The Aligned Rank Transform and discrete Variables - a Warning

Version 2
(15.7.2016)

Haiko Lüpsen

Regionales Rechenzentrum (RRZK)

Kontakt: Luepsen@Uni-Koeln.de

Universität zu Köln

The Aligned Rank Transform and discrete Variables - a Warning

Abstract

For two-way layouts in a between subjects anova design the aligned rank transform (ART) is compared with the parametric F-test as well as six other nonparametric methods: rank transform (RT), inverse normal transform (INT), a combination of ART and INT, Puri & Sen‘s L statistic, van der Waerden and Akritas & Brunners ATS. The type I error rates are computed for the uniform and the exponential distributions, both as continuous and in several variations as discrete distribution. The computations had been performed for balanced and unbalanced designs as well as for several effect models. The aim of this study is to analyze the impact of discrete distributions on the error rate. And it is shown that this scaling impact is restricted to the ART- as well as the combination of ART- and INT-method. There are two effects: first with increasing cell counts their error rates rise beyond any acceptable limit up to 20 percent and more. And secondly their rates rise when the number of distinct values of the dependent variable decreases. This behaviour is more severe for underlying exponential distributions than for uniform distributions. Therefore there is a recommendation not to apply the ART if the mean cell frequencies exceed 10.

1. Introduction

The Aligned Rank Transform (ART) seems to be a very popular method for a nonparametric analysis of variance (anova) judging from the number of publications. This procedure dates back to Hodges & Lehmann (1962) and had been made popular by Higgins & Tashtoush (1994) who extended it to factorial designs.

In order to avoid an increase of type I error rates for the interaction in case of significant main effects, as it is observed for the simple rank transform procedure (RT, for details see next chapter), an alignment is proposed: all effects that are not of primary interest are subtracted before performing an anova. 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 of interest, transforming this sum into ranks and finally performing the parametric anova to them. As the normal theory F-tests are used for testing these rank statistics, the question arises if their asymptotic distribution is the same. Salter & Fawcett (1993) showed that at least for the ART these tests are valid.

There exist numerous studies on the ART-technique. But in most papers its behaviour is examined only for the comparison of normal and continuous nonnormal distributions in relation to the parametric F-test and the RT-method. And in general it is estimated rather well, by Lei, Holt & Beasley (2004), Wobbrock et al. (2011) and Mansouri, Paige & Surles (2004) to name only a few. Higgins & Tashtoush (1994) as well as Salter & Fawcett (1993) showed that the ART procedure is valid concerning the type I error rate and that it is preferable to the F-test in cases of outliers or heavily tailed distributions as in these situations the ART has a larger power than the F-test. Richter & Payton (1999) compared the ART with the F-test and with an exact test of the ranks using the exact permutation distribution, but only to check the influence of violation of normal assumption. For nonnormal distributions the ART is superior especially using the exact probabilities.

Nevertheless there are also a couple of critical results. Some authors investigated the behaviour of the ART in heteroscedastic conditions. Among those are Leys & Schumann (2010) and Carletti & Claustriaux (2005). The first analyzed 2*2 designs for various distributions with and without homogeneity of variances. They found that in the case of heteroscedasticity the ART has even more inflated type I errors than the F-test and that concerning the power only for the main effects the ART can compete with the classical tests. Carletti & Claustriaux (2005) who used a 2*4 design with a relation of 4 and 8 for the ratio of the largest to the smallest variance came to the same results. In addition the type I error increases with larger cell counts. But they proposed an amelioration of the ART technique: to transform the ranks obtained from the ART according to the INT method, i.e. transforming them into normal scores (see next chapter). This method leads to a reduction of the type I error rate, especially in the case of unequal variances. Another study to mention in this context is the one by Danbaba (2009). He compared for a simple 3*3 two-way design 25 rank tests with the parametric F-test. He considered 4 distributions but unfortunately not the case of heterogeneous variances. His conclusion: among others the RT, INT, Puri & Sen and ATS fulfill the robustness criterion and show a power superior to the F-test (except for the exponential distribution) whereas the ART fails.

Of course the ART-procedure is designed for continuous dependent variables (dv). But in practice data are often available as integers only which may cause ties when transforming the data into ranks. And the number of ties increases if the data range is rather limited or the total n is fairly large. On the other side anova users prefer nonparametric methods if data are ordinal or look ordinal, i.e. data consist of integers with a limited range. And because of the popularity of the ART they are motivated to apply this method.

There are only few studies considering discrete distributions in their simulations. One of them is the one by Mansouri et al. (2004) in which they studied the ART-procedure for continuous and discrete variables. They found no remarkable differences in the performance. But it has to be mentioned that they studied only designs with cell sizes n_i up to 10. Another study by Kaptein et al (2010) showed, unfortunately only for a 2*2-design, the power of the anova type statistic (ATS, see next chapter for details), being superior to the F-test in the case of Likert scales.

2. Methods to be compared

Beside the ART-technique also other methods are to be compared for discrete dependent variables. It follows a brief description of them.

The anova model shall be denoted by

$$x_{ijk} = \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}$$

with fixed effects α_i (factor A), β_j (factor B), $\alpha\beta_{ij}$ (interaction AB) and error e_{ijk} .

RT (rank transform)

The rank transform method (RT) is just transforming the dependent variable (dv) into ranks and then applying the parametric anova to them. This method had been proposed by Conover & Iman (1981). Blair et al (1987), Toothaker & Newman (1994) as well as Beasley & 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. At least Hora & Conover (1984) proved that the tests of the main effects are correct. A good review of articles concerning the problems of the RT can be found in the study by Toothaker & Newman.

INT (inverse normal transform)

The inverse normal transform method (INT) consists of first transforming the dv into ranks (as in the RT method), then computing their normal scores and finally applying the parametric anova to them. The normal scores are defined as

$$\Phi^{-1}(R_i/(n+1))$$

where R_i are the ranks of the dv and n is the number of observations. It should be noted that there exist several versions of the normal scores (see Beasley, Erickson & Allison (2009) for details). This results in an improvement of the RT procedure as could be shown by Huang (2007) as well as Mansouri and Chang (1995), though Beasley, Erickson & Allison (2009) found out that also the INT procedure results in slightly too high type I error rates if there are significant main effects.

ART combined with INT (ART+INT)

Mansouri & Chang (1995) 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 RT as well as for the ART procedure, at least in the case of underlying normal distributions.

