. The authors compared the use of the sample means, as estimates of the location of the effects, to the use the medians and trimmed means. They advocated the use of the means because it is easy to calculate and is equally powerful as its alternatives. Peterson (2002) compared six alternatives for the estimation of location and concluded that the samples means and medians are the best estimators.

Repeated measures

For split-plot or repeated measures designs, the aligned data for testing for interactions is given by Higgins and Tashtoush (1994, p. 208):

$$GT_{ijk} = Y_{ijk} - \bar{Y}_{i.k} - \bar{Y}_{.j.} + \bar{Y} \ , \tag{3}$$

where $i = 1,\ldots, r$ (between Groups), $j = 1,\ldots, c$ (within Time), $k = 1,\ldots, n$ (observations/subjects/blocks), and GT_{ijk} represents the Group × Time interaction. The mean $\bar{Y}_{i.k}$ of observations of the k^{th} row (observation nested within the i^{th} Group) and the mean $\bar{Y}_{.j.}$ of observations of the j^{th} column (Time) are subtracted from the individual observation Y_{ijk}, and the overall mean is added to the difference.

Three-way designs

For three-way interactions in experiments involving three factors, the alignment is given schematically as

$$\begin{aligned} ABC_{ijkl} = Y_{ijkl} \\ -(\text{sum of 2-way means involving } i, j, k) \\ +(\text{sum of 1-way means involving } i, j, k) \\ -\text{overall mean} \end{aligned} \tag{4}$$

(Higgins & Tashtoush, 1994, p. 209). These formulae can be applied with a little programming skill in a data manipulation step of any kind of statistical package. Once this is done, ranking and testing for interactions should be a routine job for everyone. SAS statistical software (SAS Institute Inc., 2008) is used here for implementing these formulae in all subsequent examples. Leys and Schumann (2010) used (2) for the A × B interaction above on ordered Likert-type scale data.

Randomization or Permutation

Randomization or resampling tests have become popular since they were included in some procedures or functions within common statistical software. Cassell (2002) developed randomization test wrapper macros for the SAS statistical software procedures. With these macros wrapped around common parametric procedures, these can be replicated for a large number of random data permutations. Then the number of times the obtained *p*-values are equal to or smaller than the parametric test *p*-value are counted. The result of this count divided by the number of random permutations (usually 10000) is the randomization test *p*-value. In the examples below, I used this randomization test wrapper for comparison with the parametric and nonparametric procedures.

New Rank-Based Methods, Brunner SAS and R Macros, Gao and Alvo

Brunner, Domhof, and Langer (2002) developed macros with SAS and R (R Core Team, 2012) for the applications of the new rank-based methods. Noguchi, Gel, Brunner, and Konietschke (2012) published the nparLD R package which provides researchers an easy and user-friendly access to these methods. Along with a macro for calculating confidence intervals, the macros offered range from a within

(repeated measures or longitudinal) one-way design and a between one-way design, up to three-way designs with one or two within factors. The original data are always first ranked. All macros accept data only in long (or multiple record) format. Therefore, if the data are in short (or multiple variable) format, with one record per observation (per row) and the repeated measures in the columns, they are stretched out to separate records (rows) for each repeated measure and only one response (dependent variable), and a variable is added to identify the repeated measure. In each example below, an appropriate macro was used.

Shah and Madden (2004), who illustrated the usefulness of several of the SAS macros in plant disease epidemiology, also presented the theory in a succinct but clear way. In this methodology, based upon the seminal paper by Aktritas and Arnold (1994), the hypotheses are not formulated in terms of expectations of treatment effects (e.g., difference between means), but rather in reference to the distribution of variables measured in the experiment. Marginal or treatment (the authors use both terms interchangeably) effects are quantified by the appropriate estimates on mean ranks after extensive calculations. Noguchi et al. (2012) noted "the rank-based methodology is not restricted to data on a continuous scale and enables to analyze ordered categorical, dichotomous, and heavily skewed data," (p. 2). They added that the methods are robust to outliers and are appropriate for small sample sizes. Akritas et al. (1997) specified that these methods are suitable for unbalanced designs. Kaptein, Nass, and Markopoulos (2010) demonstrated the power of this approach for the analysis of Likert-type rating scales.

In an electronic supplement, Shah and Madden (2004) explained how to apply the new rank-based methods using SAS for mixed (i.e., fixed and random effects) models. The data are not aligned; the MIXED procedure is applied directly on ranked data. It allows different covariance structures for all factor level combinations by specifying the type = UN (unstructured variance-covariance) option. The use of the so-called minimum variance quadratic unbiased estimation method is recommended instead of the default restricted maximum likelihood. We used this procedure in Example 1.

Gao and Alvo (2005a; 2005b) added a new rank statistic to test for interactions in two-way layouts by comparing the sum of row ranks with the sum of column ranks. It is unclear how to apply this statistic to two-way layouts with repeated measures.

