Austrian Journal of Statistics
2023, Volume 52, 101–130.
http://www.ajs.or.at/
doi:10.17713/ajs.v52i5.1545

Generalizations of the Tests by Kruskal-Wallis, Friedman and van der Waerden for Split-plot Designs

Haiko Luepsen
University of Cologne

Abstract

Generalizations of the 1-factorial tests by Kruskal-Wallis and Friedman, as well as of the van der Waerden test are proposed for factorial split-plot designs, both allowing interactions. They are compared in regard to the type I error control and the power with the parametric F test, including the Huynh-Feldt adjustment, the inverse normal transform (INT), the ANOVA type statistic by Brunner et al. (ATS), the aligned rank transform (ART), the L statistic by Puri & Sen and a procedure by Koch. The two methods proposed show a perfect type I error control, except for two situations, and an attractive power, particularly in case of nonnormal distributions. The charm and advantage of these procedures are the possibility to apply them with statistical standard tools using only variable transformations and data management, and to receive results from well-known methods which are easy to understand.

Keywords: ANOVA, split-plot design, Kruskal Wallis, Friedman, van der Waerden, Puri & Sen, Koch, ART, ATS.

1. Introduction

Split plot designs, repeated measures designs involving two or more independent groups, are among the most common experimental designs in educational, psychological, medical and many other fields of scientific research. For analyzing data from such designs, in general the parametric ANOVA model is applied, which requires normality of the residuals, sphericity and homogeneity of the covariance matrices as well as the independence of the observed units (see e.g. Beasley and Zumbo 2009; Winer, Brown, and Michels 1991). Most people trust in the robustness of the parametric tests though it is not as great as for between subject designs. But during the last decades a number ANOVA procedures have been suggested for situations, when the assumptions of the parametric model are not met, also for analyzing data from split-plot designs. Here to mention first the rank based methods: the rank transform method RT by Conover and Iman (1981), the inverse normal transform method INT (see e.g. Mansouri and Chang 1995), the tests by Puri and Sen (1985), here denoted by *PS* and often referred as L statistic (see e.g. Harwell and Serlin 1989) and Koch's ANOVA for split-plot designs (Koch 1969).

A number of authors express concern about ranking methods for factorial ANOVA designs. Toothaker and Newman (1994) as well as Beasley and Zumbo (2009), to name only a few, found out that the type I error rate of the interaction can reach beyond the nominal level, if there are significant main effects because the effects are confounded. Headrick and Sawilowsky (2000) demonstrated this phenomenon computationally. On the other hand, the RT lets sometimes vanish an interaction effect, as Salter and Fawcett (1993) had shown in a simple example. The reason: *„additivity in the raw data does not imply additivity of the ranks, nor does additivity of the ranks imply additivity in the raw data*", as Hora and Conover (1984) found out. Payton, Richter, Giles, and L. (2006) pointed out this problem in connection with split-plot designs. There are a couple of remedies. First and often used, the aligned rank transform ART with an alignment for the interaction (see e.g. Beasley and Zumbo 2009). Secondly the anova type statistic ATS by Brunner, Munzel and Akritas (see e.g. Brunner, Munzel, and Puri 1999; Bathke, Schabenberger, Tobias, and Madden 2009), a method with a nonparametric model based on relative effects. Thirdly to mention the generalized linear models GLM, from which GEE (*Generalized Estimating Equations*), established by Liang and Zeger (1986), and GLMM (*Generalized Linear Mixed Models*, sometimes also called *MLM, multi level models*) by Harville (1977) are probably the most popular. They allow correlated responses and nearly arbitrary covariance matrices, but are based on large sample asymptotic theory and need large $n_i \geq 50$ (see e.g. Stiger, Kosinski, Barnhart, and Kleinbaum 1998). These methods as well as some more are discussed in detail e.g. by Algina (1994), Stiger *et al.* (1998) and Keselman, Algina, and Kowalchuk (2001).

The most well-known nonparametric ANOVA tests perhaps, the Kruskal-Wallis H-test (K-W) for between-subject designs and the Friedman ANOVA for repeated measures designs, do not appear in the list above, because they are designed for the analysis of 1-factorial designs. It is shown here that also a corresponding procedure for factorial split-plot designs can be defined, which incorporates these two methods as special cases, and will be denoted here by *KWF*. Another well-known procedure, the van der Waerden test (van der Waerden 1953), has been developed so far only for factorial between subject designs (Mansouri and Chang 1995) and for 1-factorial repeated measures designs (Marascuilo and McSweeney 1977). Here, an extension for the analysis of split-plot designs is proposed and will be named *van der Waerden scores test* and abbreviated *vdWS*. The van der Waerden test is closely related to the K-W and the Friedman tests. If the inverse normal transformation (for a definition see next section) is applied to the ranks, on which these tests are based, before applying the χ^2 -test, i.e. transforming them into normal scores, then the van der Waerden test is the result. Finally to mention that (Beasley 2000) suggested an interaction in split-plot designs based on the Friedman test. But he also pointed to several problems with this solution, so that it is not generally advisable. Own simulations, generally showing type I error rates of 30 percent for a nominal level of 5 percent, confirmed these problems.

The main motivation for the proposition of the KWF and vdWS methods arises from their popularity, particularly of the Kruskal-Wallis and the Friedman tests, the basis of the KWF, as well as the positive evaluation of these two tests and the van der Waerden test. Here to mention the articles by Feir and Toothaker (1974), Lix, Keselman, and Keselman (1996) and Sawilowsky (1990) considering the K-W test, Marascuilo and McSweeney (1977), Harwell and Serlin (1994) and Ernst and Kepner (1993) investigating the Friedman test, and finally Dijkstra (1987), Sheskin (2004) and Luepsen (2018) who studied the van der Waerden test. They all mention the robustness against nonnormality and heterogeneity. In the case of between subject designs, the van der Waerden test scored overall the best in view of type I error control and power, compared with other rank based procedures (see e.g. Luepsen 2018). One attractive feature of the two methods suggested here: both can be computed using standard statistical software like SAS and SPSS together with some variable transformations (see e.g. Luepsen 2020). This enables researchers to analyze split-plot designs even if programs for more sophisticated methods like the above quoted ATS, GEE or GLMM are not available.

A view words to the composition of this study. In section 2 the rank based ANOVA procedures

considered here will be described. Particularly, at the end of section 2 the proposed KWF and vdWS methods will be presented. As the main objective it will be proved in section 3 that the KWF and the vdWS methods are generalizations of the well-known classical tests, even though there is no complete coincidence for the latter one in the case of ties. In the same section it will be demonstrated that the differences are neglectable. To show the reliability of the proposed methods, they will be compared with the F test and other rank based ANOVA methods using a Monte Carlo simulation (section 4). Special attention has to be laid upon the interaction effect, whereas the main effects coincide with the established Kruskal Wallis, Friedman and van der Waerden tests. But also the models with nonnull main effects have to be examined, because first, such situations are not covered by the 1-factorial analyses, and secondly, according to the above concerns about interactions a control of their type I error seems reasonable. It is neither intended to present a statistical factorial model for the procedures proposed, nor a complete comparison of the procedures quoted above, which will be subject of a larger study.

2. The methods to be considered

2.1. The parametric F-test

At first the model for the mixed design and the parametric F-test will be given with the corresponding sum of squares, as some formulae will refer to these later. Here the classical approach will be used (see e.g. Winer *et al.* 1991), though in recent publications often mixed models, considering e.g. covariance structures, are preferred. For one grouping factor A and one repeated measures factor B the 2-factorial ANOVA model for a dependent variable y shall be denoted by

$$\gamma_{ijm} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \tau_{im} + (\beta\tau)_{ijm} + e_{ijm} \tag{1}$$

with fixed effects α_i (factor A, $i = 1, .., I$), β_j (factor B, $j = 1, .., J$), $(\alpha\beta)_{ij}$ (interaction AB), n_i subjects per group ($i = 1, .., I$), a subject specific variation τ_{im} ($m = 1, .., n_i$), multivariate normal distributed error e_{ijm} with covariance matrices Σ_i, and N=$\sum n_i$. Additionally, the F-test assumes the pooled covariance matrix to be spherical and equal for $i = 1, .., I$. The parameters α_i , β_j and $(\alpha\beta)_{ij}$, with the restrictions $\sum \alpha_i = 0$, $\sum \beta_j = 0$, $\sum (\alpha\beta)_{ij} = 0$, can be estimated by means of a linear model $\mathbf{y}^\top = \mathbf{X}\mathbf{p}^\top + \mathbf{e}^\top$ using the least squares method, where $\mathbf{y}$ are the values of the dependent variable, $\mathbf{p}$ the vector of the parameters, $\mathbf{X}$ a suitable design matrix and $\mathbf{e}$ the random variable of the errors. If the contrasts for the tests of the hypotheses $\mathrm{H}_A(\alpha_i = 0), \mathrm{H}_B(\beta_j = 0)$ and $\mathrm{H}_{AB}((\alpha\beta)_{ij} = 0)$ are orthogonal, the resulting sum of squares SS_A, SS_B, SS_{AB} of the parameters are also orthogonal and commonly called type III SSq. The sums of squares and mean squares of the effects are computed as follows:

$$SS_A = J\sum_i n_i(\bar{y}_{i..} - \bar{y})^2 \quad SS_B = N\sum_j (\bar{y}_{.j.} - \bar{y})^2 \quad SS_{AB} = \sum_i \sum_j n_i(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y})^2$$
$$MS_A = SS_A/(I-1) \qquad MS_B = SS_B/(J-1) \qquad MS_{AB} = SS_{AB}/((I-1)(J-1)) \tag{2}$$

where $\bar{y}_{i..}, \bar{y}_{.j.}$ are the level means of factor A and B, $\bar{y}_{ij.}$ are the cell means and $\bar{y}$ is the grand mean. Finally the sums of squares and mean squares of the error terms:

$$MS_{between} = J\sum_i \sum_m (\bar{y}_{i.m} - \bar{y})^2/(N-1)$$
$$MS_{within} = J\sum_i \sum_m \sum_j (y_{ijm} - \bar{y}_{i.m})^2/(N(J-1))$$
$$MS_{error(between)} = J\sum_i \sum_m (\bar{y}_{i.m} - \bar{y}_{i..})^2/(N-I)$$
$$MS_{error(within)} = J\sum_i \sum_m \sum_j (y_{ijm} - \bar{y}_{i.m} - \bar{y}_{ij.} + \bar{y}_{i..})^2/((N-I)(J-1)) \tag{3}$$

and the F-ratios as

$$F_A = MS_A/MS_{error(between)} \quad F_B = MS_B/MS_{error(within)} \quad F_{AB} = MS_{AB}/MS_{error(within)}$$

which are F distributed as long as the assumptions mentioned above are fulfilled. To make up for nonspherical data, e.g. heterogeneous variances on factor B, an appropriate adjustment of the degrees of freedom for the F-test is applied. Here the Huynh-Feldt adjustment including the correction by Lecoutre, abbreviated H-F, is chosen (see e.g. Winer, Brown, and Michels 1991; Quintana and Maxwell 1994).

2.2. Nonparametric model

The null hypotheses for the nonparametric methods have to be formulated in a different way. The model underlying a split-plot design can be described by independent continuous random vectors $\mathbf{Y}_{im} = (Y_{i1m}, \ldots, Y_{iJm})^\top$ with marginal distributions $Y_{ijm} \sim F_{ij}$ for the jth observation of subject m in group i. Whereas Kruskal & Wallis as well as Friedman state in their original work simply "equal distribution functions" as H_0, Koch (1969) and Gibbons and Chakraborti (2020) modify the hypotheses in order to restrict differences to location differences: if $F_i(\mathbf{Y}) = G(\mathbf{Y} - \boldsymbol{\mu_i})$ is the distribution function of $\mathbf{Y}$ in group i with vectors $\boldsymbol{\mu_i} = (\mu_{i1}, ..., \mu_{ij})$ of J location parameters (e.g. medians) and a distribution function G characterizing their shape, then

$$\begin{aligned} \mathrm{H}_A : \boldsymbol{\mu_1} = ... = \boldsymbol{\mu_I} \quad \mathrm{H}_B : \mu_{i1} = ... = \mu_{iJ} \quad \text{for} \quad i = 1, .., I \quad \text{and} \\ \mathrm{H}_{AB} : \gamma_{ij} = 0 \quad \text{for} \quad i = 1, .., I \quad \text{and } j = 1, .., J \end{aligned} \tag{4}$$

if $\gamma_{ij} = \mu_{ij} - \bar{\mu_{.j}} - \bar{\mu_{i.}} + \bar{\mu}$ where $\bar{\mu_{.j}}$ and $\bar{\mu_{i.}}$ are the location parameters related to the effects of A and B, and $\bar{\mu}$ the grand mean. In contrast, Brunner et al. (1999) use a different model for their ATS method, which does not require equal distribution shapes: if $\bar{F}_{i.}(y) = (1/J)\sum_j F_{ij}(y)$, $\bar{F}_{.j}(y) = (1/N)\sum_i n_i F_{ij}(y)$ and $\bar{F}_{..}(y) = (1/NJ)\sum_i \sum_j n_i F_{ij}(y)$ then the null hypotheses can be expressed in terms of marginal distribution functions (see e.g. Noguchi, Gel, Brunner, and Konietschke 2012):

$$\mathrm{H}_A : \bar{F}_{1.} = ... = \bar{F}_{I.} \quad \mathrm{H}_B : \bar{F}_{.1} = ... = \bar{F}_{.J} \quad \mathrm{H}_{AB} : \bar{F}_{ij} = \bar{F}_{i,} - \bar{F}_{.j} + \bar{F}_{..}$$

Here the normalized distribution function $F_{ij}(y) = [P(Y_{ij} \leq y) + P(Y_{ij} < y)]/2$ is used to allow tied values.

2.3. RT (rank transform) and INT (inverse normal transform)

The rank transform method (RT), proposed by Conover and Iman (1981), comprises just the transformation of all NJ values of y into ranks $R(y)$, so-called *Wilcoxon* ranks, before applying the parametric F-test to them, as described above. Here midranks are used in case of tied values. Huang (2007) as well as Mansouri and Chang (1995) showed, that applying the inverse normal transform (INT) to the ranks $R(y)$, i.e. computing their normal scores before computing the parametric F-test, results in an improvement of the RT procedure. The normal scores are defined as $z(y) = \phi^{-1}(R(y)/(n+1))$, where ϕ^{-1} denotes the inverse cumulative normal distribution function and n the number of observations, here n=NJ. Huang (2007) showed that the classical F statistic applied to normal scores $z(y)$ has the same limiting null distribution as when applied to normal data.

2.4. ART (aligned rank transform) and ART INT

For using the RT method an alignment is proposed in order to avoid an increase of type I error rates for the interaction due to nonnull main effects: all effects not to be tested are subtracted from y before performing the parametric analysis of variance, a method dating back to Hodges and Lehmann (1962). The procedure consists of first computing the residuals,

either as differences from the cell means or by means of a regression model, then adding the effect to be tested, transforming this sum into ranks and finally performing the parametric F test on them. For the alignment in the 2-factorial model (1) first the error e_{ijm} is computed as the residuals from a parametric between subject ANOVA including all effects and the subject specific variation τ_{im}. In the next step the means corresponding to the effect being tested (A, B or AB) are added:

$$y_{ijm}^{(A)} = e_{ijm} + a_i \quad y_{ijm}^{(B)} = e_{ijm} + b_j \quad y_{ijm}^{(AB)} = e_{ijm} + ab_{ij}$$

where a_i, b_j, ab_{ij} are the means of y corresponding to the effect. Then the aligned variables $y^{(A)}, y^{(B)}, y^{(AB)}$ are each transformed into Wilcoxon ranks $R(y^{(A)}), R(y^{(B)}), R(y^{(AB)})$. To test one effect the parametric F test is applied to the corresponding aligned variable, where only that effect is examined ignoring the other two. Salter and Fawcett (1993) showed that the normal theory F tests used for testing these rank statistics are valid, because their asymptotic distributions are the same.

Mansouri and Chang (1995) as well as Luepsen (2018) suggested to apply the normal scores transformation INT (see above) to the ranks obtained from the ART procedure. They showed that the transformation into normal scores improves the type I error rate for the ART procedure, too, at least in the case of underlying normal distributions. Computationally the ranked aligned variables $R(y^{(A)}), R(y^{(B)}), R(y^{(AB)})$ are to be transformed into normal scores, as described for the INT method, before applying the parametric F test on them. This procedure will be denoted here by *ART INT*. Unfortunately these methods react on discrete outcomes with increasing error rates (see e.g. Luepsen 2017).