Puri & Sen tests (L statistic)

These are generalizations of the well known Kruskal-Wallis H test (for independent samples) and the Friedman test (for dependent samples) by Puri & Sen (1985), often referred as L statistic. A good introduction offer Thomas et al (1999). The idea dates back to the 60s, when Bennett (1968) and Scheirer, Ray & Hare (1976) as well as later Shirley (1981) generalized the H test for multifactorial designs. It is well known that the Kruskal-Wallis H test as well as the Friedman test can be performed by a suitable ranking of the dv, conducting a parametric anova and finally computing χ^2 ratios using the sum of squares. In fact the same applies to the generalized tests. In the simple case of only grouping factors the χ^2 ratios are

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

where SS_{effect} is the sum of squares of the considered effect and MS_{total} is the total mean square. The major disadvantage of this method compared with the four ones above is the lack of power for any effect in the case of other nonnull effects in the model. The reason: In the standard anova the denominator of the F values is the residual mean square which is reduced by the effects of other factors in the model. In contrast the denominator of the χ^2 tests of Puri & Sen‘s L statistic is the total mean square which is not diminished by other factors. A good review of articles concerning this test can be found in the study by Toothaker & De Newman (1994).

van der Waerden

At first the van der Waerden test (see Wikipedia and van der Waerden (1953)) is an alternative to the 1-factorial anova by Kruskal-Wallis. The procedure is based on the INT transformation (see above). But instead of using the F-tests from the parametric anova, χ^2 ratios are computed using the sum of squares in the same way as for the Puri & Sen L statistics. Mansouri and Chang (1995) generalized the original van der Waerden test to designs with several grouping factors. Marascuilo and McSweeney (1977) transferred it to the case of repeated measurements. Sheskin (2004) reported that this procedure in the 1-factorial version outperforms the classical anova in the case of violations of the assumptions with regard to the power. On the other hand the van der Waerden tests suffer from the same lack of power in the case of multifactorial designs as the Puri & Sen L statistic.

Akritas, Arnold and Brunner (ATS)

This is the only procedure considered here that cannot be mapped to the parametric anova. Based on the relative effect (see Brunner & Munzel (2002)) the authors developed two tests to compare samples by means of comparing these relative effects: ATS (anova type statistic) and WTS (Wald type statistic). The ATS has preferable attributes e.g. more power (see Brunner & Munzel (2002) as well as Shah & Madden (2004)). The relative effect of a random variable X_1 to a second one X_2 is defined as $p^+ = P(X_1 \le X_2)$, i.e. the probability that X_1 has smaller values than X_2 . As the definition of relative effects is based only on an ordinal scale of the dv this method is suitable also for variables of ordinal or dichotomous scale. The rather complicated procedure is described by Akritas, Arnold and Brunner (1997) as well as by Brunner & Munzel (2002).

3. The Problem

While comparing several nonparametric methods in a Monte Carlo study for several distributions under different conditions in a simple two factor between subjects design it became evident that the type I error rate rose beyond any limit in the case of an underlying discrete distribution for increasing cell counts n_i. But not for all methods, only for the ART- and the ART+INT-procedures (see figure 1). Further simulations showed that the rate increases also if the number of distinct values of the dependent variable is becoming smaller (see figure 2). Only for 2*2 designs there are a couple of situations in which the error rate behaves inconspicuously, e.g. for the test of the main effects in a null model with an underlying uniform distribution (see tables A 7.8.3 and A 7.8.4 in appendix 7).

There is an explanation for the increase of the type I error rate when the number of distinct values gets smaller or the sample size larger: due to the subtraction of the other effects - a linear combination of the means - from the observed values even tiny differences between the means lead to large differences in the ranking. While for instance for the RT-procedure a value *m* is transformed into the same rank R_m for all groups, using the ART-procedure this value *m* is first transformed into different values (m-d_i) depending on the cell *i* and then each (m-d_i) is transformed into a different rank R_{mi} though the differences between the d_i and therefore between the (m-d_i) may be very small. And as there are only a few distinct values, there is a high number of ties causing large differences between the average ranks R_{mi} . So tiny mean differences result in larger differences between the mean ranks and therefore in significant results. And for larger sample sizes this effect is multiplied.

A simple example shall demonstrate this computational procedure. Let us consider a dichotomous dependent variable *y* with values 1 and 2 in a 2*4-design (factors A and B) with equal cell counts n_i=20.

	B_1	B_2	B_3	B_4
A_1	10*1 10*2 $\bar{x} = 1.50$	12*1 8*2 $\bar{x} = 1.40$	13*1 7*2 $\bar{x} = 1.35$	13*1 7*2 $\bar{x} = 1.35$
A_2	7*1 13*2 $\bar{x} = 1.65$	13*1 7*2 $\bar{x} = 1.35$	13*1 7*2 $\bar{x} = 1.30$	12*1 8*2 $\bar{x} = 1.40$

Table 1: data and means

B_1	B_2	B_3	B_4
10*47 10*127 $\bar{x} = 87$	12*47 8*127 $\bar{x} = 79$	13*47 7*127 $\bar{x} = 79$	13*47 7*127 $\bar{x} = 75$
7*47 13*127 $\bar{x} = 99$	13*47 7*127 $\bar{x} = 75$	13*47 7*127 $\bar{x} = 75$	12*47 8*127 $\bar{x} = 79$

Table 2: standard ranks with mean ranks

Table 1 shows the raw data which are transformed into standard ranks (table 2): the 1 into 47 (average rank of 1,..,93) and the 2 into 127 (average rank of 94,..,160).

	B_1	B_2	B_3	B_4
A_1	10*-0.556 10*0.444	12*-0.356 8*0.644	13*-0.331 7*0.669	13*-0.356 7*0.644
A_2	7*-0.594 13*0.406	13*-0.394 7*0.606	13*-0.369 7*0.631	12*-0.394 8*0.606

Table 3: aligned data

B_1	B_2	B_3	B_4
10*12.5 10*111.5 $\bar{x} = 62$	12*68 8*146 $\bar{x} = 99.2$	13*87 7*157 $\bar{x} = 111.5$	13*68 7*146 $\bar{x} = 95.3$
7*4 13*100 $\bar{x} = 66.4$	13*30 7*124 $\bar{x} = 62.9$	13*49 7*135 $\bar{x} = 79.1$	12*30 8*124 $\bar{x} = 67.6$

Table 4: aligned ranks with mean ranks

Table 3 shows the data aligned for the two main effects, computed as

$$x'_{ijk} = e_{ijk} + (ab_{ij} - a_i - b_j + 2\bar{x})$$

or equivalently as

$$x'_{ijk} = x_{ijk} - a_i - b_j + 2\bar{x}$$

where e_{ijk} are the residuals, ab_{ij} the cell means, a_i, b_j the marginal means corresponding to A and B and x the grand mean. In table 4 these values are transformed into ranks. Apparently the differences between the mean ranks are much larger for the ART than for the simple RT-procedure which is confirmed by the anova results (table 5), where additionally to the alignment for the interaction the same was done for the main effects. In opposition to the RT the ART shows both main effects as significant and a comparatively lower p-value for the interaction. This should give an impression how the ART-technique leads to an inflation of the error rates.

effect	parametric	RT	ART A, B aligned	ART A, AB aligned	ART B, AB aligned
A	0.633	0.633			0.037
B	0.152	0.152		0.019	
AB	0.828	0828	0.143		

Table 5: p-values of several anovas applied on the data of table 1

Therefore additional investigations on this relation seemed to be reasonable.