Aligned Friedman Ranks, Beasley and Zumbo SAS/IML

Beasley and Zumbo (2003; 2009) investigated the usefulness of aligned rank tests for interactions in split-plot or repeated measures designs. The alignment procedure they advocated for such designs is the Higgins and Tashtoush (1994; hereafter called H&T) method whereby the alignment takes into account that the observations (subjects/blocks) are nested within the between factor Group. The aligned data for testing the interaction are obtained by applying the above mentioned (3) for split-plot or repeated measures designs. The authors methodically compared the regular rank test, across observations, on the aligned data with the aligned Koch ranks based on ranking the K^2 pairwise differences among the K levels of the repeated measures, regardless of Group membership, and aligned Friedman ranks, based on ranking of the data from 1 to K across the levels of the repeated measures factor within each observation. In their 2009 article, they included the SAS/IML (Interactive Matrix Language) syntax code to perform the aligned regular rank, the aligned Friedman rank, and the aligned Koch rank tests. The data are first aligned and then ranked according to each of these methods. After these two steps, parametric procedures with all effects included in the model (i.e., a full-factorial model repeated measures ANOVA) are applied to the three versions of aligned ranks. In most examples below, all with repeated measures, this IML script was applied to obtain the desired aligned ranks tests along with calculations based on the H&T formulae as checks.

ARTool

Wobbrock, Findlater, and Higgins (2011) proposed the ARTool, a tool for calculating the aligned rank transform. The ARTool generalizes aligned rank transform for nonparametric factorial data analysis to N factors and can therefore be used for higher-order interactions. The alignment of the data (in long format: with the observation identifier in the first column and the response in the last column, and all intervening factors in-between) is made in five steps. First (step 1), the residuals (observations – cell mean) are computed, and then (step 2) the estimated effects for all main and interaction effects are computed. For example, for a two-way design, the estimated effect for an A × B interaction response is achieved by

$$Y_{ij} = \overline{A_i B_j} - \bar{A}_i - \bar{B}_j + \mu \ . \tag{5}$$

In a third step, the aligned response is computed:

$$Y' = \text{residual+estimated effect}\left(\text{result from step 1+step 2}\right).$$

These aligned data (Y') are ranked to create Y'', the aligned ranks. Finally, a full-factorial ANOVA is performed on Y''. Wobbrock et al. (2011) noted that "alignment works best for completely randomized designs; it also works for other designs, but effects may not be entirely stripped out" (p. 146). This application was used in all examples below.

Examples

Example 1: A Pretest-Posttest Design

The first example was taken from Bonate (2000, p. 106). The data in the table resulted from a design with two Groups (control and treatment; $n = 10$ and $n = 9$, respectively) and two repeated (pre- and post-) measures, denoted as Time factor. In the treatment Group, there was an outlier on the post-measures: a value of 19 between values quite larger than 60 in the whole table. With Dixon's test for a single outlier, this very low value was flagged with a test probability (one sided) of $p < 0.001$.

According to Bonate (2000), pretest-posttest data can be analyzed in several ways: ANOVA on final scores alone, on difference scores, on percentages change scores, by means of an analysis of covariance (ANCOVA) with the pretest as covariate for the predicting Group factor and the posttest as outcome variable, blocking by initial scores (stratification), and as repeated measures. In this example, focus on difference or gain scores, repeated measures, and ANCOVA. The resulting p-values for the interaction with the different ways of analysis are reported in Table 1.

Gain Scores

Gain scores are obtained by computing post – pre difference scores. Differences in gain scores (i.e., in difference scores) between Groups, if any, should reveal the interaction Group × Time. If there is more gain in one Group than in the other, this would correspond to the interaction.

Parametric tests. The distribution of the gain scores over the 2 Groups was not normal, according to the Shapiro-Wilk test, $p < 0.0122$. The means and standard deviations (between brackets) of the gain scores, in the control Group and in the

treatment Group respectively, were: -1.60 (7.31) vs. 15.56 (29.77). So, the difference in gain scores was: 17.16 (21.10) in favor of the treatment Group. This seemed to indicate that there was an interaction between the Time (within) and the Group (between) factor. The variances of the gain scores were not equal between Groups, $p = 0.0003$. The t-test on the difference in gain scores between Groups (i.e., Group × Time interaction), for equal variances (pooled), was not significant, $p = 0.0947$; for unequal variances, this test (Satterthwaite corrected df's) was also not significant, $p = 0.1270$.

Table 1. Resulting *p*-values with different way of analysis for the Group × Time interaction in Example 1.