2.5. Puri & Sen tests (L statistic)

The tests by Puri and Sen (1985), often referred as L statistic, offer a nonparametric test statistic for the General Linear Model (see e.g. Harwell and Serlin 1989). The resulting test statistics are asymptotically χ^2 distributed. They can be seen as a generalization of the well-known Kruskal-Wallis H test (for independent samples). The χ^2-ratios are computed in the case of only grouping factors as

$$\chi^2_{effect} = SS_{effect}/MS_{total}$$

and in the case of a mixed design for the tests of A, B and AB as

$$\chi^2_A = SS_A/MS_{between} \quad \chi^2_B = SS_B/MS_{within} \quad \chi^2_{AB} = SS_{AB}/MS_{within}$$

Here SS_A, SS_B, SS_{AB} or generally SS_{effect}, are the sum of squares as outlined before (2), but computed for *R(y)*, the Wilcoxon ranks of y, in the same way as for the RT procedure. $MS_{between}$ and MS_{within} are the mean squares previously defined, and MS_{total} the variance of *R(y)*. The degrees of freedom are those of the numerator of the corresponding F test. The major disadvantage of this method is the lack of power for any effect in the case of other nonnull effects in the model.

Similar to the modification of the RT and the ART techniques, the PS method can be improved by the INT transformation of R(y) into normal scores $\phi^{-1}(R(y)/(n+1))$ of y (n=NJ) before applying the χ^2 tests in the same way as for the L statistic. This leads to a procedure, denoted by *PS INT*, which has been proposed already by Puri and Sen (1969).

2.6. Koch's ANOVA

Koch (1969) proposed a couple of nonparametric procedures for split-plot designs based on a multivariate version of the Kruskal-Wallis test and a nonparametric analogue of the one-way MANOVA based on the trace (see e.g. Chatterjee and Sen 1966). The resulting test statistics are approximately χ^2 distributed. There are several variants for the cases with and without

compound symmetry, as well as with and without independence of the factors A and B. The version used here assumes an interaction, but no compound symmetry. A detailed description of the method and the extensive computational procedure can be found in Koch (1969) and shall not be reproduced here.

2.7. Tests by Brunner, Munzel and Puri (ATS)

The authors reflect the relative effect of a random variable X_1 to a second one X_2 , i.e. the probability that X_1 has smaller values than X_2 , which is defined as $p = P(X_1 < X_2) + P(X_1 = X_2)/2$, considering the case $P(X_1 = X_2) > 0$. As the definition of relative effects is based only on an ordinal scale of y, this method is suitable also for variables of ordinal or even dichotomous scale (see e.g. Noguchi *et al.* 2012). Based on above model (4) they developed two tests to compare samples by means of comparing the relative effects: the approximately F distributed ATS (*anova-type statistic*) and the asymptotically χ^2 distributed WTS (*Wald type statistic*). In contrary to the WTS, the ATS accounts for the sample sizes that makes it attractive for small cell counts. These tests have been extended to repeated measures designs by Brunner *et al.* (1999). Bathke *et al.* (2009) described the procedures, which involve a lot of matrix algebra and shall not be reproduced here.

2.8. The generalized Kruskal & Wallis and Friedman procedure (KWF)

First the well-known rank tests by Kruskal & Wallis and by Friedman will be quoted, because these are essential for the proposed algorithms. Let $x_{im} (i = 1, .., I \text{ and } m = 1, .., n_i)$ be the observations in a 1-way layout with I independent groups of factor A, R_{im} the Wilcoxon rank of x_{im}, $R_{i.} = \sum_m^{n_i} R_{im}$ and $N = \sum n_i$. Then the Kruskal-Wallis H

$$H = \frac{12}{N(N+1)} \sum_i^I \frac{R_{i.}^2}{n_i} - 3(N-1) \tag{5}$$

tests the identity of the I distribution functions, where H is χ^2-distributed with (*I-1*) df. The test assumes equal shapes of the distribution function for all groups. It is well-known that the H test can be performed by ranking y (using midranks for ties), conducting a parametric ANOVA and finally computing χ^2 ratios using the sum of squares (see e.g. Winer *et al.* 1991):

$$H = SS_A / MS_{total} \tag{6}$$

where $SS_A = \sum n_i (\bar{R}_{i.} - (N+1)/2)^2$ is the sum of squares related to A, MS_{total} the variance, both based on the ranks R_{im} of x, and $\bar{R}_{i.} = R_{i.}/n_i$. If additionally the ranks R_{im} are transformed into normal scores

$$z_{im} = \phi^{-1}(R_{im}/(N+1)) \tag{7}$$

before applying the parametric ANOVA, then $SS_A = \sum n_i (\bar{z_i}. - \bar{z})^2$ and above H coincides with the statistic W of the van der Waerden test (van der Waerden 1953).

Now let x_{jm} (j=1,..,J and m=1,..,N) be the observations in a 1-way layout with J repeated measurements of factor B, R_{im} the rank of x_{jm} within subject m, so-called Friedman ranks, and $R_{j.} = \sum_m^N R_{jm}$. Then the Friedman statistic

$$FR = \frac{12}{NJ(J+1)} \sum_j^J R_{j.}^2 - 3N(J+1) \tag{8}$$

tests the identity of the J distribution functions, where FR is χ^2-distributed with (J-1) df. The derivation of the Friedman test assumes that the raw scores have equal variances and covariances (Lehmann 1975). It is well-known that, similar to the H test, Friedman's test

can be performed by ranking x (using midranks for ties) within each subject, conducting a parametric repeated measurements ANOVA and finally computing χ^2 ratios using the sum of squares (see e.g. Winer *et al.* 1991):

$$FR = SS_B/MS_{within} \tag{9}$$

where $SS_B = \sum n(\bar{R}_{j.} - (J+1)/2)^2$ is the sum of squares related to B, MS_{within} the within subject mean squares of R_{jm} , the Friedman ranks of x, and $\bar{R}_j = R_{j.}/J$. If additionally the ranks R_{jm} are transformed into normal scores for each subject m

$$z_{jm} = \phi^{-1}(R_{jm}/(J+1)) \tag{10}$$

before applying the parametric ANOVA, then $SS_B = \sum n(\bar{z}_{j.} - \bar{z})^2$ and above FR coincides with the statistic W of the van der Waerden test for dependent samples (Marascuilo and McSweeney 1977).

Now the KWF procedure shall be presented. Let s_{im} denote the sum of the J observed values of y for subject m in group i:

$$s_{jm} = \sum_j Jy_ijm \tag{11}$$

and $R_A(s_{im})$ their ranks ($i = 1, .., I$ and $m = 1, .., n_i$), corresponding to Wilcoxon ranks. For subject m in group i $R_B(y_{ijm})$ (j=1,...,J) shall be the ranks of the J observed values y_{ijm}, corresponding here to Friedman ranks. Then

$$r_{ijm} = (R_A(s_{im}) - 1)J + R_B(y_{ijm}) \tag{12}$$

comprehends a transformation of y_{ijm} into ranks 1,...,JN. Computing the sum of squares for a parametric ANOVA with r_{ijm} as described in (2) and (3), then

$$\chi^2 = SS_A/MS_{between}$$

is identical to H in (5) and (6) and the test of main effect A (H_A) in (4),

$$\chi^2 = SS_B/MS_{within}$$

is identical to FR in (8) and (9) and the test of main effect B (H_B) in (4), and

$$\chi^2 = SS_{AB}/MS_{within}$$

is the test of the interaction effect AB (H_AB) in (4), with $SS_A, SS_B, MS_{between}$ and MS_{within} as described in (2). Proofs are in section 3. Therefore the ANOVA using the ranks in (12) is a generalization of both the Kruskal-Wallis and the Friedman test and can be extended to designs with more than one between subject or within subject factors.

2.9. The van der Waerden procedure

Similar to the KWF procedure, a generalization of the van der Waerden method can be obtained. Let $N_A(s_{im})$ be the normal scores of s_{im} (see (11)), and $N_B(y_{ijm})$ the normal scores of y_{ijm} for subject m in group i ($j = 1, ..., J$), and

$$z_{ijm} = N_A(s_{im}) + N_B(y_{ijm}) \tag{13}$$

the combined scores. If a parametric ANOVA is performed with z_{ijm} , then

$$\chi^2 = SS_A/MS_{between}$$

is the test of main effect A (H_A) in (4) and identical to the van der Waerden test for independent samples,

$$\chi^2 = SS_B/MS_{within}$$

is the test of main effect B (H_B) in (4) and identical to the van der Waerden test for dependent samples, and

$$\chi^2 = SS_{AB}/MS_{within}$$

is the test of the interaction effect AB (H_{AB}) in (4), with $SS_A, SS_B, MS_{between}$ and MS_{within} as described in (2) and (3). However under two restrictions: first $R_A(s_{im})$, the ranks of the sums s_{im} , have no ties, and second, for each subject $m(m = 1, .., n_i \text{ and } i = 1, .., I)$ there are no ties within $R_B(y_{ijm})$, the ranks of y_{ijm}. Proofs are in section 3, where it is shown that in fact the tests of both factors A and B are affected by ties, but in a neglectable magnitude. Therefore the ANOVA using the scores in (13) can be seen as a generalization of the van der Waerden test. Finally it should be noted that Lu and Smith (1979), who investigated the distribution of the normal scores, recommend to apply the parametric F tests directly to the normal scores without computing the above described χ^2 ratios, simply because the χ^2 tests are too inaccurate for small sample sizes.

It is easy to see, that both, the KWF and the vdWS procedures, are applicable in factorial designs, also for more than one repeated measures factor.

2.10. Some notes on the selected methods

The basis for a comparison of the procedures is the parametric F test. To consider nonspherical covariance matrices, additionally the Huynh-Feldt adjustment of the degrees of freedom for the F test is chosen together with the correction by Lecoutre. Furthermore those methods have been preferred which are well known and applicable in standard software. At first there is the INT method which is for the reasons mentioned above preferred to the RT. As, on the other hand, often the ART technique is chosen instead of the RT method to prevent confounding effects, this one is included, too, but also in conjunction with the INT transformation. All these methods apply the F test and hence assume sphericity of the covariance matrix. The Puri & Sen test as well as the version based on normal scores PS INT have been taken into consideration, because there is only a minor difference to the proposed KWF method in view of the ranking. Thompson (1991) showed that under the less stringent condition of equal correlations of the J repeated measurement variables both procedures lead to valid χ^2 tests. The less known procedure by Koch seems to be attractive, because, first, it does not assume sphericity, and secondly, there are a few studies, amongst others by Tandon and Moeschberger (1989) and Ernst and Kepner (1993), who attested this one a good performance in terms of both type I error control and power. Also the ATS does not require sphericity of the covariance matrix (see e.g. Bathke *et al.* 2009). As this method is based on a completely nonparametric model, it is perfectly suited for this comparison. As the proposed KWF and vdWS methods are both based on the Friedman test, they have of course the same assumption: equal variances of the J repeated measurement variables and equal covariances, as mentioned above. But Harwell and Serlin (1994) attested the Friedman test a strong robustness against this assumption, which therefore should also apply to the KWF and vdWS methods.

A few words to the differences between the proposed KWF and vdWS methods and some of the other procedures. While the PS and PS INT methods use overall Wilcoxon ranks, computed over all JN y values, the KWF and vdWS methods apply Wilcoxon ranks only for ranking cases, but Friedman ranks for ranking the repeated measurements within a case. The difference between the PS and the KWF methods on one side and the PS INT and vdWS methods on the other side is the transformation of the ranks into normal scores before applying the χ^2 tests as demonstrated in the definition of the KWF and vdWS methods above.

2.11. Methods not selected

As this study has not been conceived as a general comparison of ANOVA methods for split-plot designs, a number of well-known methods have not been considered, among them the

GLM techniques (GEE and GLMM), which both performed unsatisfactorily in a similar study (Luepsen 2021), and the multivariate methods, which are restricted to the tests of the repeated measurements effects: e.g. the tests by Hotelling-Lawley, Pillai, Wilks and the nonparametric equivalent by Agresti & Pendergast (see e.g. Beasley and Zumbo 2009), for an extension to split-plot designs).

3. Proofs

First it will be proved by means of algebraic transformations that the proposed KWF and the vdWS procedures coincide with the K-W and the Friedman test, respectively the 1-factorial van der Waerden tests for the tests of the main effects. As ties may lead to small differences between the exact 1-factorial van der Waerden and the scores test, a simulation study is performed to give an impression of the magnitude of the deviation.

3.1. KWF method

First the proof that both the Kruskal-Wallis test, applied in a 2-factorial split-plot design, and the KWF test return identical results for factor A. As remarked (see (6)) both methods use the ratio of SS_A and MS_{total}. Therefore it is sufficient to show that SS_A and MS_{total} differ for both tests only by the same factor.

To use the Kruskal-Wallis test for the test of factor A, first for each subject m the sum s_{im} (see 11) is to be computed, before the test is applied on the s_{im}. Therefore SS_A and the variance are calculated for $R_A(s_{im})$ according to (6) noting that their mean rank is $(N+1)/2$:

$$SS_A = \sum_i n_i(\bar{R_A}(s_{i.}) - (N+1)/2)^2$$

where $\bar{R_A}(s_{i.})$ is the mean rank of R_A in group i. On the other hand, when performing a split-plot ANOVA on the combined ranks r_{ijm} (12) and computing SS_A, according to (2) these ranks are averaged over the J measurements:

$$\begin{aligned} t_{im} := \frac{1}{J}\sum_j^J r_{ijm} \quad &= \quad \frac{1}{J}\sum_j^J[(R_A(s_{im})-1)J + R_B(y_{ijm})] \\ &= \frac{1}{J}[J(R_A(s_{im})-1)J + (J+1)J/2] \\ &= J(R_A(s_{im})-1) + (J+1)/2 \\ &= JR_A(s_{im}) - (J-1)/2 \end{aligned}$$

remembering that $R_B(y_{ijm})$ has the values $1,..,J (i = 1,..,I$ and $m = 1,..,n_i)$ and therefore their sum is $(J+1)J/2$. Thus according to (2), using the definition of t_{im} and remembering that the mean rank of $R_A(s_i)$ is $(N+1)/2$

$$\begin{aligned} SS_A &= J\sum_i^I n_i(\bar{t_{i.}} - \bar{t})^2 \\ &= J\sum_i^I n_i[\bar{R_A}(s_{i.}) - (J-1)/2) - (J(N+1)/2 - (J-1)/2)] \\ &= J^3\sum_i^I(\bar{R_A}(s_{i.}) - (N+1)/2) \end{aligned}$$

which is identical to the SS_A above, differing only by the factor J^3. Finally, it is obvious that the denominators of the resulting χ^2 ratios, the variance of $R_A(s_{im})$ and $MS_{between}$ (see (3)), are the same, in this case also differing by the factor J^3.

Next to the proof that both the Friedman ANOVA and the KWF test for factor B yield the same result. To use the Friedman test for the effect of factor B, according to (9) it will do, to compute SS_B for $R_B(y_{ijm})$, the ranks of y_{ijm}, and MS_{within} e.g. by means of a repeated measures ANOVA:

$$SS_B = \sum_j^J N[\bar{R_B}(y_{.j.}) - (J+1)/2]^2$$

where $\bar{R_B}(y_{.j.})$ is the mean rank for level j of factor B.

Now to the split-plot ANOVA on the combined ranks r_{ijm} (12) and computing SS_B according to (2), noting the average of r_{ijm} is $(NJ+1)/2$ and $\bar{R_A}(s_{i.}) = (N+1)/2$:

$$\begin{aligned} SS_B &= N\sum_j^J[[\bar{r}_{.j.} - (NJ+1)/2]^2 \\ &= N\sum_j^J[J(\bar{R_A}(s_{i.}) - 1) + \bar{R_B}(y_{.j.}) - (NJ+1)/2]^2 \\ &= N\sum_j^J[J((N+1)/2 - 1) + \bar{R_B}(y_{.j.}) - (NJ+1)/2]^2 \\ &= N\sum_j^J[\bar{R_B}(y_{.j.}) - (J+1)/2]^2 \end{aligned}$$

which is identical to the SS_B above. Similarly it can be shown that MS_{within} are the same in either computational method.