4. The Study

This is a Monte Carlo study. That means a couple of designs and theoretical distributions had been chosen from which a large number of samples had been drawn by means of a random number generator. These samples had been analyzed for the various anova methods. The number of samples had been restricted to 2000, and by means of a unique starting value they are identical for each of the situations below in order to make the results better comparable.

In the current study only grouping (between subjects) factors A and B are considered. It examines:

- Four layouts:
  - a 2*4 balanced design with 10 observations per cell (total n=80) and
  - a 4*5 unbalanced design with an unequal number of observations n_i per cell (total n=100), as well as
  - a 2*2 balanced design with 20 observations per cell (total n=80) and
  - a 2*2 unbalanced design with an unequal number of observations n_i per cell (total n=100).
  For the unbalanced designs there is a ratio max(n_i)/min(n_i) of 4. These differ not only regarding the cell counts but also the number of cells, though the df of the error term in all designs are nearly equal.
- Two underlying distributions:
  - uniform distribution in the interval (0,5) and
  - exponential distribution (parameter λ=0.4) with μ=2.5 which is extremely skewed (S=2).
- Several effect sizes for the main effects as well as for the interaction effect.

(In the following sections the terms *unbalanced design* and *unequal cell counts* will be used both for the corresponding design, being aware that they have different definitions. But the special case of a balanced design with unequal cell counts will not be treated in this study.)

Originally there had been a study in which the seven nonparametric methods listed above had been compared together with the parametric F-test for the uniform and the exponential distribution, both as continuous and as discrete distribution with the values rounded to integers, as well as for 12 other distributions. And the average cell sizes had been varied from 5 to 50 in steps of 5. After it became evident that there are large differences between the error rates for continuous and discrete distributions, especially for large n_i, further investigations seemed to be needed.

Therefore additionally the results for a continuous variable were compared with those of several discrete variables: in case of the uniform distribution with values rounded to integers 1 to 7, 1 to 4 as well as 1 and 2. And in case of the exponential distribution with values were rounded to integers which results in values 1 to 18, transformed to integers 1 to 10 and to integers 1 to 5. Finally, to be able to explain differences in the results for the 2*4 balanced and the 4*5 unbalanced design, additionally a 2*4 unbalanced and a 4*5 balanced design were analyzed.

The main focus had been laid upon the control of the type I error rates for α=0.05 for the various methods and situations. For the computation of the random variates level/cell means had to be added corresponding to the desired effect sizes. These are denoted by a_i and b_j for the level means of A and B corresponding to effects α_i and β_j , and ab_{ij} for the cell means concerning the interaction corresponding to effects $\alpha_i + \beta_j + \alpha\beta_{ij}$. In this study only one moderate effect size had been chosen for each of the models below (see e.g. Danbaba, 2012 and Shoemaker, 1986), though other authors are varying the effect size from small to large values, e.g. from 0.25 up to over 2.0 (see e.g. Salter & Fawcett, 1993 and Mansouri & Chang, 1995). The reason is obvious:

only the type I error rates are investigated and not the power in realtion to the effect size. Furthermore there are so many different influencing factors in this study that a restriction seemed necessary. And the main focus had been laid on the magnitude of n_i and the number of disctinct values of the dv.

For the subsequent specification of the effect sizes the following abbreviations are used (s being the standard deviation):

- A(d):
  $a_1=d*s$, $a_2=0$ for a 2*4 and a 2*2 plan,
  respectively $a_1= a_2= d*s$, $a_3= a_4= 0$ for a 4*5 plan
- B(d):
  $b_1= b_2= d*s$, $b_3= b_4= 0$ for a 2*4 plan,
  $b_1= d*s$, $b_2= 0$ for a 2*2 plan,
  respectively $b_1= b_2= d*s$, $b_3= b_4= b_5= 0$ for a 4*5 plan
- AB(d):
  $ab_{11}= ab_{12}= ab_{23}= ab_{24}= d*s$, $ab_{21}= ab_{22}= ab_{13}= ab_{14}= 0$ for a 2*4 plan,
  $ab_{11}= ab_{22}= d*s$, $ab_{21}= ab_{12}= 0$ for a 2*2 plan,
  respectively $ab_{11}= ab_{12}= ab_{21}= ab_{22}= ab_{34}= ab_{35}= ab_{44}= ab_{45}= d*s$,
  $ab_{31}= ab_{32}= ab_{41}= ab_{42}= ab_{14}= ab_{15}= ab_{24}= ab_{25}= 0$ and $ab_{13}= ab_{23}= ab_{33}= ab_{43}= d*s/2$
  for a 4*5 plan

The error rates had been checked for the following effects:

- main effects and interaction effect for the case of no effects (null model, equal means),
- main effects and interaction effect for the case of one significant main effect A(0.6)
  i.e. a weak impact of significant main effects,
- main effect for the case of a significant interaction AB(0.6)
  i.e. a weak impact of significant interaction effect,
- main effect for the case of a significant main and interaction effect A(0.6) and AB(0.6)
  i.e. a weak impact of significant effects.
- interaction effect for the case of both significant main effects A(0.8) and B(0.8)
  i.e. a strong impact of significant main effects.

These are 7 effect models which are analysed for both a balanced and an unbalanced design. So there are all in all 14 models which are tested for a small 2*2 design as well as for larger 2*4 and 4*5 designs.

In case of effects the uniform distribution has been transformed that $x+d*s$ lies still in the given interval.

In the case of the exponential distribution for the analysis of the influence of effects d it is not reasonable to add a constant $d*s$ to the values x of one group. In order to keep the type of exponential distribution with parameter λ for the alternative hypothesis (H_1) a parameter λ‘ had to be chosen so that the desired mean difference $1/\lambda - 1/\lambda'$ is $d*s$ where in this case $s=(1/\lambda + 1/\lambda')$. As a consequence the H_1-distribution has a larger variance.

In the original study the error rates were computed for n_i varying from 5 to 50 in steps of 5, whereas for the detailed comparison of continuous and discrete distributions the rates were computed only for $n_i = 5$, 10 and 50.