Gain scores	***p*-values**	**Rep. measures**	***p*-values**	**(R)ANVOVA**	***p*-values**
	Parametric/randomization		**Parametric/randomization**		**Parametric/randomization**
t-test		F-test		ANCOVA posttest,	
pooled	0.0947	short format	0.0947	pretest as covariate	
Satterthwaite	0.127	long format	0.0947	F-test	0.0576
Wrapper t-test		Wrapper F-test	0.0863	Wrapper F-test	0.0474
pooled	0.0862				
Satterthwaite	0.1202				
Permutation option,					
NPAR1WAY	0.0892				
	Nonparametric		**Nonparametric**		**Nonparametric**
on ranked gain scores		on ranked pre-post measures		on residual ranks	
t-test, pooled	0.0015	F-test	0.0615	Quade's	
				RANCOVA	0.0048
Wrapper t-test	0.0022	new rank-based methods		Mant.-Heanszel	0.0088
Permutation option,		(Shah & Madden)		Wilcox. exact	0.0057
NPAR1WAY	0.0025	F-appr. large N.	0.0484	Permutation options,	
		F-appr. small N	0.067	NPAR1WAY	0.0061
on gain scores, ranked		on pre-post measures,		ranked posttest,	
in procedure		ranked in procedure		ranked pretest as covariate	
Mant.-Heanszel	0.0014	Brunner macro F1_LD_F1		RANCOVA	
Wilcox. exact	0.0027	F-appr. large N	0.0484	F-test	0.0031
Brunner OWL		F-appr. small N	0.067		
Exact test	0.0027	aligned ranks			
F-approximation	0.0017	regular	0.0014		
		Friedman	0.0011		
		Koch	0.0015		
		ARTool	< 0.0001		

The Cassell randomization wrapper around this t-test procedure, showed almost the same *p*-values as these two values for the equal or unequal variances, respectively: $p = 0.0862$ and $p = 0.1202$.The permutation test option in the SAS nonparametric one-way procedure (NPAR1WAY) on the gain score also revealed a non-significant *p* value: 0.0892.

Nonparametric tests. The gain scores were ranked over the 2 Groups, thus ignoring the Group factor, and several parametric tests were applied on these ranks. The Shapiro-Wilk test for normality of distribution on the ranks was not significant, $p = 0.5236$. The mean ranks (SD's) were, for the control Group: 6.50 (3.00) and, for the treatment Group: 13.89 (5.34). The variances between Groups were equal, p = .1051. The (pooled) t-test for the difference of 7.39 (4.27) between Groups (i.e., Group × Time interaction) was significant, p = .0015. The randomization wrapper around this t-test upon the ranked gain scores resulted in a quasi-equally significant value, p = .0022. The permutation test option in the SAS nonparametric one-way procedure also yielded a significant *p*-value for the difference in ranked gain scores, p = .0025.

Typical nonparametric tests can rank the scores within the procedure itself. One of these is the Mantel-Heanszel statistic on the differences between the two Groups in gain scores, ranked within the SAS FREQ procedure with the scores = ranks option. This was significant, $p = 0.0014$. The exact Wilcoxon two-sample test on gain scores (nonparametric one-way procedure) was also significant, $p = 0.0027$. The Brunner macro One-Way Layout (OWL) for the Group factor on the gain scores, ranked within this macro, resulted in similar *p*-values; the exact $p = 0.0027$, and the value for the F-approximation was $p = .0017$.

Repeated Measures

Parametric tests. The means and standard deviations (between brackets) of the pretest measures for the control Group and treatment Group, respectively, were about equal: 75.00 (4.50) vs. 78.78 (5.89). For the posttest measures, these values were quite different: 73.40 (7.09) vs. 94.33 (28.74), respectively. Again, as for the differences in gain scores, this seemed to indicate an interaction. For a Time (pre-post) × Groups design, a repeated measure ANOVA F-test for the interaction corresponds to t-test on difference or gain scores, as pointed out by Dimitrov and Rumrill (2003). The *p*-value for this F-test on the data in short format was the exactly the same as the *p*-value for the pooled (i.e. df's for equal variances) t-test, namely $p = 0.0947$. Of course, this *p* value for the F-test was also the same when

calculated on the data in long format. The randomization test with the Cassell wrapper around this repeated measures ANOVA procedure was also not significant for this interaction, F-test, $p = 0.0863$, which was about the same value as with the randomization wrapper pooled t-test on gain scores in Table 1.

Nonparametric tests. The pretest-posttest measures were ranked on the data in long format, thus over Groups and Times. The mean ranks and their standard deviations (between brackets) of the pretest measures for the control Group and treatment Group, respectively, were: 14.85 (7.93) vs. 20.11 (7.19). For the posttest measures, respectively: 13.45 (9.48) vs. 30.78 (11.40). The Shapiro-Wilk test for normality of distributions was not significant, $p = 0.1223$. The F-test for the Group × Time interaction on these ranks was not significant, $p = 0.0615$. Applying the parametric SAS procedure MIXED on these ranks to obtain the estimated treatment effect of the new rank-based methods, as described in the electronic supplement to Shah and Madden (2004), resulted in a significant p value for the F-approximation for large samples, $p = 0.0484$, but not for small samples, $p = 0.0670$.