3.2. van der Waerden method

The argumentation is similar to the proof above. To apply the van der Waerden test for factor A, here too, for each subject m the sum s_{im} of $y_{i1m}, ..., y_{ijm}$ is computed before transforming them into normal scores $N_A(s_{im})$ for $i = 1, .., I \text{and} m = 1, .., n_i$. Here two assumptions have to be made. At first, there are no ties within $R_A(s_{im})$, thus $R_A(s_{im})$ is a permutation of 1,..,N, and hence the $N_A(s_{im})$ symmetric around 0. Therefore the mean of the $N_A(s_{im})$ is 0. According to (7) it will do, to compute SS_A and the variance for $N_A(s_{im})$:

$$SS_A = \sum_i^I n_i[\bar{N_A}(s_{i.})]^2$$

On the other hand, when performing a split-plot ANOVA on the scores z_{ijm} (13) and computing SS_A, these values have to be averaged over the J measurements (see (2)):

$$\begin{aligned} t_{im} := \frac{1}{J}\sum_j^J z_{ijm} \quad &= \quad \frac{1}{J}\sum_j^J[(N_A(s_{im}) - 1)J + N_B(y_{ijm})] \\ = N_A(s_{im}) + \frac{1}{J}\sum_j^J N_B(y_{ijm}) \end{aligned}$$

Now, the second assumption: y has no ties, i.e. for each subject m the ranks of y_{ijm} are exactly $1, \ldots, J$, hence the sum of the $N_B(y_{ijm})$ is 0 and independent of i and m. In this case $t_{im} = N_A(s_{im})$. Therefore $\bar{t} = 0$, and hence

$$SS_A = J\sum_i^I n_i(\bar{t_{i.}} - \bar{t})^2 = J\sum_i^I n_i(\bar{N_A}(s_{i.}) - 0)^2 \tag{14}$$

and identical to the above SS_A disregarding the factor J. Finally, it is obvious that the denominators of the resulting χ^2 ratios, the variance of $R_A(s_{im})$ and$MS_{between}$ (see (3)), are the same, in this case differing by the factor J.

Next the application of the van der Waerden test for factor B. Here too, it will be assumed that there are no ties within $R_B(y_{i.m})$ for $i = 1, .., I$ and $m = 1, .., n_i$, thus $R_B(y_{i.m})$ is a permutation of 1,..,J, and the $N_B(y_{i.m})$ symmetric around 0. Therefore the means of $N_B(y_{i.m})$ are 0. According to (9) and (10) it will do to compute SS_B for $N_B(y_{ijm})$, the normal scores of y_{ijm} , and MS_{within} , e.g. by means of a repeated measures ANOVA with $\bar{z}$ being the overall mean of z_{ijm}:

$$SS_B = \sum_{j}^{J} N[\bar{N_B}(y_{.j.}) - \bar{z}]^2$$

Now to the split-plot ANOVA on the scores z_{ijm} and computing SS_B. As under the above stated assumption of no ties the following equation holds: $\bar{N_{A(}}s_{i.}) = 0$, as previously noted:

$$\begin{aligned} SS_B &= N \sum_{j}^{J} (\bar{z_{.j.}} - \bar{z})^2 \\ &= N \sum_{j}^{J} [\bar{N_A}(s_{i.}) + \bar{N_B}(y_{.j.}) - \bar{z}]^2 \\ &= N \sum_{j}^{J} [\bar{N_B}(y_{.j.}) - \bar{z}]^2 \end{aligned}$$

Similarly it can be shown that MS_{within} are the same in either computational method.

3.3. van der Waerden method: difference between the exact 1-factorial and the scores test

As mentioned in section 2 and shown above, the proof, that the tests of the main effects using the van der Waerden scores and the exact test are the same, is based on a restriction concerning tied ranks. Therefore it is essential to check the difference of the two test results for both main effects, for those cases when the assumption is not fulfilled. This is done by means of a Monte Carlo simulation considering the following influencing factors: distribution of y (2, 4 or 6 integer values with 2 different distribution shapes), factor A (2, 4 or 5 groups with a total of 20, 40 or 100 subjects), factor B (2, 4 or 6 repeated measurements) and design (equal or unequal sample sizes of the groups of factor A), making all in all 108 different conditions, which were replicated 100 times each. All generated data sets fulfilled the desired condition of tied ranks. Of special interest were those, which resulted in significant p values for either factor A or B. d_A and d_B shall denote the differences between p-values from the exact and the scores test for factor A respectively factor B.

Table 1: Statistics of d_A and d_B for several restricted regions of p

	factor A			factor B		
region	minimum	median	maximum	minimum	median	maximum
0 <p <1.0	- 0.2642	- 0.0021	0.0359	- 0.0049	0.0019	0.0142
0 <p <0.1	- 0.0200	0.0001	0.0055	- 0.0048	0.0019	0.0141
0 <p <0.01	- 0.0021	0.0000	0.0003	- 0.0040	- 0.0015	- 0.0002

The most important result: in the critical region $0<p<0.1$, where most decisions are made, the differences d_A and d_B are approximately normally distributed with mean near 0, and lie in the range [- 0.0021, 0.0141] (see table 1). The histograms (figure 1) and the boxplots (figure 2 and 3) indicate that for factor A extreme differences have a slight tendency towards negative

values, i.e. the p values from the scores test are sometimes larger than the exact values. The distribution and the design show no impact on the differences (figure 3). But apparently the size of the design has an effect (figure 2). Concerning d_A the number of negative outliers increases with rising sample sizes, while mean and dispersion stay unchanged, which means that for larger samples the p values for the test of A from the scores test are sometimes larger than the exact values, whereas for d_B mean and dispersion decrease with rising sample sizes, which means that for smaller samples the p values for the test of B from the scores tend to be smaller than the exact values. The number of repeated measurements shows no remarkable tendencies.

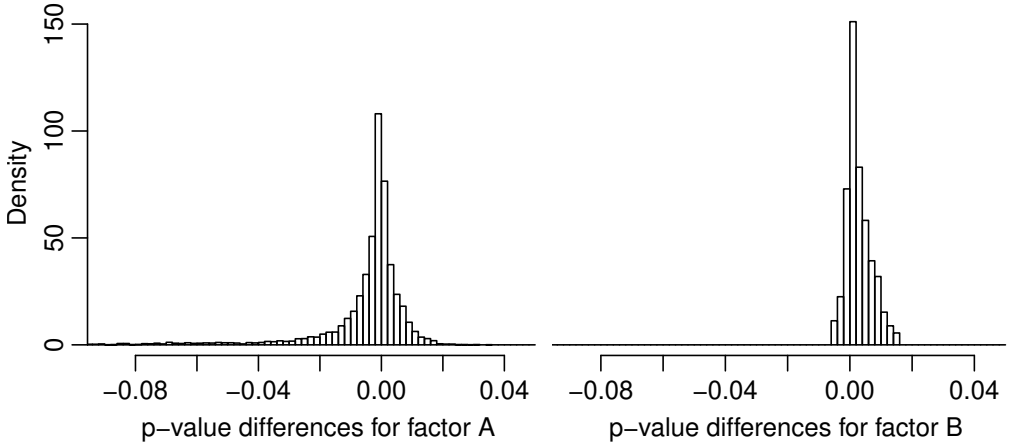

Figure 1: Histograms of d_A (left) and d_B (right)

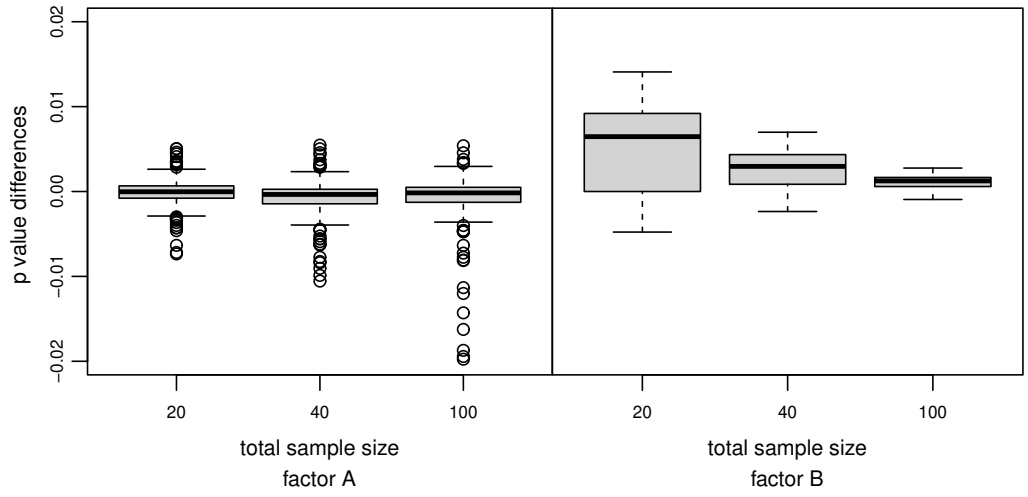

Figure 2: Boxplots of d_A (left) and d_B (right) in regard to the total sample size, restricted to the region $0<p<0.1$

4. A Monte Carlo simulation

The aim of this simulation study is to demonstrate the reliability of both proposed methods. These are compared with the parametric F test, including the Huynh-Feldt adjustment, and the rank based methods quoted in section 2. For this reason the type I error rates at 5% and the power, both as percentages of rejected null hypotheses, are investigated by means of a

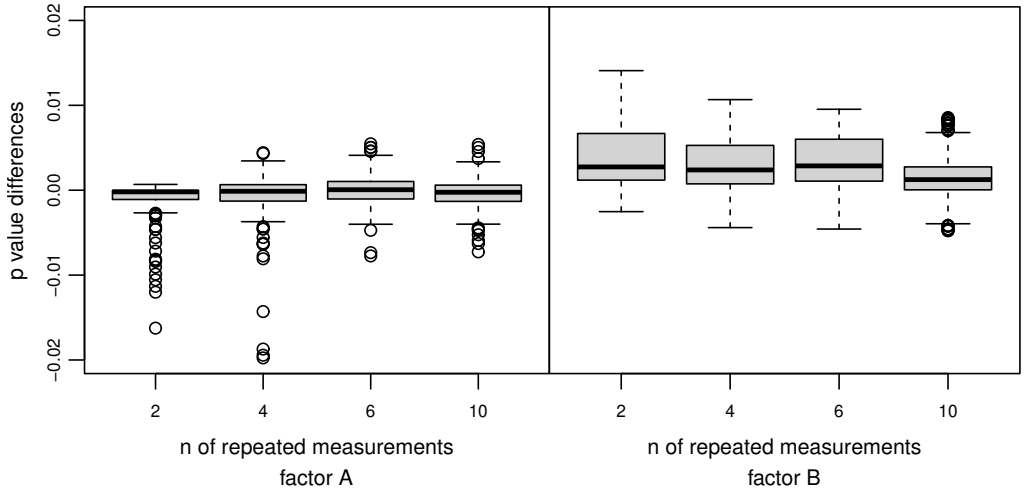

Figure 3: : Boxplots of d_A (left) and d_B (right) in regard to the number of repeated measurements, restricted to the region 0<p<0.1

Monte Carlo study, where several settings of a split-plot ANOVA layout are varied.

4.1. Methodology

The parameters of this study are the size (number of cells), cell frequencies (equal, unequal), cell counts (5,10,...,50), pairing of n_i and s_i^2 (independent, positive, negative), model (effect of factors and interaction), as well as correlation structure (equal or unequal correlations). This should cover all important situations. One remark to the pairing problem: the parametric F test tends to be conservative, if cells with larger n_i have also larger variances s_i^2 (*positive* or *direct pairing*), and reacts liberal, if cells with larger n_i have the smaller variances (*negative* or *inverse pairing*), (see e.g. Feir and Toothaker 1974).The resulting sample sizes N vary from 10 to 1200. For each cell count and situation there are 2000 replications, so that there is a total of 20.000 replications for each scenario. Without loss of generality the layout will be restricted to two factors A and B, and for each factor only one vector of effect sizes has been chosen, which should suffice to see, if one factor has at all an impact on the results. The schema below gives an overview of the design of this study, where e.g. A(B) denotes the test of factor A with a nonnull effect of factor B.

Table 2: An overview of the design

size	design	cell counts	pairing $n_i s_i$	corr structure	model	distribution
small (3*3) large (4*6)	equal n_i unequal n_i	n_i=5,...,50	independent positive negative	equal r descending r	A(-), A(B), A(AB) B(-), B(A), B(AB) AB(-),AB(A),AB(B)	1. multivar. normal ..... 14. contaminated III

Four designs are analyzed:

- a 3*3 design ("small design"), one with equal cell counts (balanced), and one with unequal cell counts having a ratio $max(n_i)/min(n_i)$ of 3.5 (unbalanced), and
- a 4*6 design ("large design"), one with equal cell counts (balanced), and one with unequal cell counts having a ratio $max(n_i)/min(n_i)$ of 4 (unbalanced),

14 different models of multivariate distributions with mean vectors μ_i and covariance matrices $\Sigma^{(i)}$ (i=1,..,I) have been chosen. The variances will be denoted by $s_{jj}^{(i)}$, the correlations by

$r^{(i)}_{j_1 j_2}$ $(j, j_1, j_2 = 1, ..., J)$.
The following two correlation structures are used, which are assumed equal for all groups:

- exchangeable (equal covariances, compound symmetry) with $r^{(i)}_{j_1 j_2} = 0.3$, a value that seems realistic and had often been chosen (see e.g. Emrich and R. 1992), and
- descending correlations $r^{(i)}_{j_1 j_2}$= (0.7, 0.5, 0.4, 0.2, 0.1) for large designs, respectively $r^{(i)}_{j_1 j_2}$= (0.7, 0.5) for small designs, which is similar to the AR(1) structure (unequal covariances, no sphericity).

The following distributions have been selected:

1. multivariate normal with equal variances $s^{(i)}_{jj}$ $(i = 1, .., I)$,
2. multivariate normal with unequal variances and $\max(s^{(i)}_{jj})/\min(s^{(i)}_{jj}) = 4$ $(i = 1, .., I)$ on factor A (n_i and $s^{(i)}_{jj}$ independent, correlation n_i with $s^{(i)}_{jj}$ r=0.07), denoted by *V(A)*
3. multivariate normal with unequal variances and $\max(s^{(i)}_{jj})/\min(s^{(i)}_{jj}) = 4$ $(j = 1, .., J)$ on factor B, denoted by *V(B)*
4. multivariate normal with unequal variances and $\max(s^{(i)}_{jj})/\min(s^{(i)}_{jj}) = 4$ $(i = 1, .., I$ and $j = 1, ., J)$ on both factors, (correlation n_i with$s^{(i)}_{jj}$ r=0.07), denoted by *V(A,B)*
5. multivariate normal with unequal variances and $\max(s^{(i)}_{jj})/\min(s^{(i)}_{jj}) = 4$ $(i = 1, .., I)$ on factor A, where small n_i correspond to small variances $s^{(i)}_{jj}$ (*positive pairing*, r=0.98).
6. multivariate normal with unequal variances and $\max(s^{(i)}_{jj})/\min(s^{(i)}_{jj}) = 4$ $(i = 1, .., I)$ on factor A, where small n_i correspond to large variances $s^{(i)}_{jj}$ (*negative pairing*, r=-0.87).
7. multivariate exponential ($\lambda = 0.4$) with $\mu = 2.5$, which is highly skewed (skewness=2),
8. multivariate exponential ($\lambda = 0.4$) with $\mu = 2.5$, rounded to integer values 1,2,..,
9. multivariate uniform in the interval [0,5],
10. multivariate uniform in the interval [0,5], rounded to integer values 1,..,5,
11. multivariate lognormal ($\mu = 0$ and $s^{(i)}_{jj}$=0.25) which is slightly skewed (skewness=0.778),
12. multivariate normal $N(\mu_i, \Sigma^{(i)})$ with equal variances where at random 20% of the N subjects were contaminated with $N(\mu_i, 4\Sigma^{(i)})$ (only some subjects are affected, but potentially all values of that subject), denoted by *contaminated I*,
13. multivariate normal $N(\mu_i, \Sigma^{(i)})$ with equal variances where at random 30% of all JN subjects were contaminated with $N(\mu_i, 4\Sigma^{(i)})$ (all values are incidentally affected, and potentially any subject with any measurement), denoted by *contaminated II*,
14. multivariate normal $N(\mu_i, \Sigma^{(i)})$ with equal variances where at random 15% of all JN subjects were contaminated with $N(\mu_i, 4\Sigma^{(i)})$ (all values are incidentally affected, but outliers are only on the right side), denoted by *contaminated III*,

Skewed distributions with unequal variances are not considered, because a number of studies show that nonparametric procedures cannot handle skewed distributions in the case of heteroscedasticity (see e.g. Vallejo, Ato, and Fernandez 2010; Keselman, Carriere, and Lix 1995; Tomarken and Serlin 1986). A more precise investigation of the error rates for rank

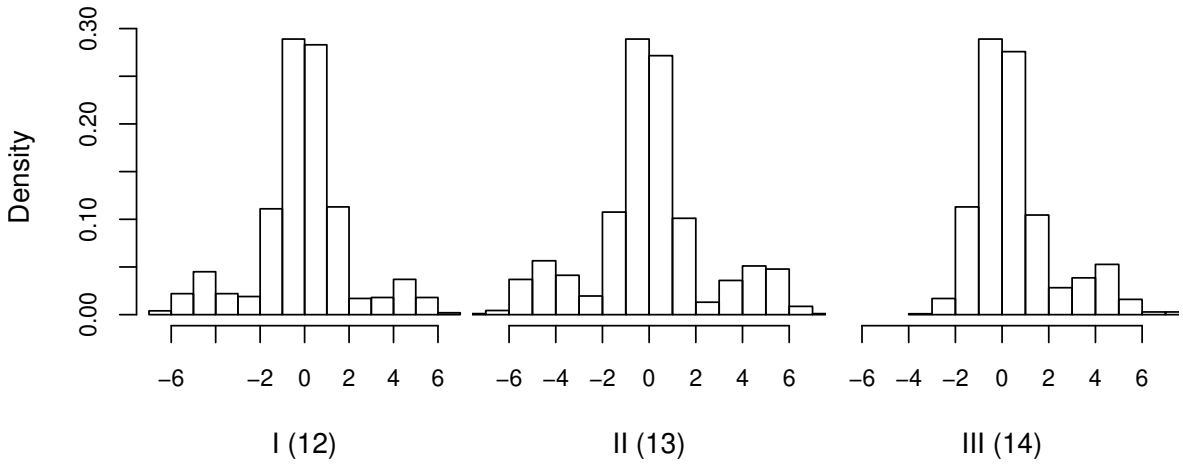

Figure 4: Histograms of the contaminated normal distributions I, II and III (no 12-14 in the list above)

based procedures applied on the lognormal distribution has been done by Luepsen (2016), who confirmed earlier results by Carletti and Claustriaux (2005).