5. Results

All tables and corresponding graphical illustrations are available online:
http://www.uni-koeln.de/~luepsen/statistik/texte/comparison-tables/
Each table and graphic includes the results for all 8 methods and report the proportions of rejections of the corresponding null hypothesis:

- Appendix 2: type I error rates (α=0.05) for large n_i (5 to 50 in steps of 5) and fixed effect sizes, for all 16 distributions, 7 different effect models, equal and unequal cell frequencies.
- Appendix 7: type I error rates (α=0.05) for n_i = 5, 10, 50 and fixed effect sizes, for the exponential and the uniform distributions with the continuous and several discrete versions, 7 different effect models, equal and unequal cell frequencies.

All references to these tables and graphics will be referred as A *n.n.n*. The most important tables and some graphics are included in this text.

Concerning the type I error rate a deviation of 25 percent ($\alpha + 0.25\alpha$) - that is 6.25 percent for α=0.05 - can to be treated as a moderate robustness (see Peterson (2002). It should be mentioned that there are other studies in which a deviation of 50 percent, i.e. ($\alpha \mp 0.5\alpha$), Bradleys liberal criterion (see Bradley, 1978), is regarded as robustness. 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.

Comparing all 8 methods with regard to the behaviour in the case of underlying discrete distributions the tables and graphics in Appendix 2 show that the type I error rates rise only for the ART- and the ART+INT-procedures for increasing cell counts n_i, in most cases beyond 10 percent, but sometimes even up to 20 percent (see e.g. A 2.4.6) or 50 percent (see A 2.4.9). But at second sight it must be admitted that there are a couple of situations where the rates rise also for the corresponding continuous distribution. But in any case the error rates for the discrete distribution lie considerably above those for the continuous distribution, on average between 10 and more than 100 percent. With one exception: the RT- as well as the ATS-method exhibit increasing error rates for the interaction in the case of a discrete exponential distribution and both significant main effects (see A 2.13.6 and 2.14.6).

Concerning the exponential distribution the rates of the ART and the ART+INT are under control for the tests in the null model with a continuous distribution, but exceed the acceptable limit with values of 8 percent and above for $n_i \geq 20$ in the case of a discrete distribution. For the tests in the nonnull models the values increase in any case above the limit of about 6 percent, while the rates for the discrete distribution exceed those for the continuous ditribution by approximately 10 percent. See table 6 for details which represents a summary of the results for the ART-method tabulated in A 2.

In the case of the uniform distribution the situation is more transparent because for the continuous distribution the error rates are always under control except for the test of a main effect if the other main effect is nonnull. For all other models the rates for the discrete distribution stay below 6 percent as long as $n_i \leq 15$ and rise up to values between 6 and 8 if n_i increases up to 50.

But it has to be noted that at least for equal cell counts the rates keep acceptable for most models, especially for the test of the interaction, though they lie between 10 and 20 percent above those for the continuous distribution. See table 7 for details.

Having now shown that discrete distributions have an impact on the error rates for the ART- and the ART+INT-method, the focus of the second part of the study lies on the question: how large is the effect of the number of distinct values of the dependent variable. And the target is here the case of small cell counts, as for large cell counts the impact has been already proven for any situation. Here a general tendency is obvious: for a decreasing number of distinct values the type I error rates for the ART- and the ART+INT-method rise.

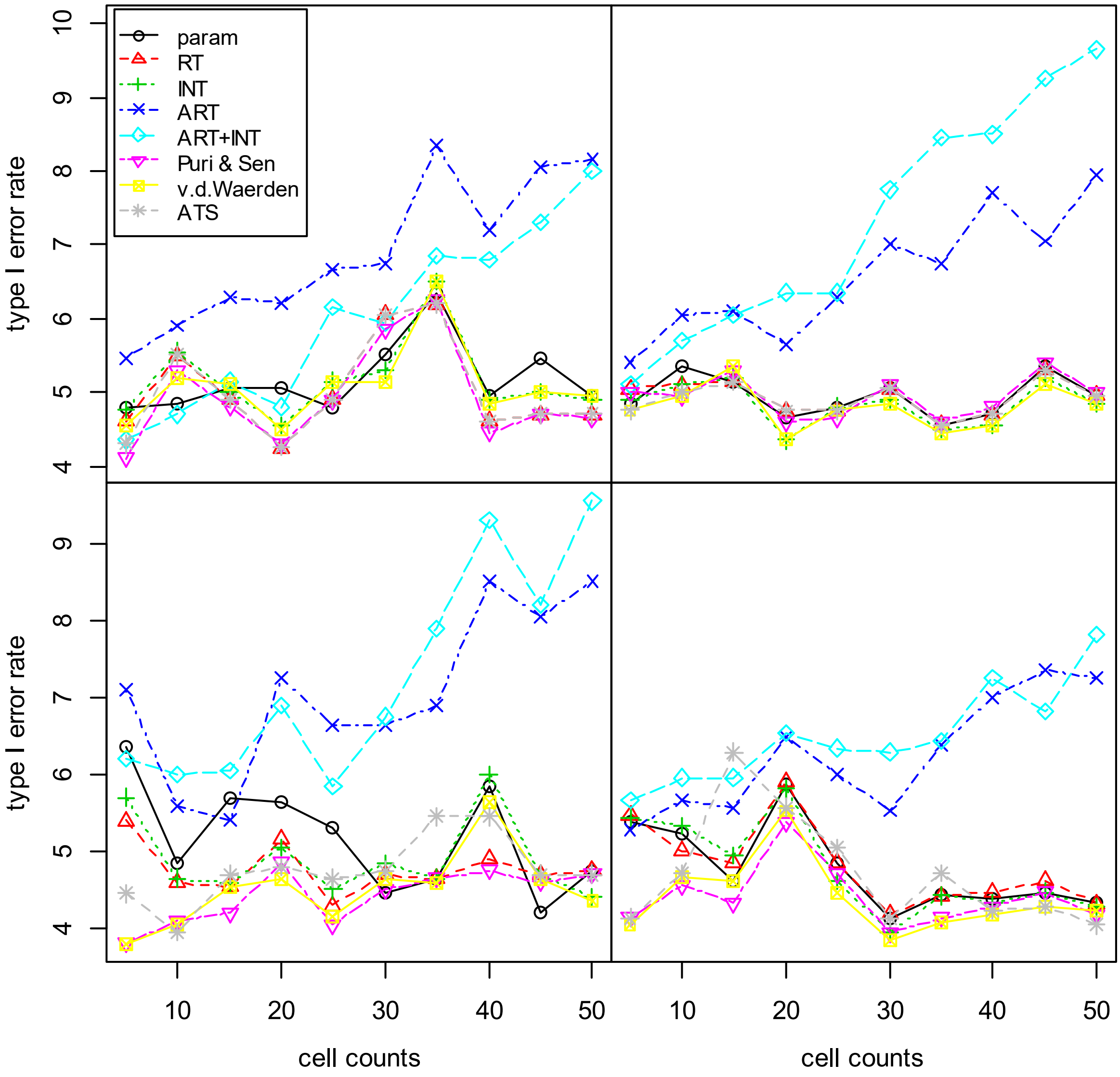

Figure 1: type I error rates for 8 methods in the case of the null model,
1st row: main effect A in a balanced design,
2nd row: interaction effect in an unbalanced design
left: exponential distribution, right: uniform distribution

As in the analysis above, for increasing n_i here also the behaviour of the exponential distribution is more critical. In any of the 14 models the rates for the ART increase to values between 10 and 20 percent for the cases of 10 or 5 distinct values. But the additional application of the INT-transformation to the ART-method is able to damp this effect though there are still a couple of situations where the rates for the ART+INT rise to values between 7 and 10, e.g. for the main and interaction effect in the null model (see A 7.1 and 7.2 as well as A 7.9 and 7.10) and for the main effect if the interaction is significant (see A 7.5 and 7.6). See table 8 for details.