Here too, some nonparametric procedures include ranking. The Brunner macro suited for this design (F1_LD_F1) was applied to the repeated measures data in long format, which yielded exactly the same p-values as obtained with the Shah and Madden (2004) parametric SAS procedure for large and small samples, respectively: 0.0485 and 0.0670.

The Beasley and Zumbo (2009) SAS/IML syntax code includes the calculations based on the H&T formula (3) aimed at the alignment for the interaction with split plot or repeated measures in short format. Applying this script (paralleled with the author's own calculations) yielded a significant p value for the regular ranking across observations of the aligned data: 0.0014. For the Friedman ranking (within observations), as well as for the Koch rankings, the resulting p-values were also quite significant, respectively: 0.0011 and 0.0015.

The ARTool can only be applied on data in long format. The ARTool procedure was applied to the data in this format and a significant p-value for the interaction was obtained, $p < 0.0001$. The Pearson correlation between the data aligned with the H&T formula and the ARTool method was: $r\ (N = 38) = 0.7386$, which indicated that the alignments were not the same.

Analysis of Covariance (ANCOVA)

The interaction Group × Time can also be tested with an ANCOVA test by specifying the pretest measure to be the covariate and the Group variable to be the

predictor in the regression with the posttest measure as dependent variable (criterion).

Parametric tests. This ANCOVA of the posttest measures on the pretest measures and the Group variable did not reveal significant *p* values, neither for the covariate, $p = 0.9558$, nor for the Group variable, $p = 0.0576$. The absence of a significant Group effect here indicated that the Group × Time interaction was not significant. The wrapper F-value for this test was just significant, $p = 0.0474$.

Nonparametric tests. Residual rank scores were calculated. First, the pretest and posttest measures were ranked separately, ignoring Groups. Then a regression was run of the posttest ranks on the pretest ranks, again ignoring Groups, and residuals (residual rank scores) were saved. An ANOVA on such residuals corresponds to the rank analysis of covariance (RANCOVA) proposed by Quade (1967). The *p* value for the Group effect on the residuals (i.e., Group × Time interaction) was significant, $p = 0.0048$. The Mantel-Heanszel statistic on the differences in mean residual rank scores between the two Groups was also significant, $p = 0.0088$, as was the exact Wilcoxon two-sample test on these residual ranks, $p = 0.0057$. The permutation test option within the NPAR1WAY procedure applied to these residual ranks was also significant, $p = 0.0061$. A RANCOVA of the ranked posttest on the ranked pretest as covariate for the Group factor also returned a significant F-value, $p = 0.0031$.

Example 2: A 3 × 5 Between × Within Design with Data on an Ordinal Scale

The second example, from Shah and Madden (2004), is the powdery mildew of wheat data. This corresponds to a 3 between × 5 within factors design with three wheat cultivars (Groups) and five severity of decay (mildew) assessments (Times) and with four replications ($n = 4$ observations) of each cultivar ($N = 12$ in total). The assessments made were on a 0-to-10 ordinal (ordered categories) scale for the severity of decay. The resulting *p*-values for the Group × Time interaction obtained by Shah and Madden (2004) with the new rank-based methods were compared to those obtained by parametric tests, randomization tests, and two nonparametric procedures. The Beasley and Zumbo (2009) SAS/IML code was run on the data to obtain the aligned regular rank test, the aligned Friedman rank test, and the aligned Koch rank test. The ARTool was also applied on these data. The resulting *p* values are displayed in the first part of Table 2.

Table 2. Resulting *p*-values with different methods of analysis for the interaction in Examples 2, 3, and 4

3 B × 5 W	*p*-values	2 W × 3 W	*p*-values	3 B × 2 B × 2 W	*p*-values[1]
Parametric/randomization		**Parametric/randomization**		**Parametric/randomization**	
F-test	0.0710	F-test	0.0043	F-test	0.0692
Wrapper F-test	0.0735	Wrapper F-test	0.0046	Wrapper F-test	0.0728
Nonparametric		**Nonparametric**		**Nonparametric**	
Brunner F1_LD_F1		Brunner LD_F2		Brunner F2_LD_F1	
F-appr. small N	0.1421	F-appr. large N	< 0.0001	F-appr. large N	< 0.0001
		F-appr. small N	0.0015	F-appr. small N	0.0412
Aligned ranks		Aligned ranks		Aligned ranks	
regular	0.0427	regular	0.0099	regular	
Friedman	0.0054	Friedman	0.0065	short format	0.0753
Koch	0.0339	ARTool	0.0099	long format	0.1528
ARTool	0.0544			Friedman	0.0133
				ARTool	0.1425