The type I error rates of all main and interaction effects are checked in the range of n=5,...,50 for the case of the null model, the case of one significant main effect $A_{0.5}$ or $B_{0.5}$, and the case of a significant interaction $AB_{0.5}$. Here e.g. A_d denotes an effect of size d for factor A, corresponding to effect vectors $(-sd/2, 0, ..., 0, sd/2)$ with the standard deviation s of y. Analog definitions for B_d and AB_d. To eliminate spurious oscillations within the range of n=5,...,50 the type I error rates are smoothed by means of simple two-sided moving averages. The power is computed in relation to n=5,...,50, but for slightly different effect sizes: $A_{0.4}$, $B_{0.3}$ and $AB_{0.5}$, because the power rates come to close to 100 % for larger d.

For unbalanced designs the interaction effects $(ab)_{ij}$ had to be adjusted respecting the different cell counts, in order to avoid impacts on the main effects. It should be remarked that most ANOVA procedures are based upon LS estimation, which corresponds to weighted means analysis, where the cell counts n_i have a larger impact on the results than with the unweighted means analysis. The latter assumes equal cell counts by design and allows only a couple of missing observations (see e.g. Winer *et al.* 1991). Unfortunately the ATS method for split-plot designs, as implemented in the R package `nparLD`, is based on the unweighted means analysis (see e.g. Noguchi *et al.* 2012), which may lead to results, which are not comparable with those from the other analyses. For the ATS method also unadjusted rates are added in the tables. All results report the percentage of rejections of the corresponding null hypothesis. Since only the correct adherence to the α-level for given effects is of interest here, the slight differences in the null hypotheses of the individual nonparametric procedures are irrelevant.

4.2. Criteria

A deviation of 25 percent $(\alpha + 0.25\alpha)$ - that is 6.25 percent for $\alpha = 0.05$ - can be regarded as a moderate definition of robustness (see Peterson 2002), whereas 50 percent $(\alpha + 0.50\alpha)$ - that is 7.50 percent for $\alpha = 0.05$ - will be treated as liberal robustness, according to Bradleys liberal criterion (see Bradley 1978), which is often used in other studies. As a large amount of the results concerns the error rates for 10 sample sizes $n_i = 5, ..., 50$, it seems reasonable to allow a couple of exceedances within this range.

4.3. Tables and graphical illustrations

The results quoted in the next part of this section represent only a small extract from the

numerous tables and graphics produced in this study and will concentrate on essential results. All tables and corresponding graphical illustrations report the percentage of rejections of the corresponding null hypothesis, for different models and $n_i = 5, 10, .., 50$, small and large designs as well as balanced and unbalanced designs. They are divided into the chapters C1 for type I error rates at $\alpha = 0.05$ for fixed n_i and C2 for the power in relation to n_i. Both are available online on the server of the university of Cologne under http://www.uni-koeln.de/∼ luepsen/statistik/texte/comparison-tables/. An extract of the type I error rates is given in the appendix: for the interaction AB in case of the null model and the case with one nonnull main effect in tables 2-4, for factor B in table 5, and for the case of pairing in table 6. Comparisons of the power are visualized in figures 6 and 7.

4.4. Type I error rates

The type I error rates of the two tests proposed are nearly identical for all distributions, situations and effects considered. Overall they stay in the range of moderate robustness, for the interaction AB as well as for both main effects. There are only two exceptions in the case of heterogeneous variances of factor A: first, the test of the interaction if B has a nonnull effect, and secondly, the test of B if there is a nonnull interaction. In both situations the type I error rates rise up to 10 or more for increasing cell counts n_i depending on the degree of heterogeneity (see tables 4 and 5), a problem which Lei, Holt, and Beasley (2004) as well as Payton, Richter, Giles, and L. (2006) already pointed to. And the rates are even more upsetting for unequal correlations. On the other side, these two tests are able to control the type I error rates completely in many situations, where the parametric F test, and often also the H-F adjustment, fails with rates up to 10 and beyond. Here to mention:

- First the tests for factor B and the interaction AB, when there are heterogeneous variances (see figure 5 as well as tables 2, 3, 4 and 5), even though the Huynh-Feldt adjustment reduces the rates mostly to an acceptable range. Overall the KWF and the vdWS procedures are more robust regarding heterogeneous variances compared with the parametric F test, especially in situations where the error rates of the F test are oscillating around the upper limit of the interval of robustness, e.g. for the test of the interaction when A is nonnull (see table 3).

- Secondly the cases of negative and positive pairing: when large variances are paired with small sample sizes, both procedures control the error rates for the test of the interaction (see figure 5), except the situation mentioned above, if B is nonnull. And for the test of factor A the KWF method is able to damp the type I error rates down to values below 7 for moderate $n_i \leq 30$ (see table 6) and below 9.5 for larger cell counts. Also when large variances are paired with large sample sizes, the rates are under control in most instances, so that the power of both methods is the best. The type I error rates are displayed in table 6, where for comparative purposes also those of the ATS are shown, because this is often described as retaining the error level close to the nominal level in all situations (see e.g. Noguchi *et al.* 2012; Luepsen 2018). Perhaps to complete: the pairing problem has generally no impact on the test of B. Compared with the F test, including the H-F adjustment, and the ATS method, the KWF procedure seems to be best in this scenario.

- Thirdly the tests for factor B, when the underlying distribution is normal but contaminated (see e.g. table 5). Here the H-F adjustment is unable to reduce the error rates.

Power

Concerning the power, the F test including the H-F adjustment performs better in many situations for all three effects, with rates about 10-20% above those of the KWF and the vdWS procedures, especially for small n_i and for an underlying exponential distribution (see

figure 6) which had to be expected from the studies by Feir and Toothaker (1974) and Ellis and Haase (1994). But these two procedures can keep up with the F test in most cases of heterogeneous variances and perform often even better (see e.g. figure 6). On the other side it will be not surprising that the KWF and the vdWS methods show higher power rates than the F test in most cases of contaminated normal distributions (see figure 7). For factor A the two tests proposed cannot outperform the F test, but achieve at least the same level of power. But things look better for factor B and the interaction. In fact the F test performs slightly better than the KWF and the vdWS methods for small $n_i \leq 20$ in some instances, but generally these two procedures belong to the best performing, with rates between 30% and 100% above those of the F test. And for unequal correlations their performance is even better with rates between 80% and 150% above those of the F test, and clearly better than all other procedures (see figure 7). In fact the PS INT and the INT procedures score slightly better for small designs, but their insufficient type I error control has to be taken into account. Finally one more positive feature of the KWF and the vdWS methods: in the case of positive pairing they are the best performing methods, particularly for the interaction with rates about 20-50% above those of other tests for small $n_i \leq 30$ (see e.g. figure 6). To make power differences better visible in the area of small n, where the graphs of the absolute power lie close together, a *relative power* is used. The *relative power* of a method is computed as its (absolute) power divided by the 25% trimmed mean of the power of all 8 methods in percent.

4.5. Comparisons

How do the two procedures perform compared with other rank based ANOVA procedures? A stable method, particularly in regard of the type I error control, is the Hyunh-Feldt adjustment to the F test. It is the favorite for split-plot designs in many studies (see e.g. Stiger *et al.* 1998). But in this study the H-F adjustment cannot hold completely the error rates in the interval of moderate robustness, in contrary to the two procedures under consideration, e.g. in the cases of negative pairing (see table 6) and for contaminated normal distributions (see e.g. table 5). Furthermore, the H-F adjustment has also lower power rates in case of contaminated distributions. The best one from the other rank based procedures listed in section 2 seems to be Koch‘s method. It keeps the type I error rates completely in the interval of moderate robustness, even in the situations where the KWF and the vdWS methods fail. But unfortunately Koch‘s procedure has two problems: extreme reactions in the situations of positive and negative pairing, in contrary to the proposed methods, and the poor power of the tests for B and AB for $n_i \leq 20$. Else the normal scores method (INT): in nearly all situations with heteroscedasticity the error rates exceed the limits of robustness, particularly for the tests of B and AB with rates rising to 10 and beyond (see e.g. tables 2, 3 and 5), as well as for the case of negative pairing. Also the ART INT and both Puri & Sen procedures (PS and PS INT) have severe deficiencies concerning the type I error rates when there are heterogeneous variances, for the test of B (see e.g. table 5) as well as for the test of AB (see tables 2 and 3), showing rates up to 10 and more. With regard to the power, the PS INT procedure is generally able to keep up with the F test in many situations (see e.g. figures 6 and 7). Particularly for the test of the interaction its performance lies above average. In addition the ART INT method reacts on discrete outcomes with increasing error rates, particularly for the tests of the main effects (see e.g. table 5). This phenomenon has already been observed for between subject designs by Luepsen (2017). Finally, the ATS shows exceeding error rates generally for small and medium cell counts $n_i \leq 30$, starting often with rates about 15 (n_i=5) and 12 (n_i=10), particularly for the test of factor A, but sometimes also for AB and B (see e.g. tables 3, 5 and 6). This deficiency has been previously noted by Tian and Wilcox (2007). Furthermore to mention the deficient type I error control in the case of negative pairing (see table 6). In all of these situations both the KWF and the vdWS methods keep the type I error under complete control. Advantages of the other methods regarding the power are at the expense of their increased type I error rates.

Although both, the KWF and the vdWS procedures, show nearly an identical behaviour,

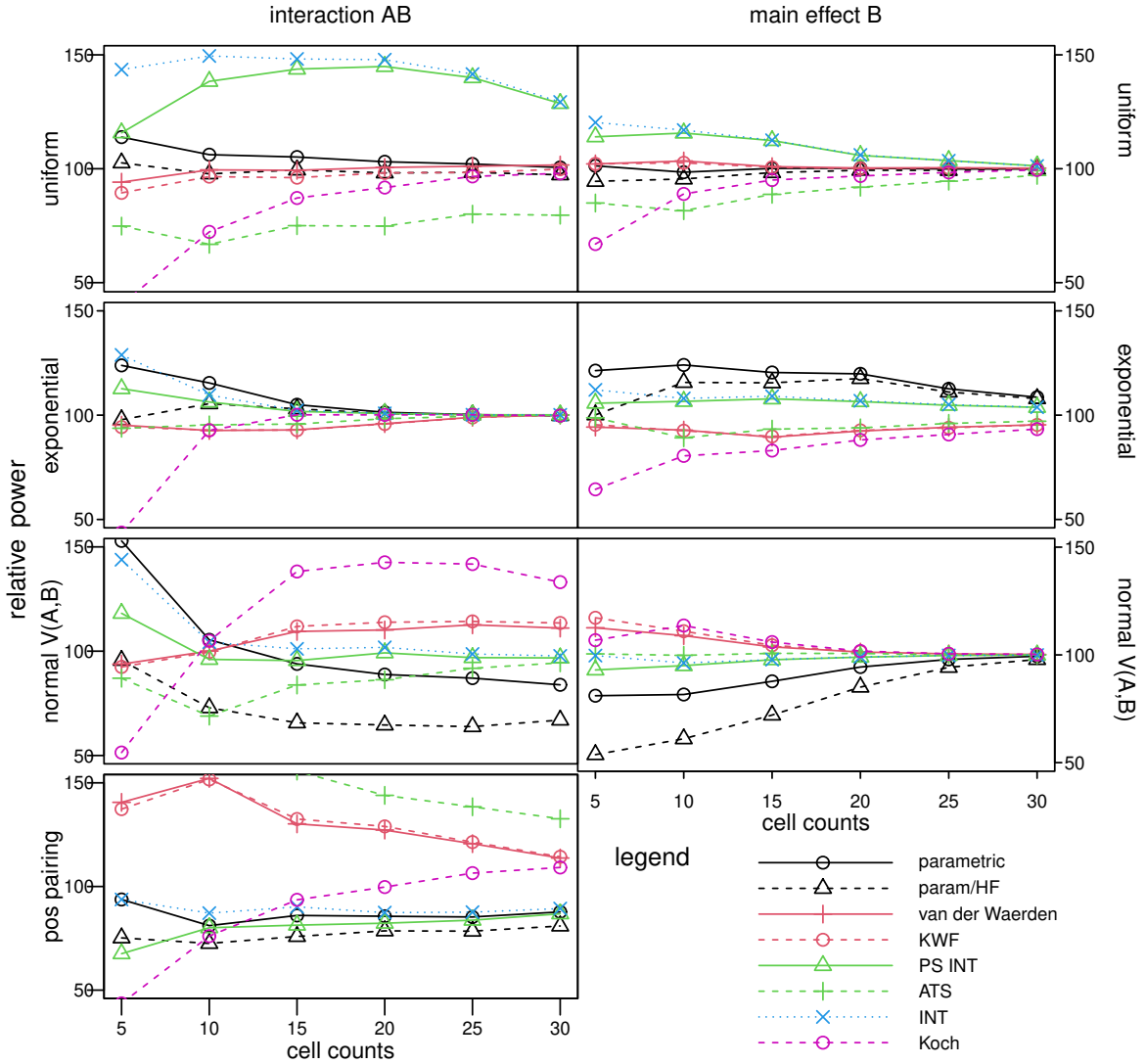

Figure 5: Relative powerin the range of n_i=5,..,30 for the interaction AB and main effect B (null model, balanced large design, unequal correlations) for 4 different distributions: uniform, exponential, normal with unequal variances on both factors, and for the case of positive pairing

there are some minor differences. Concerning the type I error the KWF method is able to control the situation of negative pairing better than the vdWS method. With regard to the power there are only very few situations where the vdWS outperforms the KWF method, e.g. for underlying uniform distributions, while both show equal rates else. Here, compared with the study for between subject designs (see e.g. Luepsen 2018), a generally better power performance of the van der Waerden test has been expected.

5. Conclusion

The study showed that both, the KWF and the vdWS procedures, have the same attributes, that were noted previously by several authors for 1-factorial analyses and mentioned in section 1: robustness in cases of departures from normality and homoscedasticity, but on the other side, in some situations the need for large samples to achieve a satisfying power. Similarly this is valid for the test of the interaction, also in the cases of a nonnull model. Both methods are able to score in cases of heterogeneous conditions such as positive and negative pairing, as well as in cases of nonnormal distributions such as contaminated normal distributions, where they belong to the best performing procedures in terms of power and type I error control. Especially the test of the interaction shows no conspicuous findings. On the other side, the tests of the effects in case of other nonnull effects revealed a precarious, but possibly rare problem, which remain covered in a 1-factorial analysis: the tests of B and the interaction are confounded in the case of heterogeneous variances of factor A. A comparison with other procedures showed that only the Huynh-Feldt adjustment of the F test and Koch‘s method are able to compete on the whole. The final deduction: the vdWS method, and even more the KWF method are reliable tools for analyzing data from split-plot designs for nearly all kinds of underlying distributions, as long as not the tests of both repeated measures effects show a remarkable impact. And both methods can be applied using standard statistical software providing familiar tests.