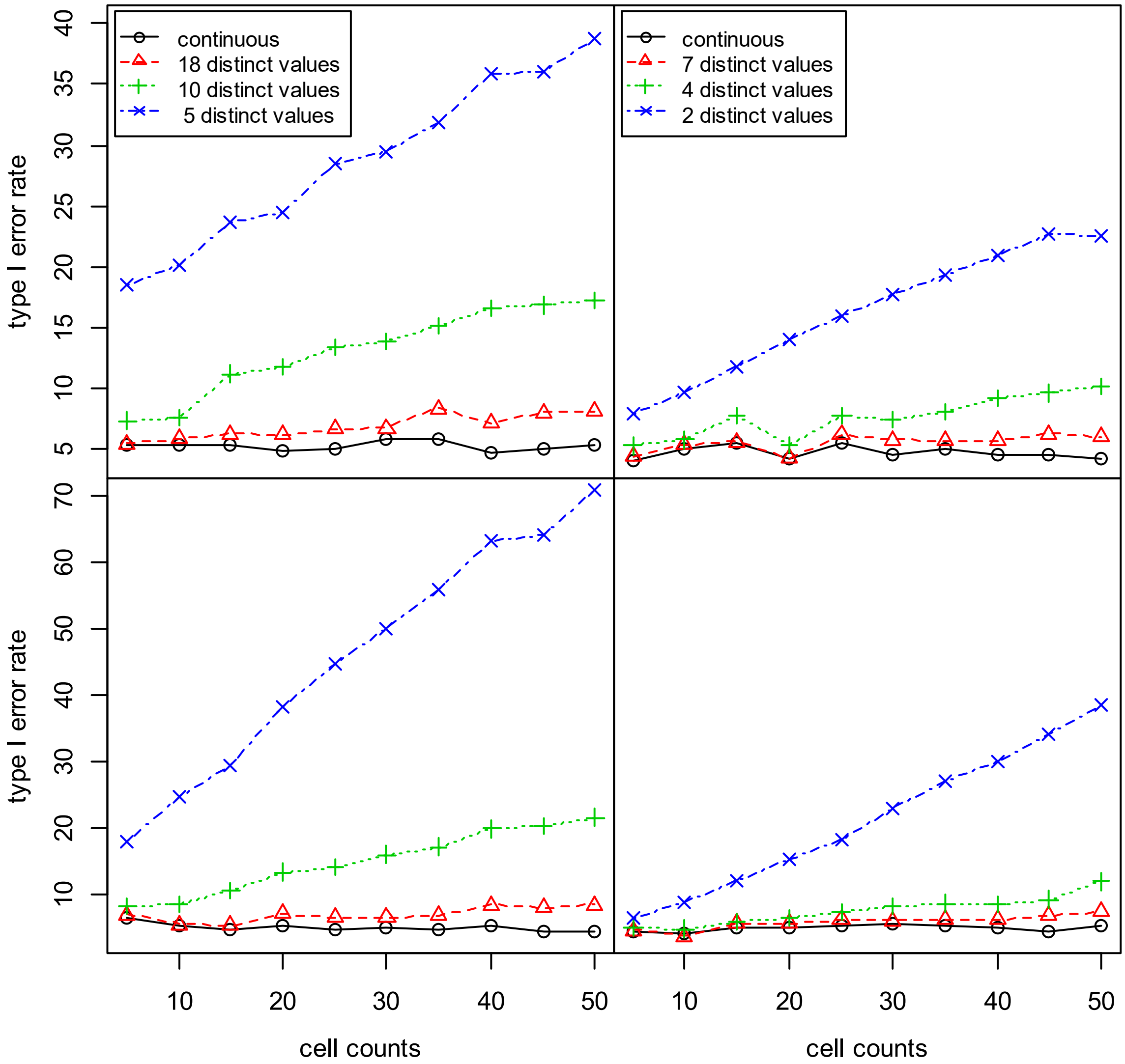

Figure 2: type I error rates for 4 different scalings of the dependent variable in the case of the null model,
1st row: main effect A in a balanced design,
2nd row: interaction effect in an unbalanced design
left: exponential distribution, right: uniform distribution

In contrast for the uniform distribution the error rates exceed the acceptable limit with values between 8 and 10 percent only when the dependent variable is dichotomous. And this only in 6 of the 14 models. Furthermore it is remarkable that the rates for the unbalanced 4*5-design lie below those for the balanced 2*4-design. As the designs differ not only with regard to the balance but also to the number of cells, additionally a balanced 4*5- and an unbalanced 2*4-design were studied. Their results lie in between. See table 9 for details.

6. Conclusion

As the type I error rates for the ART- as well as for the ART+INT-method rise above any limit with increasing cell counts n_i in the case of discrete dependent variables a general warning has to be stated for this combination. For exponential distributed discrete variables this is more severe because even for small cell counts (5 or 10) the rates lie often beyond an acceptable limit of 6 percent whereas for uniform distributed discrete variables the error rates are under control at least for cell counts of 15 and below.

The behaviour of the error rate worsens when the number of distinct values of the dependent variable decreases. Here also the situation is more severe for the exponential distribution. For 5 or 10 distinct values the rates of the ART lie above 10 percent. In contrast for the uniform distribution only the case of 2 distinct values is critical.

Therefore the application of the ART and the ART+INT-methods should be restricted to the cases when first

- the cell frequencies stay below 15

and additionally

- either the underlying distribution is approximately uniform,
- or the number of distinct values exceeds 10.

The application of the INT-transformation to the ART is able to damp the high error rates when there are fewer distinct values.

The conclusion: The ART as well as the ART+INT cannot be applied to Likert and similar metric or ordinal scaled variables, e.g. frequencies like the number of children in a family or the number of goals, or ordinal scales with ranges from 1 to 5.