[1]Note: Only the triple interaction *p*-values are reported

Parametric tests

For the data in long format, Levene's test for homogeneity of variance between Groups was not significant, $p = 0.0669$. For the data in short format, the Mauchly sphericity test (for transformed variates) was also not significant, $p = 0.2053$. The Greenhouse-Geiser (G-G) epsilon was $\varepsilon = 0.6220$. O'Brien and Kaiser (1985) suggested that, when you have a large violation of sphericity (e.g. $\varepsilon < 0.70$) and your sample size is greater than $k + 10$ (i.e., the number of levels of the repeated measures factor + 10), then a MANOVA is more powerful; in other cases, the repeated measures design should be selected. Here, the $N = 12$ was not larger than 15 (5 time measures + 10), so we further concentrated on the repeated measures design. The *p* value for the F-test on the Group × Time interaction was not significant, $p = 0.0710$. The wrapper randomization test resulted in a similar *p*-value: 0.0735.

Nonparametric tests

Brunner's macro F1_LD_F1 also resulted in a *p*-value which was not significant for small samples, $p = 0.1421$. The value for large samples could not be calculated because the covariance matrix was singular and hence could not be inverted.

The Beasley and Zumbo (2009) procedure yielded all significant *p*-values, 0.0427, 0.0054, and 0.0339, respectively, for the aligned regular, Friedman, and Koch rank tests. The ARTool resulted in a non-significant value $p = 0.0544$. The correlation of the alignment with the ARTool and the alignment with the H&T formula was $r\ (N = 60) = 0.7811$, again indicating that the alignments were not the same.

Example 3: A Doubly Repeated Measures, 2 × 3 Within × Within Design, Small Sample Size

The aligned data for testing within × within, A × B interactions is given by the same H&T formula (2) as for the between × between design. In this example, I fabricated data for a 2 within × 3 within design. There were six repeated measures; one record per observation, with six columns as doubly repeated measures A and B. The means of repeated measures 1 + 2 + 3 corresponded to A level 1 and the means of the measures 4 + 5 + 6 corresponded to A level 2. The means of measures 1 + 4, 2 + 5, and 3 + 6 related to the B levels 1, 2, and 3, respectively. The results of the different tests for this example are reported in the middle part of Table 2.

Parametric tests

For this example the data, with $N = 12$ observations, were made up to yield a significant interaction, $p = 0.0043$. The randomization test wrapper resulted in an equally significant interaction value, $p = 0.0046$. The G-G ε was quite large: 0.9270.

Nonparametric tests

Brunner's macro LD_F2 is suited for doubly repeated measures. When applied on these data, both F-approximation tests, for large and small sample sizes, were quite significant interaction: $p < 0.0001$ and $p = 0.0015$, respectively.

The Beasley and Zumbo (2009) script conducted on these data confirmed the significant interaction, with *p* values = 0.0099 and 0.0065, respectively, for the aligned regular rank and the aligned Friedman rank tests. To achieve the Friedman rank test, for each observation, the aligned data were ranked from 1 to 6, and then a 2 × 3 repeated measures ANOVA was applied on these ranks. Note that Koch's ranking method is not applicable with doubly repeated measures because there are no between Groups. The ARTool returned a *p* value which was the same as with

the aligned regular rank test, $p = 0.0099$. The correlation of the alignment with the ARTool and the alignment with the H&T formula was perfect here, $r\ (N = 72) = 1$.

Example 4: A Three-Way 3 × 2 × 2 Between × Between × Within Design

The fourth example was taken from Cody and Smith (1987, p. 159), a three-factor experiment with repeated measures on the last factor. They invented data for a market experiment on male and female (Sex factor) subjects who were offered three different Brands of coffee. Each brand was tasted twice (Time factor); once after breakfast, once after dinner. The preference of each brand was measured on a 10 point scale. The Shapiro Wilk test on this taste measures, over Brand, Sex, and Time combinations (in long format), was not significant, $p = 0.9465$. There were three subjects in each Brand × Sex condition combination, resulting in a total $N = 18$.

Parametric tests

There were very significant Time and Brand effects, both with $p < 0.0001$, but the triple interaction Brand × Sex × Time was not significant, $p = 0.0692$. The two-way interactions, Brand × Sex, Time × Brand and Time × Sex, were also not significant. In this example we concentrated only upon the triple interaction. The results of the different ways of analysis for the triple interaction are presented in the last part (on the right) of Table 2. The randomization test wrapper also showed a non-significant $p = 0.0728$ for this interaction.

Nonparametric tests

Brunner's macro F2_LD_F1 is suited for this design with two between factors and one repeated measures factor. When applied on these data, both F-approximation tests for the triple interaction Brand × Sex × Time, for large and small sample sizes, were significant: $p < 0.0001$ and $p = 0.0412$, respectively.