6. Programming

This study has been programmed in R (version 3.5.3). For the data generation the function `mvrnorm` from the package `MASS` (see e.g. Ripley 1987) has been used to receive multivariate normal distributed variates. Other multivariate distributions were obtained by suitable data transformations. Various functions have been chosen to analyze the simulated data: the function `aov` in combination with `drop1` (to receive type III sum of squares estimates in the case of unequal cell counts) for the standard ANOVA F-test, an own function `np.anova` for the factorial Puri & Sen-tests as well as for the KWF and the vdWS tests, and an own function `koch.anova` for Koch‘s nonparametric analysis of split-plot designs. Finally the function `nparLD` from the package `nparLD` has been applied for the ATS method. For the own functions see Luepsen (2014). Some of the computations have been performed on a Windows notebook, but for the major part the high performance cluster CHEOPS of the Regional Computing Centre (RRZK) of the university of Cologne has been used. I would like to thank the staff of the RRZK for their technical support as well as Prof. Unkelbach for his organizational support.

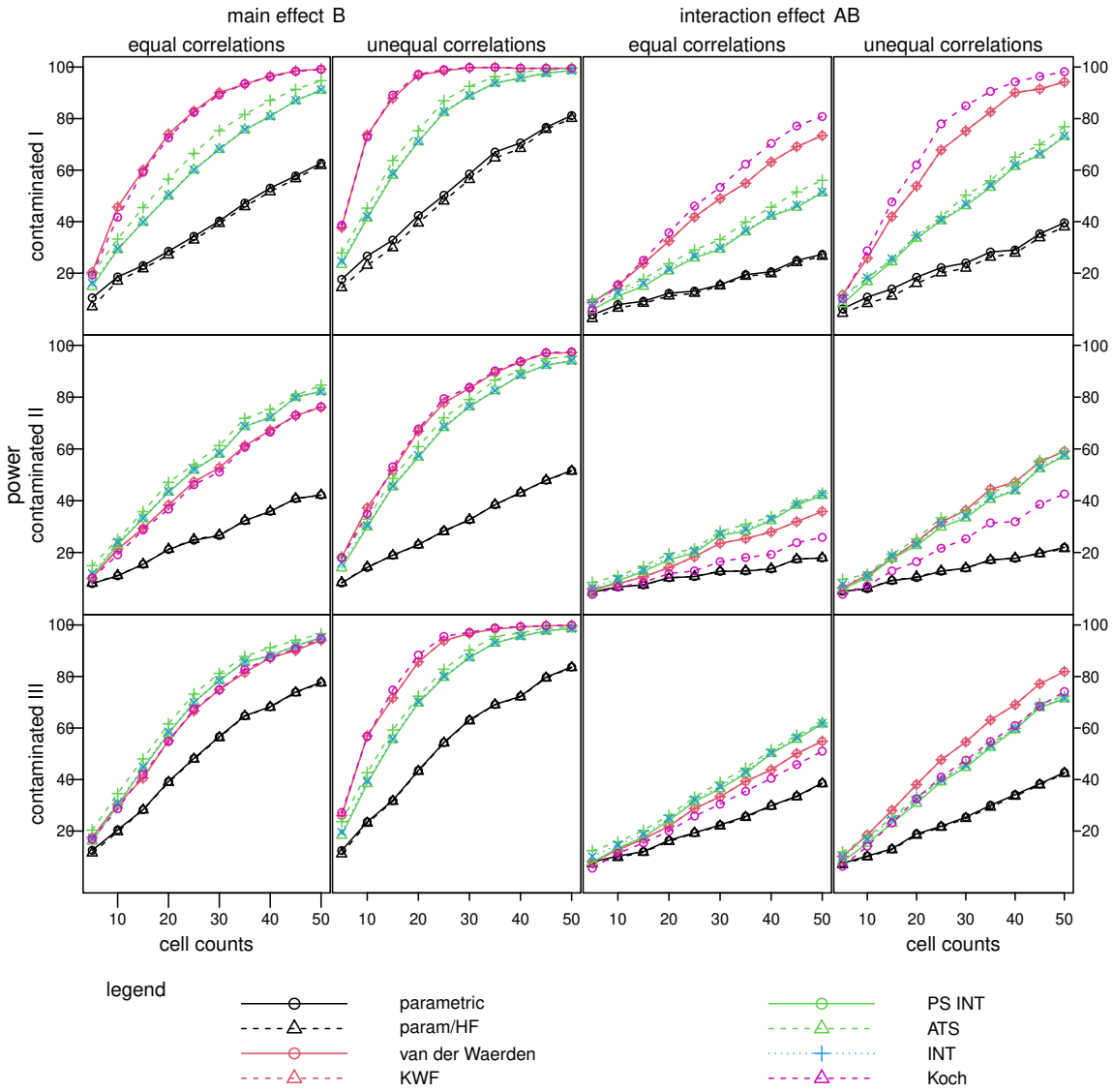

Figure 6: Power for the test of main effect B and the interaction for multivariate normal contaminated distributions I and II (unbalanced small design), for equal and unequal correlations

References

Algina J (1994). "Some Alternative Approximate Tests for a Split Plot Design." *Multivariate Behavioral Research*, **29**(4), 365–384. doi:10.1207/s15327906mbr2904_3.

Bathke AC, Schabenberger O, Tobias RD, Madden LV (2009). "Greenhouse–Geisser Adjustment and the ANOVA-Type Statistic: Cousins or Twins?" *The American Statistician*, **63**(3), 239–246. doi:10.1198/tast.2009.08187.

Beasley TM (2000). "Nonparametric Tests for Analyzing Interactions Among Intra-Block Ranks in Multiple Group Repeated Measures Designs." *Journal of Educational and Behavioral Statistics*, **25**(1), 20–59. doi:10.3102/10769986025001020.

Beasley TM, Zumbo BD (2009). "Aligned Rank Tests for Interactions in Split-Plot Designs: Distributional Assumptions and Stochastic Heterogeneity." *Journal of Modern Applied Statistical Methods*, **8**(1), 16–50. doi:10.22237/jmasm/1241136180.

Bradley JV (1978). "Robustness?" *British Journal of Mathematical and Statistical Psychology*, **31**, 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x.

Brunner E, Munzel U, Puri ML (1999). "Rank-Score Tests in Factorial Designs with Repeated Measures." *Journal of Multivariate Analysis*, **70**(4), 286–317. doi:10.1006/jmva.1999.1821.

Carletti I, Claustriaux JJ (2005). "Anova or Aligned Rank Transform Methods: Which One Use when Assumptions Are Not Fulfilled ?" *Buletinul USAMV-CN*, **62**, 1454–2382.

Chatterjee SK, Sen PK (1966). "Non-parametric Tests for the Multivariate Multisample Location Problem." *S. N. Roy Memorial Volume, edited by R.C. Bose, et al.*

Conover WJ, Iman RL (1981). "Rank Transformations as a Bridge between Parametric and Nonparametric Statistics." *American Statistician*, **35**(3), 124–129. doi:10.1080/00031305.1981.10479327.

Dijkstra JB (1987). "Analysis of Means in Some Non-standard Situations." *Technische Universiteit, Eindhoven*. doi:10.6100/IR272914.

Ellis VE, Haase RF (1994). "Rank Transformation and Puri-Sen Nonparametric Analysis of Split-Plot and Repeated Measures Designs." *Meeting of the American Statistical Association, Toronto, Ontario, Canada*.

Emrich LJ, R PM (1992). "On Some Small Sample Properties of Generalized Estimating Equation Estimates for Multivariate Dichotomous Outcomes." *Journal of Statistical Computation and Simulation*, **41**, 19–29. doi:10.1080/00949659208811388.

Ernst MD, Kepner Jl (1993). "A Monte Carlo Study of Rank Tests for Repeated Measures Designs." *Communications in Statistics - Simulation and Computation*, **22**(3), 671–678. doi:10.1080/03610919308813115.

Feir BJ, Toothaker LE (1974). "The ANOVA F-Test versus the Kruskal-Wallis Test: A Robustness Study." *Paper presented at the 59th Annual Meeting of the American Educational Research Association in Chicago, IL*.

Gibbons JD, Chakraborti S (2020). *Nonparametric Statistical Inference*. CRC press.

Harville DA (1977). "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems." *Journal of the American Statistical Association*, **72**(358), 320–338.

Harwell MR, Serlin RC (1989). "A Nonparametric Test Statistic for the General Linear Model." *Journal of Educational Statistics*, **14**(4), 351–371. doi:10.3102/10769986014004.

Harwell MR, Serlin RC (1994). "A Monte Carlo Study of the Friedman Test and Some Competitors in the Single Factor, Repeated Measures Design with Unequal Covariances." *Computational Statistics & Data Analysis*, **17**, 35–49. doi:10.1016/0167-9473(92)00060-5.

Headrick TC, Sawilowsky SS (2000). "Properties of the Rank Transformation in Factorial Analysis of Covariance." *Communications in Statistics - Simulation and Computation*, **29**(4), 1059–1088. doi:10.1080/03610910008813654.

Hodges JL, Lehmann EI (1962). "Rank Methods for Combination of Independent Experiments in Analysis of Variance." *Annals of Mathematical Statistics*, **33**, 482–497.

Hora SC, Conover WJ (1984). "The F Statistic in the Two-Way Layout with Rank-Score Transformed Data." *Journal of the American Statistical Association*, **79**(387), 668–673. doi:10.1080/01621459.1984.10478095.

Huang ML (2007). "A Quantile-Score Test for Experimental Design." *Applied Mathematical Sciences*, **1**(11), 507–516.

Keselman HJ, Algina J, Kowalchuk RK (2001). "The Analysis of Repeated Measures Designs: A Review." *British Journal of Mathematical and Statistical Psychology*, **54**, 1–20. doi:10.1348/000711001159357.

Keselman HJ, Carriere KC, Lix LM (1995). "Robust and Powerful Nonorthogonal Analyses." *Psychometrika*, **60**, 395–418. doi:10.1007/BF02294383.

Koch GG (1969). "Some Aspects of the Statistical Analysis of Split Plot Experiments in Completely Randomized Designs." *Journal of the American Statistical Association*, **64**(326), 485–504. doi:10.1080/01621459.1969.10500989.

Lehmann EL (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

Lei X, Holt JK, Beasley TM (2004). "Aligned Rank Tests as Robust Alternatives for Testing Interactions in Multiple Group Repeated Measures Designs with Heterogeneous Covariances." *Journal of Modern Applied Statistical Methods*, **3**(2), 17. doi:10.22237/jmasm/1099268220.

Liang KY, Zeger SL (1986). "A Comparison of Two Bias-Corrected Covariance Estimators for Generalized Estimating Equations." *Biometrika*, **73**, 13–22. doi:10.1111/j.1541-0420.2007.00764.x.

Lix LM, Keselman JC, Keselman HJ (1996). "Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test." *Review of Educational Research*, **66**(4), 579–619. doi:10.3102/0034654306600.

Lu HT, Smith PJ (1979). "Distribution of the Normal Scores Statistic for Nonparametric One-Way Analysis of Variance." *Journal of the American Statistical Association*, **74**(367), 715–722. doi:10.1080/01621459.1979.10481676.

Luepsen H (2014). "R Functions for the Analysis of Variance." URL http://www.uni-koeln.de/~luepsen/R/.

Luepsen H (2016). "The Lognormal Distribution and Nonparametric Anovas - A Dangerous Alliance." *Technical report*, University of Köln.

Luepsen H (2017). "The Aligned Rank Transform and Discrete Variables: A Warning." *Communications in Statistics - Simulation and Computation*, **46**(9), 6923–6936. doi:10.1080/03610918.2016.1217014.

Luepsen H (2018). "Comparison of Nonparametric Analysis of Variance Methods - A Vote for van der Waerden." *Communications in Statistics - Simulation and Computation*, **47**(9), 2547–2576. doi:10.1080/03610918.2017.1353613.

Luepsen H (2020). "Varianzanalysen - Prüfung der Voraussetzungen und Übersicht der Nichtparametrischen Methoden sowie Praktische Anwendungen mit R und SPSS." URL http://www.uni-koeln.de/~luepsen/statistik/texte/nonpar-anova.pdf.

Luepsen H (2021). "ANOVA with Binary Variables – the F Test and Some Alternatives." *Communications in Statistics - Simulation and Computation*. doi:10.1080/03610918.2020.1869983.

Mansouri H, Chang GH (1995). "A Comparative Study of Some Rank Tests for Interaction." *Computational Statistics and Data Analysis*, **19**, 85–96. doi:10.1016/0167-9473(93)E0045-6.

Marascuilo LA, McSweeney M (1977). *Nonparametric and Distributionfree Methods for the Social Sciences*. Brooks/Cole Pub. Co.

Noguchi K, Gel YR, 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), 1–23. doi:10.18637/jss.v050.i12.

Payton ME, Richter SJ, Giles K, L Royer TA (2006). "On Transformations of Count Data for Tests of Interaction in Factorial and Split-Plot Experiments." *Journal of Economic Entomology*, **99**(3), 1002–1006. doi:10.1093/jee/99.3.1002.

Peterson K (2002). "Six Modifications Of The Aligned Rank Transform Test For Interaction." *Journal Of Modern Applied Statistical Methods*, **1**(1), 100–109. doi:10.22237/jmasm/1020255240.

Puri ML, Sen PK (1969). "Analysis of Covariance based on General Rank Scores." *The Annals of Mathematical Statistics*, pp. 610–618.

Puri ML, Sen PK (1985). *Nonparametric Methods in General Linear Models*. Wiley, New York. doi:10.1515/9783110917819.

Quintana SM, Maxwell SE (1994). "A Monte Carlo Comparison of Seven ϵ-Adjustment Procedures in Repeated Measures Designs with Small Sample Sizes." *Journal of Educational Statistics*, **19**(1), 57–71. doi:10.3102/10769986019001057.

Ripley BD (1987). *Stochastic Simulation*. Wiley, New York.

Salter KC, Fawcett RF (1993). "The Art Test of Interaction: A Robust and Powerful Rank Test of Interaction in Factorial Models." *Communications in Statistics - Simulation and Computation*, **22**(1), 137–153. doi:10.1080/03610919308813085.

Sawilowsky SS (1990). "Nonparametric Tests of Interaction in Experimental Design." *Review of Educational Research*, **60**(1), 91–126. doi:10.3102/00346543060001091.

Sheskin DJ (2004). *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall. doi:10.1201/9781420036268.

Stiger RT, Kosinski AS, Barnhart HX, Kleinbaum DG (1998). "Anova for Repeated Ordinal Data with Small Sample Size? A Comparison of Anova, Manova, WLS and GEE Methods by Simulation." *Communications in Statistics - Simulation and Computation*, **27**(2), 357–375. doi:10.1080/03610919808813485.

Tandon PK, Moeschberger ML (1989). "Comparison of Nonparametric and Parametric Methods in Repeated Measures Designs - A Simulation Study." *Communications in Statistics - Simulation and Computation*, **18**(2), 777–792. doi:10.1080/03610918908812790.

Thompson GL (1991). "A Unified Approach to Rank Tests for Multivariate and Repeated Measures Designs." *Journal of the American Statistical Association*, **86**(414), 410–419.

Tian T, Wilcox RR (2007). "A Comparison of Two Rank Tests for Repeated Measure Designs." *Journal of Modern Applied Statistical Methods*, **6**(1), 331–335. doi:10.22237/jmasm/1177993800.

Tomarken AJ, Serlin RC (1986). "Comparison of ANOVA Alternatives under Variance Heterogeneity and Specific Noncentral Structures." *Psychological Bulletin*, **99**(1), 90–99. doi:10.1037/0033-2909.99.1.90.

Toothaker LE, Newman D (1994). "Nonparametric Competitors to the Two-Way ANOVA." *Journal of Educational and Behavioral Statistics*, **19**(3), 237–273. doi:10.3102/10769986019003237.

Vallejo G, Ato M, Fernandez MP (2010). "A Robust Approach for Analyzing Unbalanced Factorial Designs with Fixed Levels." *Behavior Research Methods*, **42**(2), 607–617. doi:10.3758/BRM.42.2.607.

van der Waerden BL (1953). "Order Tests for the Two-Sample Problem. II, III." *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Serie A*, **564**, 303–316.

Winer BJ, Brown DR, Michels KM (1991). *Statistical Principles in Experimental Design*. McGraw-Hill. doi:10.1037/11774-000.

Appendix

The following tables give an extract of the type I error rates of the interaction AB as well as of main effect B for the proposed KWF and vdWS procedures, for the parametric F test including H-F adjustment, the PS INT, Koch, INT, ART INT and ATS methods. The error rates are listed for ni=5,50, equal and unequal cell counts, small and large designs, equal and unequal correlations. Rates outside the interval of liberal robustness are marked bold.