Though between subject designs are the basis for this study it is to be expected that the results are also valid for designs with repeated measurements such as split plot designs.

7. Software

This study has been programmed in R (version 3.0.2), using mainly the standard anova function `aov` in combination with `drop1` to receive type III sum of squares estimates in the case of unequal cell counts. For the ART, ATS, factorial Puri & Sen and van der Waerden methods own functions had been written (see Luepsen, 2014). All the computations had been performed on a Windows notebook.

8. Tables

			cell count									
method	count	scale	5	10	15	20	25	30	35	40	45	50
Main effect A - null model	equal	cont.	5.40	5.40	5.35	4.90	5.00	5.80	5.80	4.65	5.00	5.30
		discr.	5.45	5.90	6.30	6.20	6.65	6.75	8.35	7.20	8.05	8.15
	unequal	cont.	4.15	3.05	3.70	2.65	3.95	3.45	3.50	3.60	3.20	4.05
		discr.	4.25	3.50	4.40	3.70	4.95	5.20	5.15	5.90	5.65	6.50
Main effect B - A significant	equal	cont.	6.00	7.05	7.70	7.70	7.20	7.60	6.40	6.90	8.40	8.90
		discr.	6.25	7.25	7.80	7.90	7.85	7.95	7.30	7.35	8.70	9.15
	unequal	cont.	6.45	7.60	8.00	10.50	13.20	13.65	15.50	17.45	18.25	19.50
		discr.	6.60	8.40	8.35	11.70	13.90	15.40	16.65	20.00	20.85	22.50
Main effect A - interact. sig	equal	cont.	6.45	5.85	7.30	7.05	7.45	6.10	6.00	6.70	6.40	6.60
		discr.	6.65	7.00	8.10	9.30	8.35	8.95	8.25	10.25	9.60	10.00
	unequal	cont.	6.20	5.40	7.35	6.80	6.85	8.45	9.10	9.50	10.30	12.15
		discr.	6.25	6.00	8.45	7.80	8.20	10.80	11.95	13.75	13.40	16.85
Main effect B - A and int. sig	equal	cont.	6.65	6.30	6.60	7.15	7.85	7.40	7.45	7.30	7.20	6.50
		discr.	7.05	6.40	7.30	7.70	7.90	8.10	7.65	7.55	7.25	7.15
	unequal	cont.	6.85	6.35	7.25	8.00	8.60	9.50	9.70	11.65	11.65	13.40
		discr.	7.00	6.75	8.05	8.35	9.40	11.55	12.30	13.00	14.80	16.90
Interaction - null model	equal	cont.	5.85	5.70	5.65	5.75	5.60	5.60	5.60	5.90	5.75	6.00
		discr.	6.00	6.15	6.25	6.30	6.15	6.80	6.75	7.00	6.70	6.85
	unequal	cont.	6.65	5.35	4.85	5.55	4.70	5.20	4.85	5.55	4.65	4.65
		discr.	7.10	5.60	5.40	7.25	6.65	6.65	6.90	8.50	8.05	8.50
Interaction - A significant	equal	cont.	7.65	7.90	8.40	7.40	8.50	7.20	8.55	8.25	7.80	7.40
		discr.	8.20	7.65	8.95	8.25	9.25	7.60	9.25	8.70	8.65	9.40
	unequal	cont.	5.10	6.40	5.65	5.90	4.60	5.30	5.90	4.95	6.00	5.20
		discr.	5.80	6.65	6.25	6.20	5.85	5.95	6.60	5.90	6.95	7.05
Interaction - A and B sig	equal	cont.	7.10	8.15	8.05	7.85	7.65	8.90	7.65	9.25	8.55	9.15
		discr.	7.55	8.05	8.45	8.30	8.10	9.60	8.10	9.25	9.05	9.10
	unequal	cont.	5.30	5.65	7.10	6.10	6.90	7.15	6.80	7.60	7.55	6.70
		discr.	5.25	6.05	6.50	6.05	6.75	7.90	7.75	7.95	8.35	7.75

Table 6: type I error rates of the ART-method for 7 different effect models, equal and unequal cell counts and two different scalings of the exponential distribution (continuous and discrete). (This is a summary of the results for the ART-method tabulated in Appendix 2.)

			cell count									
method	count	scale	5	10	15	20	25	30	35	40	45	50
Main effect A - null model	equal	cont.	5.05	5.40	5.40	4.75	4.65	5.05	4.40	4.20	5.75	5.15
		discr.	5.40	6.05	6.10	5.65	6.30	7.00	6.75	7.70	7.05	7.95
	unequal	cont.	2.70	2.30	2.60	2.25	2.60	2.55	2.35	2.90	2.30	2.40
		discr.	2.85	2.40	3.40	3.10	3.30	3.90	4.30	4.90	4.85	5.05
Main effect B - A significant	equal	cont.	4.60	4.55	4.00	4.65	4.60	5.55	5.40	4.65	4.00	5.05
		discr.	4.40	4.50	4.30	4.65	5.05	5.40	5.00	5.10	4.55	5.70
	unequal	cont.	3.50	4.20	4.80	5.10	7.15	7.80	9.00	8.90	9.55	11.05
		discr.	3.75	4.65	6.25	7.10	10.05	12.20	15.75	22.25	25.10	32.80
Main effect A - interact. sig	equal	cont.	4.45	5.15	4.85	4.90	4.75	5.35	5.30	4.90	4.75	4.45
		discr.	5.10	5.25	5.90	5.80	5.95	6.70	6.60	6.90	8.15	6.80
	unequal	cont.	3.05	2.85	2.80	3.50	3.20	4.00	3.75	4.80	4.90	5.30
		discr.	3.20	3.45	2.90	3.45	4.65	4.65	4.80	6.15	5.55	5.90
Main effect B - A and int. sig	equal	cont.	4.60	4.55	4.00	4.65	4.60	5.55	5.40	4.65	4.00	5.05
		discr.	4.40	4.40	4.50	5.05	5.45	5.70	5.10	4.90	4.20	5.70
	unequal	cont.	3.40	3.40	3.60	3.60	3.90	4.15	4.40	4.90	4.50	5.55
		discr.	3.50	3.75	3.65	4.05	4.65	4.80	4.95	6.05	5.30	6.15
Interaction - null model	equal	cont.	5.25	5.30	4.10	5.00	5.05	4.60	4.70	4.25	5.20	4.55
		discr.	5.75	4.50	4.85	5.00	5.95	5.15	4.85	5.40	6.30	6.15
	unequal	cont.	5.50	5.55	5.30	5.80	4.35	4.45	4.60	4.20	4.65	4.40
		discr.	5.55	5.95	5.85	6.80	6.30	5.80	6.70	7.35	7.70	7.60
Interaction - A significant	equal	cont.	5.25	5.30	4.10	5.00	5.05	4.60	4.70	4.25	5.20	4.55
		discr.	5.75	5.00	4.75	5.25	5.85	4.85	5.05	5.35	5.60	5.90
	unequal	cont.	5.50	5.65	5.05	5.80	4.35	4.50	4.50	4.30	4.60	4.60
		discr.	5.55	6.00	6.40	7.45	6.30	6.10	6.85	6.55	7.00	7.20
Interaction - A and B sig	equal	cont.	5.25	5.30	4.10	5.00	5.05	4.60	4.70	4.25	5.20	4.55
		discr.	5.75	4.50	4.85	5.00	5.95	5.15	4.85	5.40	6.30	6.15
	unequal	cont.	5.10	5.60	4.95	5.90	4.40	4.40	4.60	4.25	4.60	4.45
		discr.	5.65	6.10	5.60	7.00	6.65	6.05	7.25	7.65	7.45	8.05