The data were aligned with the H&T schematic formula (4) for the triple interaction. The aligned data were first ranked in short format, with the two aligned repeated measures ranked separately, but over all Brand and Sex levels combinations. On these ranked aligned data, the triple interaction was not significant, $p = 0.0753$. Yet the Friedman ranking on the aligned data in short

format resulted in a significant triple interaction, $p = 0.0133$. Note that the Koch ranking was not applied because this is too complicated in this example.

When ranking the data aligned in long format (over Brands, Sex, and Time), as is done with the ARTool, the ANOVA yielded an F-value which was also not significant, but with a somewhat larger $p = 0.1528$. The same test on the ARTool ranked data was also not significant, $p = 0.1425$. The correlation between H&T formula aligned data and the ARTool alignment for the triple interaction was almost perfect here, r ($N = 36$) = 0.9997. The alignments were not exactly the same due to small differences in numerical precision. In the calculations of the means, only six decimals were used, whereas the ARTool uses nine.

Because of the discrepancy (in this example, not in the other examples) between the *p* values obtained with rankings of the aligned data in short versus long format, the ranking data were inspected and it was found that the discrepancy was due to the fact that the ranking in long format was more precise; it ranged from 1 to 36 (18 observations × 2 Times) and contained 8 ties. However, the rankings in short format twice ranged from 1 to 18 and had 10 ties and 3 doubles (equal ranks).

Discussion

Example 1: The Pretest-Posttest Design

Bonate (2000, p. 103) proposed two ways for dealing with an outlier: simply removing the outlier or applying a method to minimize the influence of an observation on parameter estimations, namely the iterative reweighted least-squares (IRWLS). Two weight functions were used for the iterations, the Huber function and the bisquare. Removing the outlier from the data in this example resulted in a p value < 0.0001, as did both weight functions. Fagerland (2012) suggested using the Yuen-Welch t-test for trimmed means in situations with outliers. Applying this test here returned a similar p-value < 0.0001.

Gain scores

All parametric tests for the interaction on the gain scores used in this example were not significant, apparently because of the outlier. The randomization tests only confirmed these values. Yet all parametric tests, after ranking the gain scores, were significant (< 0.01), and the nonparametric procedures also all resulted in significant p values (< 0.01).

Repeated measures

With respect to the repeated measures analyses of the interaction, the parametric test and randomization test both were not significant, but the parametric tests, after ranking the repeated measures, were also not significant (contrary to such tests after ranking the gain scores). The new rank-based methods yielded either barely well or barely not significant *p*-values, whereas the analyses on the aligned regular, Friedman, and Koch rank tests were clearly significant (< 0.01). The ARTool application gave a very significant *p* value, but the correlation between the ARTool aligned and the H&T aligned data was somewhat poor (0.7386), indicating that the ARTool alignment may not be flawless in this repeated measures design. The next smallest *p*-value was for the aligned Friedman ranking.

(R)ANCOVA

In the literature, there is controversy about gain scores versus repeated measures and ANCOVA (see e.g., Senn, 2006; Knapp & Schafer, 2009; Smolkowski, 2013). I do not go into this discussion here. Dimitrov and Rumrill (2003) concluded that the analysis of gain scores is okay, but that the ANCOVA should be the preferred method for analysis of pretest-posttest data. The ANCOVA (especially the wrapper version) was somewhat less affected by the outlier, which seems to support Dimitrov and Rumrill's (2003) conclusion. The RANCOVA of the ranked posttest on the ranked pretest as covariate for the Group factor and the tests on the residual ranks all gave *p* values < 0.01. Dimitrov and Rumrill (2003) also noted that the attractive characteristic of residual scores is that, unlike gain score, they do not correlate with observed pretest scores. Compared with the ANCOVA model, however, they specified that an ANOVA on residuals (e.g., Quade's RANCOVA) is less powerful (p. 161).

Example 2: The 3 × 5 Between × Within Design

This two-way mixed 3 × 5 design was on an ordinal scale. With Likert-type or ordinal data, nonparametric tests are usually advised (see e.g., Kaptein et al., 2010; De Winter & Dodou, 2012), especially with small sample sizes, as in this example. Fagerland (2012) pointed to a paradox in statistical practice in high-impact medical journals, namely that the median sample size research studies published has increased manifold, while the use of nonparametric tests has increased at the expense of t-tests. It was concluded nonparametric tests should only be used with

small samples, because such tests in large studies may provide answers to the wrong question.

The parametric tests *p*-values for the interaction were not significant. All nonparametric tests in this example revealed that the interaction was significant, except for the ARTool application and the Brunner methods. The correlation between the ARTool aligned data and the H&T aligned data was slightly poor (0.7811) again, as in Example 1, showing that the ARTool does not align the data faultlessly in this design. The aligned regular rank, the aligned Friedman rank, and the aligned Koch rank tests all returned *p*-values < 0.01.