Table 3: Type I error rates for the interaction AB (null model)

size / correlation		small / equal r				large / equal r				small / unequal r				large / unequal r			
		eq		ne		eq		ne		eq		ne		eq		ne	
distribution	method	5	50	5	50	5	50	5	50	5	50	5	50	5	50	5	50
normal	parametric	5.5	5.4	5.9	5.5	7.3	**7.6**	**7.6**	**7.8**	6.2	6.3	7.1	6.3	**9.4**	**10.5**	**9.2**	**9.6**
hetero B	param/HF	4.8	4.8	5.3	4.9	5.0	5.6	5.1	5.8	4.7	4.8	5.3	4.9	5.2	5.9	5.2	5.5
V(B)	vdWaerden	4.2	5.2	3.7	5.0	4.7	5.4	4.0	5.5	4.2	5.4	3.6	5.1	5.5	6.3	4.3	6.2
	KWF	4.2	5.2	3.7	5.0	4.6	5.3	4.2	5.8	4.2	5.4	3.6	5.1	5.3	6.3	4.3	6.2
	PS INT	4.3	5.3	4.9	5.4	5.4	6.8	5.0	7.0	5.1	6.0	5.8	6.0	6.9	**9.5**	7.0	**8.6**
	Koch	3.6	4.2	3.4	5.0	2.5	4.9	2.3	5.3	3.7	4.5	3.4	4.9	2.6	4.8	2.4	5.0
	INT	5.2	5.4	5.8	5.6	6.6	7.0	6.1	7.1	6.2	6.0	7.0	6.1	**8.4**	**9.7**	**8.2**	**8.8**
	ART/INT	5.6	5.3	5.8	5.5	7.0	7.4	6.5	7.4	6.1	6.2	6.9	6.2	**8.6**	**9.9**	**8.0**	**8.6**
	ATS	5.8	5.3	**12.1**	6.2	4.0	5.0	5.8	5.1	6.2	5.4	**10.8**	6.0	4.7	5.6	6.9	5.4
normal	parametric	6.0	5.6	5.3	4.2	6.6	6.9	5.8	5.4	6.4	6.1	6.0	4.4	**7.6**	**8.0**	6.8	6.9
hetero A	param/HF	6.0	5.6	5.3	4.2	6.2	6.8	5.5	5.4	5.9	5.8	5.6	4.2	6.2	6.8	5.5	5.9
V(A)	vdWaerden	4.5	4.3	3.5	4.8	3.9	5.2	3.6	5.2	4.5	5.2	3.9	5.0	4.6	5.3	4.2	5.8
	KWF	4.5	4.3	3.5	4.8	3.9	5.2	3.6	5.0	4.5	5.2	3.9	5.0	4.4	5.2	4.1	5.7
	PS INT	4.3	5.5	3.7	4.4	4.6	6.2	4.6	5.3	4.6	5.6	4.7	4.7	5.7	7.1	5.0	6.6
	Koch	4.3	5.4	3.7	5.2	3.0	6.2	2.6	5.5	4.1	5.6	3.9	5.6	2.9	6.0	2.8	5.8
	INT	5.6	5.6	5.0	4.6	5.8	6.2	5.6	5.5	5.8	5.7	6.0	4.8	7.5	7.2	6.4	6.8
	ART/INT	6.2	6.0	5.8	5.2	6.8	7.3	6.2	6.0	7.0	6.3	6.2	5.2	7.4	**8.0**	7.0	7.3
	ATS	5.6	5.0	**10.7**	5.5	3.5	5.1	4.1	5.0	6.3	5.1	**10.4**	5.9	4.3	5.1	4.7	5.1
normal	parametric	6.5	6.1	6.2	5.1	**8.5**	**8.9**	**7.9**	**8.0**	7.3	6.7	6.8	5.8	**11.3**	**10.9**	**9.3**	**9.6**
hetero	param/HF	5.9	5.3	5.2	4.5	6.1	6.7	5.9	5.8	5.8	5.4	5.6	4.4	6.3	7.0	5.5	5.9
A and B	vdWaerden	4.2	5.2	3.7	5.0	4.7	5.4	4.0	5.5	4.2	5.4	3.6	5.1	5.5	6.3	4.3	6.2
V(A,B)	KWF	4.2	5.2	3.7	5.0	4.6	5.3	4.2	5.8	4.2	5.4	3.6	5.1	5.3	6.3	4.3	6.2
	PS INT	4.3	5.2	5.0	5.2	5.3	6.8	5.3	6.5	5.2	5.9	5.3	5.6	7.5	**9.1**	7.0	**7.9**
	Koch	4.2	5.1	3.8	6.1	3.0	5.7	2.8	5.6	3.9	5.1	3.8	5.9	2.8	5.9	2.8	5.7
	INT	5.8	5.3	6.2	5.3	6.4	7.0	6.6	6.8	6.5	6.1	6.6	5.7	**8.9**	**9.2**	**8.1**	**8.1**
	ART/INT	6.4	6.2	6.4	6.0	**8.2**	**8.9**	**7.9**	**8.0**	7.2	7.1	6.9	6.2	**10.3**	**10.8**	**8.9**	**9.1**
	ATS	6.2	5.2	**11.6**	6.0	3.8	5.0	4.8	5.0	6.5	5.1	**10.7**	5.8	4.7	5.2	5.7	5.2
exponent	parametric	4.8	4.7	6.1	5.3	4.7	4.5	5.8	5.1	5.3	5.4	6.7	5.4	5.5	5.2	6.4	5.5
discrete	param/HF	4.4	4.6	5.7	5.3	3.8	4.5	5.0	4.9	4.3	4.8	5.9	4.8	4.3	4.4	5.1	4.8
	vdWaerden	4.6	5.0	4.2	5.0	3.9	4.7	4.0	5.0	4.8	5.2	4.5	4.6	4.6	5.3	4.1	5.4
	KWF	4.7	5.0	4.2	5.0	3.9	4.7	3.9	4.9	4.6	5.3	4.5	4.7	4.6	5.4	4.0	5.3
	PS INT	4.4	4.8	4.8	5.2	3.6	4.6	4.1	5.2	5.1	5.5	6.0	5.2	5.0	6.2	5.0	5.6
	Koch	3.8	4.8	3.6	5.5	2.7	4.6	2.8	5.0	3.8	5.3	3.4	4.9	2.5	4.7	2.6	5.0
	INT	5.5	5.1	5.8	5.4	4.8	4.8	4.9	5.3	6.6	5.6	7.0	5.4	6.0	6.5	6.2	5.8
	ART/INT	5.6	4.9	6.1	5.5	5.1	4.8	5.2	5.3	5.8	5.5	6.4	5.3	5.7	5.5	5.9	5.6
	ATS	6.1	5.2	**11.2**	5.7	3.0	4.7	5.4	5.2	6.8	5.2	**10.4**	5.3	4.0	5.2	5.4	5.0
uniform	parametric	5.1	4.6	5.9	5.8	4.8	5.0	5.2	5.4	5.6	5.5	6.4	5.6	5.8	6.2	5.9	5.8
discrete	param/HF	5.4	4.7	5.9	5.8	4.9	5.0	5.3	5.4	5.2	5.1	5.9	5.3	5.2	5.4	5.0	5.3
	vdWaerden	4.2	4.9	3.8	5.3	4.0	4.8	3.8	5.2	4.2	4.8	3.9	5.3	4.4	5.4	3.8	5.9
	KWF	4.2	5.0	3.8	5.4	4.0	4.8	3.9	5.2	4.2	4.8	3.9	5.4	4.4	5.4	3.9	5.9
	PS INT	4.2	4.7	4.8	5.7	3.9	5.0	3.9	5.2	4.5	5.2	5.2	5.3	4.8	6.2	4.5	5.8
	Koch	3.6	4.8	3.2	5.7	2.6	4.9	2.2	5.0	3.4	4.7	3.1	5.2	2.4	5.2	2.4	4.9
	INT	5.2	4.8	5.8	5.8	4.9	5.2	5.1	5.3	5.3	5.4	6.3	5.5	5.9	6.2	5.8	6.1
	ART/INT	5.4	4.8	5.7	5.6	4.6	4.9	5.2	5.4	5.7	5.5	6.3	5.6	5.8	6.3	5.8	5.9
	ATS	6.0	4.7	**11.4**	5.7	3.7	4.8	5.5	5.5	5.9	5.2	**10.0**	5.7	4.2	5.2	5.8	4.8
normal	parametric	4.9	5.1	5.4	5.9	5.5	5.3	5.7	5.4	5.1	5.3	5.4	4.8	5.4	5.5	5.7	5.4
contamintd	param/HF	4.4	4.9	5.2	5.8	5.3	5.2	5.2	5.0	4.7	5.1	5.1	4.8	4.8	5.3	5.2	5.0
III	vdWaerden	5.0	4.7	4.6	4.8	4.2	5.2	4.6	5.2	4.4	5.2	3.8	4.9	4.8	5.2	4.6	5.2
	KWF	5.0	4.7	4.6	4.8	4.2	4.9	4.5	5.2	4.4	5.2	3.8	4.9	4.8	5.2	4.5	5.2
	PS INT	4.3	4.8	4.1	5.5	3.8	5.0	4.5	5.1	4.3	4.8	4.2	4.6	4.0	5.2	4.5	5.1
	Koch	3.9	4.8	3.2	5.1	3.0	5.0	2.8	4.7	3.8	4.7	3.1	4.5	3.0	5.1	2.8	4.7
	INT	5.5	4.9	5.0	5.6	4.8	5.2	5.4	5.3	5.4	4.9	5.3	4.8	5.3	5.5	5.4	5.3
	ART/INT	5.3	4.9	5.6	5.8	5.4	5.3	6.0	5.2	5.3	5.3	5.6	4.7	5.9	5.6	6.0	5.2
	ATS	6.5	5.1	**10.4**	5.4	3.6	4.8	5.6	4.6	6.0	5.3	**8.9**	5.6	4.1	5.2	5.6	4.6

Table 4: Type I error rates for the interaction AB (model with nonnull A effect)

size / correlation		small / equal r				large / equal r				small / unequal r				large / unequal r			
		eq		ne		eq		ne		eq		ne		eq		ne	
distribution	method	5	50	5	50	5	50	5	50	5	50	5	50	5	50	5	50
normal	parametric	5.5	5.4	5.9	5.5	7.3	**7.6**	**7.6**	**7.8**	6.2	6.3	7.1	6.3	**9.4**	**10.5**	**9.2**	**9.6**
hetero B	param/HF	4.8	4.8	5.3	4.9	5.0	5.6	5.1	5.8	4.7	4.8	5.3	4.9	5.2	5.9	5.2	5.5
V(B)	vdWaerden	4.2	5.2	3.7	5.0	4.7	5.4	4.0	5.5	4.2	5.4	3.6	5.1	5.5	6.3	4.3	6.2
	KWF	4.2	5.2	3.7	5.0	4.6	5.3	4.2	5.8	4.2	5.4	3.6	5.1	5.3	6.3	4.3	6.2
	PS INT	4.2	5.2	4.9	5.5	5.4	7.2	5.1	7.2	5.2	6.2	5.9	6.4	7.5	**10.0**	7.5	**9.3**
	Koch	3.6	4.2	3.4	5.0	2.5	4.9	2.3	5.3	3.7	4.5	3.4	4.9	2.6	4.8	2.4	5.0
	INT	5.2	5.3	6.0	5.7	6.8	7.2	6.4	7.2	6.2	6.2	7.1	6.6	**8.6**	**10.2**	**8.6**	**9.4**
	ART/INT	5.6	5.3	5.8	5.5	7.0	7.4	6.5	7.4	6.1	6.2	6.9	6.2	**8.6**	**9.9**	**8.0**	**8.6**
	ATS	6.0	**8.2**	**11.6**	7.4	4.8	**10.4**	6.1	**8.1**	6.6	**9.2**	**11.5**	**8.2**	5.5	**11.2**	7.2	**8.4**
normal	parametric	6.0	5.6	5.3	4.2	6.6	6.9	5.8	5.4	6.4	6.1	6.0	4.4	**7.6**	**8.0**	6.8	6.9
hetero A	param/HF	6.0	5.6	5.3	4.2	6.2	6.8	5.5	5.4	5.9	5.8	5.6	4.2	6.2	6.8	5.5	5.9
V(A)	vdWaerden	4.5	4.3	3.5	4.8	3.9	5.2	3.6	5.2	4.5	5.2	3.9	5.0	4.6	5.3	4.2	5.8
	KWF	4.5	4.3	3.5	4.8	3.9	5.2	3.6	5.0	4.5	5.2	3.9	5.0	4.4	5.2	4.1	5.7
	PS INT	4.5	5.3	4.2	4.7	4.4	6.0	4.4	5.4	4.3	5.8	4.9	5.0	5.8	7.2	5.1	6.8
	Koch	4.3	5.4	3.7	5.2	3.0	6.2	2.6	5.5	4.1	5.6	3.9	5.6	2.9	6.0	2.8	5.8
	INT	5.4	5.5	5.4	5.0	5.7	6.2	5.6	5.5	5.7	5.9	6.2	5.1	6.9	7.3	6.2	7.0
	ART/INT	6.2	6.0	5.8	5.2	6.8	7.3	6.2	6.0	7.0	6.3	6.2	5.2	7.4	**8.0**	7.0	7.3
	ATS	5.8	5.1	**10.9**	5.6	3.8	5.1	4.0	5.0	6.2	5.1	**10.4**	5.8	4.3	5.1	4.3	5.0
normal	parametric	6.5	6.1	6.2	5.1	**8.5**	**8.9**	**7.9**	**8.0**	7.3	6.7	6.8	5.8	**11.3**	**10.9**	**9.3**	**9.6**
hetero	param/HF	5.9	5.3	5.2	4.5	6.1	6.7	5.9	5.8	5.8	5.4	5.6	4.4	6.3	7.0	5.5	5.9
A and B	vdWaerden	4.2	5.2	3.7	5.0	4.7	5.4	4.0	5.5	4.2	5.4	3.6	5.1	5.5	6.3	4.3	6.2
V(A,B)	KWF	4.2	5.2	3.7	5.0	4.6	5.3	4.2	5.8	4.2	5.4	3.6	5.1	5.3	6.3	4.3	6.2
	PS INT	4.4	5.8	4.8	6.0	5.9	**8.3**	5.3	7.5	5.5	6.8	5.7	6.6	**7.6**	**10.1**	7.1	**9.2**
	Koch	4.2	5.1	3.8	6.1	3.0	5.7	2.8	5.6	3.9	5.1	3.8	5.9	2.8	5.9	2.8	5.7
	INT	5.6	5.9	6.2	6.1	7.1	**8.5**	6.6	**7.8**	6.5	7.2	6.9	6.9	**9.2**	**10.4**	**8.2**	**9.4**
	ART/INT	6.4	6.2	6.4	6.0	**8.2**	**8.9**	**7.9**	**8.0**	7.2	7.1	6.9	6.2	**10.3**	**10.8**	**8.9**	**9.1**
	ATS	6.2	**8.8**	**11.9**	**7.6**	4.6	**10.3**	5.3	**8.4**	6.9	**10.3**	**11.2**	**8.9**	5.8	**10.6**	6.2	**8.7**
exponent	parametric	5.3	5.6	4.1	3.9	5.5	5.2	6.1	5.5	5.3	5.3	4.5	4.4	5.7	6.1	6.8	6.9
discrete	param/HF	4.9	5.4	3.8	3.6	4.5	5.0	4.8	5.3	4.4	4.7	3.8	4.1	4.2	5.2	4.9	5.6
	vdWaerden	3.9	5.2	3.2	4.6	3.9	4.7	3.6	4.5	4.4	5.7	3.5	4.2	4.6	5.1	4.2	5.8
	KWF	3.8	5.2	3.1	4.6	4.1	4.8	3.5	4.3	4.4	5.7	3.4	4.3	4.3	5.2	4.1	5.6
	PS INT	4.0	5.3	3.3	4.5	3.7	4.6	4.0	5.1	4.6	5.7	4.3	4.2	4.5	5.8	4.7	6.5
	Koch	3.5	5.1	2.7	4.7	2.8	4.4	2.5	4.7	3.6	5.4	2.6	3.9	2.8	4.5	2.7	5.3
	INT	4.7	5.4	4.4	4.6	4.9	4.6	5.1	5.2	5.6	5.8	5.2	4.4	5.3	5.8	6.0	6.6
	ART/INT	5.7	6.0	4.5	4.9	5.8	5.4	6.2	5.7	6.4	6.1	4.8	4.9	6.4	6.4	6.7	6.8
	ATS	6.1	5.0	**10.4**	4.9	3.4	4.4	4.9	4.7	6.1	5.8	**8.8**	4.6	4.0	4.7	5.3	5.3
uniform	parametric	5.2	4.8	5.6	5.3	4.7	5.2	4.3	5.2	5.3	5.1	6.1	5.4	6.3	6.4	5.3	5.8
discrete	param/HF	5.4	4.8	5.5	5.2	4.6	5.2	4.5	5.2	5.0	4.8	5.6	5.2	5.8	5.6	4.8	5.0
	vdWaerden	4.1	5.0	3.5	5.1	4.0	4.9	3.3	5.4	4.4	4.7	3.9	5.1	5.1	5.7	3.6	5.4
	KWF	4.2	4.9	3.5	4.9	4.0	5.1	3.3	5.4	4.4	4.7	3.9	5.0	5.0	5.7	3.5	5.3
	PS INT	4.2	4.6	4.2	5.2	3.5	5.5	3.5	5.1	4.2	4.9	4.6	4.9	5.3	6.2	4.2	6.1
	Koch	3.8	4.5	3.3	5.2	2.6	5.2	2.6	5.1	3.6	4.6	3.2	4.8	2.7	5.1	2.7	4.6
	INT	5.2	4.8	5.3	5.3	4.5	5.6	4.2	5.3	5.2	5.1	5.8	5.0	6.3	6.4	5.4	6.3
	ART/INT	5.2	4.8	5.1	5.3	4.7	5.3	4.4	5.2	5.5	5.1	5.7	5.5	6.2	6.4	5.4	5.8
	ATS	6.0	4.8	**10.9**	5.8	3.4	4.7	5.2	5.5	5.9	4.8	**10.3**	5.7	4.6	5.2	5.4	4.9
normal	parametric	4.9	5.1	5.4	5.9	5.5	5.3	5.3	4.9	5.1	5.3	5.4	4.8	5.4	5.5	5.7	5.4
contamintd	param/HF	4.4	4.9	5.2	5.8	5.3	5.2	4.9	4.7	4.7	5.1	5.1	4.8	4.8	5.3	5.2	5.0
III	vdWaerden	5.0	4.7	4.6	4.8	4.2	5.2	3.8	4.8	4.4	5.2	3.8	4.9	4.8	5.2	4.6	5.2
	KWF	5.0	4.7	4.6	4.8	4.2	4.9	3.8	4.8	4.4	5.2	3.8	4.9	4.8	5.2	4.5	5.2
	PS INT	4.3	4.8	4.5	5.6	4.2	5.0	3.9	4.3	4.4	4.9	5.0	5.1	4.3	5.3	4.5	5.0
	Koch	3.9	4.8	3.2	5.1	3.0	5.0	2.6	4.2	3.8	4.7	3.1	4.5	3.0	5.1	2.8	4.7
	INT	5.6	4.8	5.8	5.8	5.1	5.1	4.9	4.6	5.4	5.0	6.0	5.2	5.5	5.5	5.5	5.2
	ART/INT	5.3	4.9	5.6	5.8	5.4	5.3	5.3	4.9	5.3	5.3	5.6	4.7	5.9	5.6	6.0	5.2
	ATS	6.2	5.0	**10.6**	5.4	3.5	4.9	6.0	4.9	6.0	5.0	**9.0**	5.5	4.0	5.4	5.9	4.8