Table 7: type I error rates of the ART-method for 7 different effect models, equal and unequal cell counts and two different scalings of the uniform distribution (continuous and discrete). (This is a summary of the results for the ART-method tabulated in Appendix 2.)

			number of discrete values							
			n = 5 / n = 10				n = 50			
method	count	design	cont.	18	10	5	cont.	18	10	5
Main effect A - null model	equal	2*4	5.40	5.90	7.65	20.20	5.30	8.15	17.25	38.75
		4*5	6.90	7.45	8.60	27.60	4.95	6.60	19.80	55.45
	unequal	2*4	4.65	4.65	7.15	17.05	4.65	8.25	19.70	48.35
		4*5	4.15	4.25	4.80	19.05	4.05	6.50	25.45	72.25
Main effect B - A significant	equal	2*4	7.05	7.25	7.25	10.90	8.90	9.15	13.80	24.15
		4*5	9.15	9.60	11.00	22.80	9.70	12.05	25.90	60.45
	unequal	2*4	6.45	7.00	9.20	14.15	17.70	22.90	40.95	73.85
		4*5	6.45	6.60	7.25	14.40	19.50	22.50	37.85	69.30
Main effect A - interact. sig	equal	2*4	5.85	7.00	10.45	23.55	6.60	10.00	20.70	39.55
		4*5	7.65	8.50	10.95	33.55	8.25	10.90	23.80	59.20
	unequal	2*4	6.40	7.35	11.05	25.30	25.20	32.25	44.60	69.70
		4*5	6.20	6.25	8.35	29.20	12.15	16.85	35.90	79.05
Main effect B - A and int. sig	equal	2*4	7.45	7.80	9.20	11.10	8.10	8.80	10.05	26.45
		4*5	9.55	10.20	12.10	23.95	9.95	12.10	18.65	67.70
	unequal	2*4	7.25	7.85	9.25	16.60	13.10	15.45	15.45	54.10
		4*5	8.05	8.75	9.40	15.85	17.40	18.95	25.55	57.45
Interaction - null model	equal	2*4	5.70	6.15	6.90	12.55	6.00	6.85	12.40	31.85
		4*5	6.60	7.05	8.65	15.55	4.95	7.90	22.60	70.95
	unequal	2*4	5.75	6.15	8.35	14.70	5.10	7.20	13.60	37.40
		4*5	6.65	7.10	8.40	17.95	4.65	8.50	21.70	71.10
Interaction - A significant	equal	2*4	7.90	7.65	9.20	14.60	7.40	9.40	14.70	26.40
		4*5	8.70	8.80	9.90	16.35	7.80	10.00	18.60	44.90
	unequal	2*4	6.55	7.65	9.10	13.20	8.05	8.65	14.70	24.15
		4*5	5.10	5.80	6.35	11.30	5.20	7.05	12.30	18.60
Interaction - A and B sig	equal	2*4	8.15	8.05	8.85	11.35	9.15	9.10	11.60	21.10
		4*5	8.80	8.80	9.30	12.90	9.60	10.85	19.35	40.50
	unequal	2*4	10.00	10.40	11.25	15.75	9.90	11.45	13.55	29.10
		4*5	5.30	5.25	5.70	7.75	6.70	7.75	11.05	35.35

Table 8: type I error rates of the ART-method for 7 different effect models, equal and unequal cell counts, two different designs and 4 different scalings of the exponential distribution (continuous and discrete) for small cell counts (5 and 10) and large cell counts (50).

			number of discrete values							
			n = 5 / n= 10				n = 50			
method	count	design	cont.	7	4	2	cont.	7	4	2
Main effect A - null model	equal	2*4	5.05	5.40	5.80	9.65	4.20	6.0	10.15	22.60
		4*5	4.85	5.20	5.60	5.75	4.85	5.95	9.45	28.95
	unequal	2*4	3.20	3.70	4.25	7.05	2.80	4.95	9.30	29.05
		4*5	2.20	2.25	2.80	3.00	2.30	3.65	7.20	38.45
Main effect B - A significant	equal	2*4	4.45	4.70	5.45	6.65	5.15	5.30	5.15	8.30
		4*5	6.15	6.25	6.50	9.05	4.15	5.80	7.60	23.95
	unequal	2*4	5.55	6.20	7.20	10.70	13.30	21.85	29.65	60.50
		4*5	3.70	3.85	4.25	5.25	11.15	14.80	23.55	42.85
Main effect A - interact. sig	equal	2*4	5.05	4.85	5.85	8.85	4.20	6.55	6.60	14.05
		4*5	4.85	5.40	5.40	7.05	4.85	5.50	7.00	16.20
	unequal	2*4	4.80	4.60	4.85	9.10	12.15	14.40	13.05	40.55
		4*5	2.70	2.40	2.85	3.60	4.25	5.70	6.75	29.30
Main effect B - A and int. sig	equal	2*4	4.45	4.40	4.75	5.45	5.15	4.85	5.90	8.10
		4*5	6.15	5.90	6.00	7.00	4.15	5.65	6.30	13.45
	unequal	2*4	5.15	5.65	6.00	8.30	11.45	15.30	7.85	18.40
		4*5	3.80	4.15	4.05	5.15	8.80	11.00	12.25	12.90
Interaction - null model	equal	2*4	6.05	5.85	6.25	7.70	4.75	5.25	6.65	14.60
		4*5	5.35	5.65	5.35	7.40	5.00	6.10	9.10	35.75
	unequal	2*4	4.65	5.05	5.15	6.35	4.95	6.10	8.00	20.70
		4*5	4.40	4.70	5.05	6.45	5.55	7.50	12.05	38.75
Interaction - A significant	equal	2*4	6.05	6.00	5.95	9.05	4.75	5.65	5.50	9.20
		4*5	5.35	5.10	5.45	7.50	5.00	5.80	5.85	16.10
	unequal	2*4	4.30	4.80	4.85	7.75	5.00	5.80	7.75	10.10
		4*5	4.40	4.25	4.50	5.45	5.50	6.90	8.20	9.80
Interaction - A and B sig	equal	2*4	6.05	6.15	6.05	8.15	4.75	5.20	6.50	11.35
		4*5	5.35	5.6	5.30	7.45	5.00	5.30	9.60	16.45
	unequal	2*4	4.30	4.60	5.35	6.55	4.75	6.30	9.90	11.45
		4*5	4.60	4.95	4.85	5.15	5.60	6.50	12.25	14.95

Table 9: type I error rates of the ART-method for 7 different effect models, equal and unequal cell counts, two different designs and 4 different scalings of the uniform distribution (continuous and discrete) for small cell counts (5 and 10) and large cell counts (50).