Example 3: The Doubly Repeated 2 × 3 Within × Within Design

The data for this example were fabricated to result in a significant *p*-value for the interaction. All types of tests revealed this significant interaction with *p*-values < 0.01. The Brunner F-approximation for large samples was even < 0.0001. Here, the ARTool alignment and ranking were the same as the H&T alignment and ranking.

Example 4: The 3 × 2 × 2 Between × Between × WIthin Design

This higher order design did not have a significant F-value for the triple interaction when tested parametrically or with the randomization test wrapper. Yet, when tested with the Brunner new rank-based methods or with the aligned Friedman ranking, the resulting *p*-values were significant. All other nonparametric tests were not significant. The Shapiro-Wilk test indicated that the distribution was quite normal and the interaction should not really be tested with nonparametric tests. This implies that significant *p*-values for the triple interaction in this example may be invalid.

The small discrepancy between the *p* values obtained after ranking in short versus long format was most probably due to the fact that the ranking in long format was more precise and less affected by ties.

Conclusion

The obtained *p*-values for the parametric analyses of interactions in all examples were very comparable to the values obtained by the randomization methods. In Example 1, the outlier masked the interaction with the parametrical tests as well as with randomization tests. In my opinion, the randomization tests can be seen only

as confirmation/rejection tools, not as nonparametric alternatives; it is as if they copy the parametric results.

With respect to the ARTool application, for fully between designs and fully within designs, and especially for higher order (> 2 way) such designs, this tool is indeed very convenient, as claimed by Wobbrock et al. (2011). Yet the ARTool alignment of the data for the interactions was questionable in Examples 1 and 2. Further research could perhaps clarify way the ARTool alignment of the data in these two examples was different from the H&T alignment.

The new rank-based methods can be an alternative way of looking at the data. They test treatment effects without assuming differences in location parameters (e.g., means). They test differences among distributional characteristics. One problem might be that it is not always precisely clear how population distributions differ, as noted by Serlin and Harwell (as cited in Beasley & Zumbo, 2009, p. 19). In Examples 1 and 2, these methods showed a non-significant *p* value for the F-approximations for small samples; this, in contrast to all (except for ARTool in Example 2) other nonparametric tests' *p*-values. On the other hand, in the last example, only the Friedman aligned ranks tests and this Brunner technique showed a significant triple interaction, although the data fabricated by Cody and Smith (1987) did not have a significant triple interaction with a parametric test and the distribution was quite normal. In the examples (except for Example 3) the new rank-based methods contradicted the results of other (except for the Friedman) nonparametric tests.

All the examples showed that the aligned (regular) rank test with the Higgins and Tashtoush formulae was quite sensitive to detect the interactions which should be spotted as significant. Except for pretest-posttest design, this test should be applied in factorial designs with repeated measures. In pretest-posttest designs, the RANCOVA (not Quade's ANOVA on residuals) should be the preferred way of testing. Ranking should be done on the aligned data in long format because it is more precise than separate rankings of aligned repeated measures (short format). The aligned Friedman rank tests were significant in all examples. It is advisable to report both the aligned regular rank test and the aligned Friedman rank test. They rank the data differently; the aligned regular rank test does so across observations, the Friedman test ranks the data within observations.

References

Akritas, M. G. (1990). The rank transform method on some two factor designs. *Journal of the American Statistical Association, 85*(409), 73-78. doi: 10.1080/01621459.1990.10475308

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*(425), 336-343. doi: 10.1080/01621459.1994.10476475

Akritas, M. G., Arnold, S. F., & Brunner, E. (1997). Nonparametric hypothesis and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association, 92*(437), 258-265. doi: 10.1080/01621459.1997.10473623

Akritas, M. G., & Brunner, E. (1997). A unified approach to rank tests in mixed models. *Journal of Statistical Planning Inference, 61*(2), 249-277. doi: 10.1016/S0378-3758(96)00177-2

Beasley, T. M., & Zumbo, B. D. (2003). Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. *Computational Statistics & Data Analysis, 42*(4), 569-593. doi: 10.1016/S0167-9473(02)00147-0

Beasley, T. M., & Zumbo, B. D. (2009). Aligned rank tests for interactions in split-plot designs: Distributional assumptions and stochastic homogeneity. *Journal of Modern Applied Statistical Methods, 8*(1), 16-50. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol8/iss1/4/

Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. London, UK: Chapman & Hall/CRC.