Table 5: type I error rates for the interaction AB (model with nonnull B effect)

size / correlation		small / equal r				large / equal r				small / unequal r				large / unequal r			
		eq		ne		eq		ne		eq		ne		eq		ne	
distribution	method	5	50	5	50	5	50	5	50	5	50	5	50	5	50	5	50
normal	parametric	5.5	5.4	5.9	5.5	7.3	**7.6**	**7.6**	**7.8**	6.2	6.3	7.1	6.3	**9.4**	**10.5**	**9.2**	**9.6**
hetero B	param/HF	4.8	4.8	5.3	4.9	5.0	5.6	5.1	5.8	4.7	4.8	5.3	4.9	5.2	5.9	5.2	5.5
V(B)	vdWaerden	3.0	3.4	2.8	3.4	2.5	3.2	2.2	3.4	2.4	3.0	2.1	3.2	2.4	2.9	2.3	3.0
	KWF	3.0	3.4	2.8	3.4	2.3	3.0	2.2	3.3	2.4	3.0	2.1	3.2	2.2	2.8	2.1	2.7
	PS INT	2.6	3.5	2.9	3.3	2.9	4.0	2.6	4.2	2.9	3.5	3.2	3.5	4.0	5.8	3.7	5.3
	Koch	3.6	4.2	3.4	5.0	2.5	4.9	2.3	5.3	3.7	4.5	3.4	4.9	2.6	4.8	2.4	5.0
	INT	5.5	5.3	5.8	5.5	6.4	7.1	6.1	7.2	6.1	6.2	7.1	6.0	**8.4**	**10.0**	**8.1**	**9.0**
	ART/INT	5.6	5.3	5.8	5.5	7.0	7.4	6.5	7.4	6.1	6.2	6.9	6.2	**8.6**	**9.9**	**8.0**	**8.6**
	ATS	6.2	5.0	**11.4**	6.3	3.9	5.0	5.6	4.9	6.2	5.5	**11.0**	6.3	5.0	5.7	6.8	5.4
normal	parametric	5.4	5.3	5.4	4.5	6.7	7.3	5.6	5.8	6.0	6.3	5.8	5.0	**7.6**	**8.0**	6.4	6.6
hetero A	param/HF	5.5	5.3	5.2	4.4	6.4	7.3	5.3	5.8	5.5	5.8	5.3	4.7	6.2	6.8	5.5	5.8
V(A)	vdWaerden	4.1	**8.7**	4.0	**9.7**	4.3	**9.4**	3.7	**10.3**	4.0	**12.9**	4.4	**15.7**	3.2	**29.2**	3.8	**12.9**
	KWF	4.1	**8.7**	4.0	**9.7**	4.2	**9.5**	3.7	**10.3**	4.0	**12.9**	4.4	**15.7**	3.2	**29.6**	3.6	**12.9**
	PS INT	3.5	4.7	3.9	3.5	3.8	5.4	3.1	4.9	3.1	4.8	3.4	4.2	3.0	4.1	4.2	5.4
	Koch	3.8	5.6	3.8	5.2	3.3	6.2	3.0	5.7	3.9	5.8	3.4	5.6	2.9	6.0	3.0	5.8
	INT	5.1	5.4	5.6	4.6	6.2	6.8	5.1	5.9	5.6	6.3	5.7	5.4	7.2	7.9	6.7	7.0
	ART/INT	5.7	5.6	5.7	5.2	6.8	7.5	6.1	6.6	6.4	6.8	5.8	5.7	7.4	**8.0**	6.7	7.2
	ATS	6.1	6.2	**10.5**	6.2	4.0	6.2	3.9	6.5	5.8	7.2	**10.2**	6.5	5.2	**12.3**	4.3	6.4
normal	parametric	5.4	5.3	5.4	4.5	6.7	7.3	5.6	5.8	6.0	6.3	5.8	5.0	**11.3**	**10.9**	6.4	6.6
hetero	param/HF	5.5	5.3	5.2	4.4	6.4	7.3	5.3	5.8	5.5	5.8	5.3	4.7	6.3	7.0	5.5	5.8
A and B	vdWaerden	3.7	**13.2**	3.6	**16.7**	4.0	**14.0**	3.5	**14.1**	3.1	**18.6**	3.7	**25.5**	3.6	**28.1**	3.3	**19.0**
V(A,B)	KWF	3.7	**13.2**	3.6	**16.7**	3.8	**13.8**	3.6	**13.9**	3.1	**18.6**	3.7	**25.5**	3.5	**29.1**	3.3	**19.2**
	PS INT	2.9	3.9	3.1	2.7	3.4	4.5	2.8	4.2	2.1	3.4	2.2	2.9	4.0	6.0	3.3	4.4
	Koch	3.8	5.6	3.8	5.2	3.3	6.2	3.0	5.7	3.9	5.8	3.4	5.6	2.8	5.9	3.0	5.8
	INT	5.0	5.5	5.8	4.6	6.3	6.8	5.1	5.9	5.9	6.6	5.5	5.7	**9.6**	**10.6**	6.8	7.0
	ART/INT	5.7	5.6	5.7	5.2	6.8	7.5	6.1	6.6	6.4	6.8	5.8	5.7	**10.3**	**10.8**	6.7	7.2
	ATS	6.0	**7.8**	**11.3**	6.7	4.1	**7.6**	4.0	7.5	6.4	**9.9**	**10.9**	**7.8**	5.5	**11.6**	4.4	**7.9**
exponent	parametric	5.5	5.4	6.4	5.2	5.6	5.6	6.0	5.9	4.9	5.4	6.4	5.6	5.6	5.1	5.6	6.1
discrete	param/HF	4.3	4.8	5.5	4.7	4.0	4.6	4.5	5.0	4.2	5.1	5.8	5.4	4.2	4.3	4.6	5.2
	vdWaerden	3.9	3.9	3.4	4.2	2.7	3.7	2.3	3.7	2.9	3.1	2.9	3.3	2.5	3.6	2.2	3.7
	KWF	3.8	3.9	3.4	4.2	2.8	3.8	2.2	3.8	2.9	3.0	2.9	3.3	2.5	3.5	2.1	3.9
	PS INT	3.3	3.9	3.9	4.3	2.4	3.4	2.1	3.5	2.6	3.5	3.5	3.5	2.5	3.5	2.3	3.6
	Koch	4.0	4.5	3.6	5.3	2.5	4.4	2.4	4.6	3.6	5.1	4.0	5.0	2.5	4.4	2.0	5.0
	INT	5.4	5.1	5.9	5.5	4.8	5.3	4.6	5.0	5.9	5.8	6.7	5.1	5.3	5.6	4.7	5.7
	ART/INT	5.8	5.1	6.0	5.4	5.6	5.4	5.4	5.5	5.5	5.6	6.1	5.4	5.6	5.1	5.3	6.1
	ATS	6.4	5.0	**12.0**	5.9	3.4	4.5	4.7	4.7	6.6	5.2	**10.6**	5.2	3.8	5.1	4.8	4.7
uniform	parametric	5.5	4.9	5.8	5.2	4.9	5.3	4.2	4.9	5.5	5.0	6.6	5.8	6.0	6.4	5.2	5.7
discrete	param/HF	5.5	5.0	5.9	5.3	5.1	5.2	4.2	4.8	5.1	4.8	6.1	5.4	5.4	5.5	4.8	5.2
	vdWaerden	3.1	3.1	2.2	3.3	1.9	2.6	1.8	3.1	1.8	2.0	1.7	2.5	1.8	2.4	1.6	2.1
	KWF	3.1	3.1	2.3	3.3	2.0	2.7	1.8	3.2	1.8	2.0	1.8	2.5	1.9	2.5	1.7	2.1
	PS INT	2.1	2.9	2.7	3.1	1.4	2.3	1.5	2.5	1.7	1.9	2.0	2.3	1.8	2.3	1.7	2.1
	Koch	3.6	4.7	3.1	5.2	2.5	4.9	2.5	5.0	3.7	4.4	3.2	5.3	2.5	4.9	2.4	4.7
	INT	5.4	4.8	5.8	5.2	4.7	4.9	4.2	5.0	5.2	5.0	6.0	5.7	5.8	6.1	5.2	5.5
	ART/INT	5.5	4.9	5.6	5.2	4.9	5.2	4.3	4.9	5.6	4.9	6.2	5.7	6.1	6.3	5.2	5.7
	ATS	6.1	4.7	**11.0**	5.4	3.7	4.9	5.2	5.6	5.8	4.8	**10.8**	5.8	4.2	5.1	5.7	5.0
normal	parametric	4.9	5.1	5.4	5.9	5.5	5.3	5.3	4.9	5.1	5.3	5.4	4.8	5.4	5.5	5.7	5.4
contamintd	param/HF	4.4	4.9	5.2	5.8	5.3	5.2	4.9	4.7	4.7	5.1	5.1	4.8	4.8	5.3	5.2	5.0
III	vdWaerden	3.2	4.2	3.5	4.3	3.0	3.8	2.2	3.3	3.0	3.5	2.8	3.3	3.0	3.6	2.6	3.6
	KWF	3.2	4.2	3.5	4.3	2.9	3.9	2.2	3.0	3.0	3.5	2.8	3.3	2.9	3.6	2.8	3.7
	PS INT	3.6	4.1	3.7	4.6	3.1	3.8	2.8	3.5	3.4	4.0	3.6	3.8	3.2	4.2	3.4	3.8
	Koch	3.9	4.8	3.2	5.1	3.0	5.0	2.6	4.2	3.8	4.7	3.1	4.5	3.0	5.1	2.8	4.7
	INT	5.3	5.2	5.2	5.6	5.0	5.2	5.0	4.6	5.5	5.0	5.9	4.9	5.8	5.6	5.8	5.4
	ART/INT	5.3	4.9	5.6	5.8	5.4	5.3	5.3	4.9	5.3	5.3	5.6	4.7	5.9	5.6	6.0	5.2
	ATS	6.3	5.2	**10.4**	5.2	3.8	4.9	5.6	4.6	6.2	5.3	9.0	5.6	4.2	5.6	5.6	4.8

Table 6: Type I error rates for main effect B (model with nonnull A effect)