9. Literature

Akritas, M.G., Arnold, S.F., Brunner, E. (1997). Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs, *Journal of the American Statistical Association*, Volume 92, Issue 437, pp 258-265.

Beasley, T.M., Zumbo, B.D. (2009). Aligned Rank Tests for Interactions in Split- Plot Designs: Distributional Assumptions and Stochastic Heterogeneity, *Journal of Modern Applies Statistical Methods,* Vol 8, No 1, pp 16-50.

Beasley, T.M., Erickson, S., Allison, D.B. (2009). Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited? *Behavourial Genetics*, 39 (5), pp 380-395.

Bennett, B.M. (1968). Rank-order tests of linear hypotheses, *Journal of Stat . Society B* 30, pp 483- 489.

Blair, R.C., Sawilowsky, S.S., Higgins, J.J. (1987). Limitations of the rank transform statistic, *Communications Statististics,* B 16, pp 1133-45.

Bradley, J.V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, pp 144-152.

Brunner, E., Munzel, U. (2002). *Nichtparametrische Datenanalyse - unverbundene Stichproben*, Springer, Berlin.

Carletti, I. , Claustriaux, J.J. (2005). Anova or Aligned Rank Transform Methods: Which one use when Assumptions are not fulfilled ? *Buletinul USAMV-CN*, nr. 62/2005 and below, ISSN, pp 1454-2382.

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

Danbaba, A. (2009). A Study of Robustness of Validity and Efficiency of Rank Tests in AMMI and Two-Way ANOVA Tests. Thesis, University of Ilorin, Nigeria

Danbaba, A. (2012). Comparison of a Class of Rank-Score Tests in Two-Factor Designs. *Nigerian Journal of Basic and Applied Science*, 20 (4), pp 305-314

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

Hodges, J.L. and Lehmann, E.I. (1962). Rank methods for combination of independent experiments in analysis of variance. *Annals of Mathematical Statistics* 27, pp 324-335.

Hora, S.C., Conover, W.J. (1984). The F Statistic in the Two-Way Layout with Rank-Score Transformed Data, *Journal of the American Statistical Association*, Vol. 79, No. 387, pp. 668-673.

Huang, M.L. (2007). A Quantile-Score Test for Experimental Design. *Applied Mathematical Sciences,* Vol. 1, No 11, pp 507-516.

Kaptein, M., Nass, C., Markopoulos, P. (2010). *Powerful and Consistent Analysis of Likert-Type Rating Scales*. CHI 2010: 1001 Users, April 10–15, 2010, Atlanta, GA.

Leys, C., Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology*, Volume 46, Issue 4, July

2010, Pages 684–688

Lei, X., Holt, J.K., Beasley,T.M. (2004). Aligned Rank Tests As Robust Alternatives For Testing Interactions In Multiple Group Repeated Measures Designs With Heterogeneous Covariances. *Journal of Modern Applied Statistical Methods*, 2004, Vol 3, Issue 2.

Luepsen, H (2014). *R-Funktionen zur Varianzanalyse*.
URL: http://www.uni-koeln.de/~luepsen/R/ .

Mansouri, H. , Chang, G.-H. (1995). A comparative study of some rank tests for interaction . *Computational Statistics & Data Analysis* 19 (1995) 85-96 .

Mansouri, H. , Paige, R., Surles, J. G. (2004). Aligned rank transform techniques for analysis of variance and multiple comparisons. Missouri University of Science and Technology *Communications in Statistics - Theory and Methods* - Volume 33, Issue 9.

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

Peterson, K. (2002). Six Modifications Of The Aligned Rank TransformTest For Interaction. *Journal Of Modem Applied Statistical Methods*. Vol. 1, No. 1, pp 100-109.

Puri, M.L. & Sen, P.K. (1985). *Nonparametric Methods in General Linear Models*. Wiley, New York.

R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/ .

Richter, S.J. and Payton, M. (1999). Nearly exact tests in factorial experiments using the aligned rank transform. *Journal of Applied Statistics*, Volume 26, Issue 2.

Salter, K.C. and Fawcett, R.F. (1993). The art test of interaction: A robust and powerful rank test of interaction in factorial models. *Communications in Statistics: Simulation and Computation* 22 (1), pp 137-153.

Scheirer, J., Ray, W.J., Hare, N. (1976). The Analysis of Ranked Data Derived from Completely Randomized Factorial Designs. *Biometrics*. 32(2). International Biometric Society, pp 429−434.

Shah, D. A., Madden, L. V. (2004). Nonparametric Analysis of Ordinal Data in Designed Factorial Experiments . *The American Phytopathological Society*, Vol. 94, No. 1, pp 33 - 43.

Sheskin, D.J. (2004). *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall.

Shoemaker, L.H. (1986). A Nonparametric for Analysis of Variance. *Communications in Statistics: Simulation and Computation* 15 (3), pp 609-632.

Shirley, E.A. (1981). A distribution-free method for analysis of covariance based on ranked data. *Journal of Applied Statistics* 30: 158-162.

Thomas, J.R., Nelson, J.K. and Thomas, T.T. (1999). A Generalized Rank-Order Method for Nonparametric Analysis of Data from Exercise Science: A Tutorial. *Research Quarterly for Exercise and Sport, Physical Education, Recreation and Dance,* Vol. 70, No. 1, pp 11-23.

Toothaker, L.E. and De Newman (1994). Nonparametric Competitors to the Two-Way

ANOVA. *Journal of Educational and Behavioral Statistics*, Vol. 19, No. 3, pp. 237-273.

van der Waerden, B.L. (1953). *Order tests for the two-sample problem. II, III*, Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Serie A, 564, pp 303–310 and pp 311–316.

Wikipedia. URL: http://en.wikipedia.org/wiki/Van_der_Waerden_test .

Wobbrock, J. O., Findlater, L., Gergle, D. & Higgins, J. (2011). The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures. *Computer Human Interaction - CHI* , pp. 143-146.