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

Brunner, E., & Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers, 42*(1), 1-52. doi: 10.1007/s003620000039

Cassell, D. L. (2002). A randomization-test wrapper for SAS® PROCs. *Proceedings of the Twenty-seventh Annual SAS users Group International Conference*. Cary, NC: SAS Institute Inc. Retrieved from http://www2.sas.com/proceedings/sugi27/Proceed27.pdf

Cody, R. P., & Smith, J. K. (1987). Applied statistics and the SAS programming language (2nd ed.). New York, NY: North-Holland.

Conover, W. J., & Iman, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs. *Communication in Statistics – Theory and Methods, 5*(14), 1349-1368. doi: 10.1080/03610927608827447

De Winter, J. C. F., & Dodou, D. (2010). Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation, 15*(11). Retrieved from http://pareonline.net/getvn.asp?v=15&n=11

Dimitrov, D. M., & Rumrill, P. D., Jr. (2003). Pretest-posttest designs and measurement of change. *Work, 20*(2), 159-165.

Fagerland, M. W. (2012). t-tests, non-parametric tests, and large studies – a paradox of statistical practice? *BMC Medical Research Methodology, 12*(78). doi: 10.1186/1471-2288-12-78

Gao, X., & Alvo, M. (2005a). A nonparametric test for interaction in two-way layout. *The Canadian Journal of Statistics, 33*(4), 529-543. doi: 10.1002/cjs.5550330405

Gao, X., & Alvo, M. (2005b). A unified nonparametric approach for unbalanced factorial designs. *Journal of the American Statistical Association, 100*(471), 926-941. doi: 10.1198/016214505000000042

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

Higgins, J. J., & Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World, 1*(2), 201-211.

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

Kaptein, M., Nass, C., & Markopoulos, P. (2010). Powerful and consistent analysis of Likert-type rating scales. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM Press.

Kelley, D. L., & Sawilowsky, S. S. (1997). Nonparametric alternatives to the f statistic in analysis of variance. *Journal of Statistical Computation and Simulation, 58*(4), 343-359. doi: 10.1080/00949659708811839

Knapp, T. R., & Schafer, W. D. (2009). From gain score t to ANCOVA F (and vice versa). *Practical Assessment, Research & Evaluation, 14*(6). Retrieved from http://pareonline.net/getvn.asp?v=14&n=6

Leys, C., & Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology, 46*(4), 684 688. doi: 10.1016/j.jesp.2010.02.007

McSweeney, M. (1967). An empirical study of two proposed nonparametric tests for main effects and interaction (Doctoral dissertation). Retrieved from Dissertation abstracts international. (1968-16394-001)

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD: an R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software, 50*(12). doi: 10.18637/jss.v050.i12

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. Psychological Bulletin, 97(2), 316-333. doi: 10.1037/0033-2909.97.2.316

Peterson, K. (2002). Six modifications of the aligned rank transform test for interaction. *Journal of Modern Applied Statistical Methods, 1*(1), 100-109. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol1/iss1/13/

Puri, M. L., & Sen, P. K. (1985). *Nonparametric methods in general linear models*. New York, NY: Wiley.

Quade, D. (1967). Rank analysis of covariance. *Journal of the American Statistical Association, 62*(320), 1187-1200. doi: 10.1080/01621459.1967.10500925

R Core Team. (2012). *R: a language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Richter, S. J., & Payton, M. E. (1999). Nearly exact tests in factorial experiments using the aligned rank transform. *Journal of Applied Statistics, 26*(2), 203-217. doi: 10.1080/02664769922548

Rodriguez, O. J. C., Álvarez, G. J., & Remirez, R. J. (2009). Análisis no paramétrico de la interacción de dos factores mediante el contraste de rangos alineados [An aligned rank test for a nonparametric analysis of the two-way interaction]. *Psicothema, 21*(1), 152-158.

Salazar-Álvarez, M., Tercero-Gómez, V. G., Temblador-Pérez, M., & Conover, W. J. (2014). Nonparametric analysis of interactions: a review and gap analysis. *Proceedings of the 2014 Industrial & Systems Engineering Research Conference*. Montréal, Canada.

SAS Institute Inc. (2008). The SAS System, version 9.2 [Computer software]. Cary, NC: SAS Institute Inc.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60*(1), 91-126. doi: 10.3102/00346543060001091

Sen, P. K. (1968). On a class of aligned rank order tests in two-way layouts. *Annals of Mathematical Statistics, 39*(4), 1115-1124. Retrieved from http://www.jstor.org/stable/2239679

Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine, 25*(24), 4334-4344. doi: 10.1002/sim.2682

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

Smolkowski, K. (2013). *Gain score analysis* [Unpublished report]. Retrieved from http://homes.ori.org/~keiths/Tips/Stats_GainScores.html

Toothaker, L. E., & Newman, D. (1994). Nonparametric competitors to the two-way ANOVA. *Journal of Educational & Behavioral Statistics, 19*(3), 237-273. doi: 10.3102/10769986019003237

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedure. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM Press.