size / correlation		small / equal r				large / equal r				small / unequal r				large / unequal r			
		eq		ne		eq		ne		eq		ne		eq		ne	
distribution	method	5	50	5	50	5	50	5	50	5	50	5	50	5	50	5	50
normal	parametric	5.8	5.9	5.9	5.9	7.3	6.8	7.2	6.8	6.4	6.6	6.4	6.6	**8.5**	**8.1**	**8.4**	**8.1**
hetero B	param/HF	5.3	5.3	5.6	5.2	5.9	5.2	5.8	5.3	5.4	5.2	5.5	5.2	5.7	5.0	5.8	5.1
V(B)	vdWaerden	6.0	5.4	6.0	5.4	5.2	5.1	5.2	5.1	5.7	5.3	5.7	5.3	5.9	5.0	5.9	5.0
	KWF	6.0	5.4	6.0	5.4	5.2	5.2	5.2	5.2	5.7	5.3	5.7	5.3	5.9	5.0	5.9	5.0
	PS INT	5.4	5.3	5.4	5.5	6.1	5.6	6.1	5.6	6.2	6.3	6.0	6.1	7.4	7.3	7.4	7.3
	Koch	5.2	5.4	5.2	5.4	4.0	4.5	4.0	4.5	5.5	5.5	5.5	5.5	4.1	4.5	4.1	4.5
	INT	5.9	5.5	5.9	5.4	6.5	5.6	6.4	5.6	6.9	6.3	6.5	6.1	**7.6**	7.4	7.4	7.4
	ART/INT	5.9	5.2	6.0	5.1	6.6	5.7	6.3	5.8	6.5	6.3	6.5	6.4	**7.9**	7.3	**7.8**	7.3
	ATS	**7.6**	6.2	**11.5**	6.3	6.5	5.6	6.7	5.7	**8.4**	7.0	**11.3**	7.1	6.5	5.8	7.1	5.6
normal	parametric	5.6	5.4	5.4	5.6	5.3	5.3	5.2	5.2	6.0	6.1	6.0	6.0	6.3	6.2	6.1	6.2
hetero A	param/HF	5.6	5.4	5.5	5.5	4.9	5.2	4.8	5.2	5.4	5.7	5.5	5.5	5.5	5.6	5.3	5.4
A	vdWaerden	5.1	5.6	5.1	5.6	5.2	5.3	5.2	5.3	5.6	5.9	5.6	5.9	5.0	5.2	5.0	5.2
V(A)	KWF	5.1	5.6	5.1	5.6	5.0	5.3	5.0	5.3	5.6	5.9	5.6	5.9	4.9	5.0	4.9	5.0
	PS INT	4.8	5.6	5.0	5.6	4.4	5.1	4.3	5.2	5.3	5.6	5.3	5.8	5.5	5.8	5.4	5.9
	Koch	4.8	5.8	4.8	5.8	4.0	5.2	4.0	5.2	5.6	6.0	5.6	6.0	4.1	5.0	4.1	5.0
	INT	5.7	5.8	5.9	5.7	5.2	5.2	5.3	5.3	5.8	5.7	5.8	5.8	6.1	6.1	6.2	6.0
	ART/INT	6.3	6.4	6.8	6.7	6.5	6.7	6.8	7.0	6.7	6.8	7.1	6.8	**7.8**	7.3	**7.7**	7.4
	ATS	6.7	5.5	**9.7**	5.6	4.8	5.0	4.9	5.0	6.2	5.5	**9.4**	5.2	5.5	4.9	5.4	4.9
normal	parametric	6.3	5.8	6.5	5.6	7.0	7.1	7.2	7.2	7.3	6.6	7.1	6.5	**8.5**	**8.4**	**8.5**	**8.4**
hetero	param/HF	5.7	5.2	5.6	5.0	5.3	5.2	5.4	5.4	5.8	5.2	5.8	5.3	5.4	5.5	5.5	5.4
A and B	vdWaerden	6.0	5.4	6.0	5.4	5.2	5.1	5.2	5.1	5.7	5.3	5.7	5.3	5.9	5.0	5.9	5.0
V(A,B)	KWF	6.0	5.4	6.0	5.4	5.2	5.2	5.2	5.2	5.7	5.3	5.7	5.3	5.9	5.0	5.9	5.0
	PS INT	5.3	6.3	5.3	6.9	6.4	**8.4**	6.5	**9.0**	6.2	**8.2**	6.2	**8.5**	7.5	**10.5**	**7.6**	**10.8**
	Koch	5.2	5.4	5.2	5.4	4.0	4.5	4.0	4.5	5.5	5.5	5.5	5.5	4.1	4.5	4.1	4.5
	INT	6.0	6.4	6.0	7.0	6.8	**8.6**	7.0	9.1	7.0	**8.2**	6.8	**8.7**	**8.1**	**10.7**	**8.1**	**11.0**
	ART/INT	**7.9**	**8.4**	**8.6**	**9.4**	**8.9**	**12.5**	**9.4**	**13.5**	**9.2**	**10.8**	**9.8**	**11.8**	**11.1**	**14.7**	**11.5**	**15.0**
	ATS	7.5	7.1	**11.4**	6.3	6.6	**7.6**	6.3	7.4	**8.2**	7.7	**11.1**	7.0	6.6	7.4	6.7	7.2
exponent	parametric	4.3	4.9	5.0	5.0	4.6	4.8	4.6	4.7	4.4	4.9	5.3	5.2	5.0	6.0	5.0	6.0
discrete	param/HF	3.7	4.8	4.3	4.9	3.4	4.6	3.5	4.6	3.9	4.5	4.4	4.6	3.4	4.9	3.4	4.9
	vdWaerden	4.2	4.6	5.2	4.6	4.7	4.8	4.7	4.8	4.3	5.1	5.3	5.2	4.9	5.4	4.9	5.4
	KWF	4.3	4.6	5.2	4.5	5.0	4.9	5.0	4.9	4.5	5.1	5.4	5.2	4.9	5.4	4.9	5.4
	PS INT	4.2	4.3	4.7	4.9	4.4	5.1	4.4	5.1	4.3	5.1	5.2	4.9	4.6	5.6	4.6	5.6
	Koch	4.2	4.6	5.0	4.5	3.6	4.9	3.6	4.9	4.2	5.0	4.7	4.9	3.3	4.9	3.3	4.9
	INT	4.7	4.4	5.2	5.1	4.7	5.1	4.7	5.2	4.7	5.2	5.5	5.0	4.8	5.6	4.9	5.7
	ART/INT	5.3	7.2	5.8	7.5	4.5	4.4	5.6	5.7	6.4	**9.6**	6.5	**10.8**	5.3	6.3	6.2	**8.2**
	ATS	5.3	4.6	**8.6**	5.0	4.6	5.4	5.7	4.6	5.6	5.0	**8.2**	4.8	4.3	4.6	5.2	4.9
uniform	parametric	6.1	5.2	5.8	5.5	4.8	5.2	4.8	5.2	5.6	5.8	5.5	5.5	5.9	5.7	5.9	5.7
discrete	param/HF	6.0	5.3	6.0	5.5	4.7	5.2	4.8	5.2	5.4	5.4	5.3	5.2	5.5	5.3	5.5	5.2
	vdWaerden	5.2	5.2	5.3	5.2	5.2	5.3	5.2	5.3	4.8	5.2	4.8	5.2	5.3	5.5	5.3	5.5
	KWF	5.4	5.2	5.4	5.3	5.2	5.4	5.2	5.4	4.8	5.2	4.9	5.3	5.2	5.4	5.2	5.4
	PS INT	5.5	5.3	5.4	5.3	4.3	5.2	4.3	5.2	5.0	5.6	5.2	5.5	5.6	6.0	5.6	6.0
	Koch	5.0	5.2	5.1	5.1	3.8	5.6	3.8	5.6	4.5	5.2	4.6	5.3	3.9	5.4	3.9	5.4
	INT	6.0	5.6	5.8	5.3	4.7	5.3	4.6	5.3	5.6	5.7	5.7	5.6	6.2	6.1	6.0	6.0
	ART/INT	6.3	**11.2**	6.9	**13.8**	5.1	**9.3**	5.7	**16.3**	6.7	**14.5**	7.5	**19.1**	6.5	**11.8**	7.2	19.4
	ATS	6.7	5.2	**10.4**	6.1	4.6	5.1	5.7	5.2	6.1	5.4	**9.0**	5.5	5.4	5.0	5.9	5.3
normal	parametric	5.2	**10.2**	5.5	**10.2**	6.0	**16.1**	6.2	**16.2**	5.0	**10.8**	5.2	**10.8**	6.6	**17.7**	6.6	**17.6**
contamintd	param/HF	4.9	**9.9**	5.0	**9.8**	5.2	**15.5**	5.2	**15.5**	4.4	**10.3**	4.5	**10.4**	5.8	**16.9**	5.6	**17.0**
III	vdWaerden	5.6	5.3	5.6	5.3	4.8	5.8	4.8	5.8	5.1	4.7	5.1	4.7	4.5	5.7	4.5	5.7
	KWF	5.6	5.3	5.6	5.3	4.8	5.7	4.8	5.7	5.1	4.7	5.1	4.7	4.6	5.6	4.6	5.6
	PS INT	5.1	6.9	5.0	6.8	4.8	**9.7**	4.8	**9.7**	4.4	7.3	4.5	7.4	5.3	**10.5**	5.3	**10.5**
	Koch	5.3	5.5	5.3	5.5	4.0	5.8	4.0	5.8	5.0	4.8	5.0	4.8	3.8	5.2	3.8	5.2
	INT	5.6	7.0	5.4	6.9	5.3	9.6	5.3	9.6	4.9	7.3	5.2	7.4	5.8	10.5	5.7	10.5
	ART/INT	4.7	6.6	4.9	7.0	4.6	**9.3**	4.8	**9.4**	4.6	7.0	5.1	6.9	5.0	**9.9**	5.2	**10.2**
	ATS	6.2	5.3	**9.6**	5.8	4.9	5.6	5.8	5.4	5.7	5.2	**8.6**	5.4	4.9	6.0	5.8	6.0

Table 7: Type I error rates for all models for the cases of positive and negative pairing, here e.g. A(B) denotes the test of factor A with a nonnull effect of factor B

	pairing	positive								negative							
	covariance	equal r				equal r				unequal r				unequal r			
	size	small		large		small		large		small		large		small		large	
model	method n	5	50	5	50	5	50	5	50	5	50	5	50	5	50	5	50
A(-)	parametric	2.9	2.1	2.0	2.0	3.0	2.2	2.0	1.9	**11.1**	**11.3**	**12.6**	**12.9**	11.3	**11.2**	**12.5**	**12.8**
	vdWaerden	2.7	2.0	1.8	1.7	2.7	2.0	1.9	1.8	6.6	**9.8**	6.9	**10.2**	6.9	**9.8**	7.0	**10.3**
	KWF	2.8	3.1	2.4	2.6	2.9	3.0	2.4	2.6	5.5	**8.3**	6.2	**9.1**	5.5	**8.2**	6.3	**8.8**
	Koch	2.8	3.1	2.5	2.7	2.9	3.1	2.4	2.7	5.8	**8.2**	6.8	**9.1**	5.8	**8.2**	6.9	**8.8**
	ATS	**12.4**	5.1	**10.3**	5.7	**13.0**	5.4	**11.0**	5.4	**16.5**	6.2	**13.6**	6.4	**18.3**	6.5	**14.1**	6.3
A(B)	parametric	2.9	2.1	2.0	2.0	3.0	2.2	2.0	1.9	**11.1**	**11.3**	**12.6**	**12.9**	**11.3**	**11.2**	**12.5**	**12.8**
	vdWaerden	2.7	2.0	1.8	1.7	2.7	2.0	1.9	1.8	6.6	**9.8**	6.9	**10.2**	6.9	**9.8**	7.0	**10.3**
	KWF	2.8	3.1	2.4	2.6	2.9	3.0	2.4	2.6	5.5	**8.3**	6.2	**9.1**	5.5	**8.2**	6.3	**8.8**
	Koch	2.8	3.1	2.5	2.7	2.9	3.1	2.4	2.7	5.8	**8.2**	6.8	**9.1**	5.8	**8.2**	6.9	**8.8**
	ATS	**12.3**	5.3	**10.5**	5.7	**12.6**	5.3	**11.0**	5.4	**17.1**	6.1	**13.7**	6.3	**18.0**	6.4	**14.4**	6.3
A(AB)	parametric	2.4	2.1	2.0	2.0	2.5	2.1	2.0	1.9	**14.5**	**13.5**	**12.6**	**12.9**	**14.7**	**13.6**	**12.5**	**12.8**
	vdWaerden	2.2	2.0	1.8	1.7	2.6	1.8	1.9	1.8	**8.2**	**10.9**	6.9	**10.2**	7.5	**10.6**	7.0	**10.3**
	KWF	2.6	3.1	2.4	2.6	2.8	2.9	2.4	2.6	6.7	**9.4**	6.2	**9.1**	6.0	**9.6**	6.3	**8.8**
	Koch	2.8	3.0	2.5	2.7	2.8	2.7	2.4	2.7	7.2	**9.2**	6.8	**9.1**	6.5	**9.4**	6.9	**8.8**
	ATS	**10.6**	6.3	**10.5**	5.8	**11.5**	6.3	**10.9**	5.6	**17.8**	6.8	**13.8**	6.4	**18.8**	6.9	**14.2**	6.3
B(-)	parametric	5.1	5.8	4.2	4.8	5.7	6.0	5.2	6.2	6.3	5.8	6.4	5.5	6.8	6.0	7.4	6.6
	param/HF	4.9	5.9	4.1	4.8	5.1	5.6	4.6	5.3	6.2	5.7	6.2	5.4	6.3	5.7	6.6	5.8
	vdWaerden	4.7	5.6	5.1	5.2	5.6	6.1	4.9	5.0	4.7	5.6	5.0	5.5	5.6	6.1	5.2	5.6
	KWF	4.7	5.6	5.0	5.2	5.6	6.1	4.8	4.9	4.7	5.6	5.0	5.5	5.6	6.1	5.3	5.3
	Koch	4.8	5.8	4.0	5.2	5.6	6.0	4.1	5.0	4.8	5.8	4.0	5.2	5.6	6.0	3.7	5.1
	ATS	**8.8**	5.7	5.3	4.9	**8.6**	5.3	5.6	5.0	**10.3**	5.8	5.7	5.2	**9.0**	5.6	6.1	5.3
B(A)	parametric	5.1	5.8	4.2	4.8	5.7	6.0	5.2	6.2	6.3	5.8	6.4	5.5	6.8	6.0	7.4	6.6
	param/HF	4.9	5.9	4.1	4.8	5.1	5.6	4.6	5.3	6.2	5.7	6.2	5.4	6.3	5.7	6.6	5.8
	vdWaerden	4.7	5.6	5.1	5.2	5.6	6.1	4.9	5.0	4.7	5.6	5.0	5.5	5.6	6.1	5.2	5.6
	KWF	4.7	5.6	5.0	5.2	5.6	6.1	4.8	4.9	4.7	5.6	5.0	5.5	5.6	6.1	5.3	5.3
	Koch	4.8	5.8	4.0	5.2	5.6	6.0	4.1	5.0	4.8	5.8	4.0	5.2	5.6	6.0	3.7	5.1
	ATS	**8.8**	5.5	5.5	4.6	**8.6**	5.5	5.3	4.9	**9.8**	5.8	5.8	5.3	**9.0**	5.4	5.5	5.5
B(AB)	parametric	4.5	5.0	4.4	4.9	4.8	4.8	5.0	5.8	5.9	5.0	6.7	4.6	6.2	5.8	7.4	5.2
	param/HF	4.6	5.0	4.2	4.9	4.7	4.5	4.5	5.2	5.8	5.0	6.3	4.5	5.8	5.6	6.1	4.3
	vdWaerden	4.5	7.1	4.3	6.1	4.7	6.2	4.4	6.2	5.5	**7.8**	5.0	6.6	5.5	9.4	5.4	7.0
	KWF	4.5	7.1	4.2	6.1	4.7	6.2	4.3	6.0	5.5	**7.8**	4.8	6.4	5.5	9.4	5.4	6.9
	Koch	4.6	**12.4**	3.5	**8.0**	5.1	**21.0**	3.7	**9.5**	**12.8**	**67.6**	**7.8**	**45.3**	**17.6**	**90.3**	**9.1**	**61.2**
	ATS	**16.1**	**67.9**	**9.9**	**41.3**	**21.3**	**88.2**	**10.4**	**50.9**	**11.0**	**13.5**	6.5	**8.6**	**10.3**	**18.1**	6.9	**8.7**
AB(-)	parametric	2.1	1.9	1.0	1.3	2.5	2.1	1.7	1.7	**15.5**	**13.7**	**24.9**	**22.7**	**15.4**	**13.2**	**25.0**	**23.4**
	param/HF	2.0	1.9	0.9	1.2	2.4	1.9	1.3	1.4	**15.4**	**13.6**	**24.5**	**22.6**	**14.4**	**12.2**	**22.5**	**21.4**
	vdWaerden	3.3	4.6	3.2	4.6	4.2	4.6	3.5	5.0	3.5	4.5	3.0	4.9	4.0	4.9	3.4	6.0
	KWF	3.3	4.6	3.3	4.6	4.2	4.6	3.4	5.0	3.5	4.5	2.8	4.8	4.0	4.9	3.4	5.7
	Koch	1.6	2.7	1.0	2.1	1.7	2.6	1.1	2.1	5.8	8.2	5.3	**10.6**	5.7	**8.2**	5.3	**10.8**
	ATS	**9.2**	5.3	4.5	4.9	**8.9**	5.0	4.8	5.5	**12.3**	5.4	5.8	4.6	**10.8**	5.4	6.4	5.0
AB(A)	parametric	2.1	1.9	1.0	1.3	2.5	2.1	1.7	1.7	**15.5**	**13.7**	**24.9**	**22.7**	**15.4**	**13.2**	**25.0**	**23.4**
	param/HF	2.0	1.9	0.9	1.2	2.4	1.9	1.3	1.4	**15.4**	**13.6**	**24.5**	**22.6**	**14.4**	**12.2**	**22.5**	**21.4**
	vdWaerden	3.3	4.6	3.2	4.6	4.2	4.6	3.5	5.0	3.5	4.5	3.0	4.9	4.0	4.9	3.4	6.0
	KWF	3.3	4.6	3.3	4.6	4.2	4.6	3.4	5.0	3.5	4.5	2.8	4.8	4.0	4.9	3.4	5.7
	Koch	1.6	2.7	1.0	2.1	1.7	2.6	1.1	2.1	5.8	8.2	5.3	**10.6**	5.7	**8.2**	5.3	**10.8**
	ATS	**9.0**	5.5	4.2	5.0	**9.1**	5.2	4.9	5.6	**11.9**	5.2	5.4	4.7	**10.3**	5.1	6.2	4.7
AB(B)	parametric	2.1	1.9	1.0	1.3	2.5	2.1	1.7	1.7	**15.5**	**13.7**	**24.9**	**22.7**	**15.4**	**13.2**	**25.0**	**23.4**
	param/HF	2.0	1.9	0.9	1.2	2.4	1.9	1.3	1.4	**15.4**	**13.6**	**24.5**	**22.6**	**14.4**	**12.2**	**22.5**	**21.4**
	vdWaerden	2.2	5.8	1.7	**10.1**	1.8	5.8	1.6	**12.4**	3.0	5.6	2.4	**10.3**	2.3	5.5	2.7	**12.3**
	KWF	2.2	5.8	1.6	**10.3**	1.8	5.8	1.6	**12.6**	3.0	5.6	2.3	**10.2**	2.3	5.5	2.7	**12.3**
	Koch	1.6	2.7	1.0	2.1	1.7	2.6	1.1	2.1	5.8	8.2	5.3	**10.6**	5.7	**8.2**	5.3	**10.8**
	ATS	**9.4**	6.5	4.5	7.6	**9.2**	7.0	5.6	**9.7**	**12.6**	6.1	5.8	6.4	**11.9**	6.8	6.6	6.8

Affiliation:

Haiko Luepsen
Computer Science
University of Cologne
D 50931 Koeln, Germany
E-mail: luepsen@uni-koeln.de

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 52
2023

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2022-05-23
Accepted: 2022-10-31