Dieter Rasch and Thomas Rusch and Marie Simeckova and Klaus D. Kubinger and Karl Moder and Petr Simecek

Tests of additivity in mixed and fixed effect two-way ANOVA models with single sub-class numbers

Tests of additivity in mixed and fixed effect two-way ANOVA models with single sub-class numbers

Dieter Rasch
University of Natural Resources and Applied Life Sciences
Vienna

Thomas Rusch
WU (Wirtschafts-universität Wien)

Marie Šimečková
Biometric Unit
Prague

Klaus D. Kubinger
University of Vienna

Karl Moder
University of Natural Resources and Applied Life Sciences
Vienna

Petr Šimeček
Biometric Unit
Prague

Abstract

In variety testing as well as in psychological assessment, the situation occurs that in a two-way ANOVA-type model with only one replication per cell, analysis is done under the assumption of no interaction between the two factors. Tests for this situation are known only for fixed factors and normally distributed outcomes. In the following we will present five additivity tests and apply them to fixed and mixed models and to quantitative as well as to Bernoulli distributed data. We consider their performance via simulation studies with respect to the type-I-risk and power. Furthermore, two new approaches will be presented, one being a modification of Tukey's test and the other being a new experimental design to test for interactions.

Keywords: Additivity tests, Two-way ANOVA without replication, Mixed model, block design, Rasch model.

1. Introduction

Two-way ANOVA-type models are a well known class of linear models that allow estimation and testing of two main effects and one interaction effect. Usually, the number of replications per factor level combination or cell is greater than one, which enables estimation of the main effects and the interaction effect simultaneously. If the number of replications in each cell is equal to one, the classic way of estimating or testing the interaction effect is not applicable anymore. A number of solutions to the test problem have been developed - the most prominent by Tukey (1949) - and will be presented subsequently, as well as a modification of Tukey's test and a new way of designing experiments to test for interaction.

Unfortunately, all of these developments apply to the case of fixed effects and normally distributed data. Since many possible applications correspond to mixed effect models, it is necessary to find out if the proposed additivity tests can be used with mixed models as well and, if so, how powerful they are. Therefore results concerning the robustness and power of

This is an electronic version of an article published in *Statistical Papers, 50(4), 905-916* doi:10.1007/s00362-009-0254-4. The final publication is available at at http://link.springer.com/article/10.1007/s00362-009-0254-4

additivity tests in mixed models will be presented in the following, as well as results concerning robustness and power of additivity tests in the case of binary data.

In particular, two applications motivated the use of additivity tests for (generalized) mixed models. Mainly, in variety testing, a relatively large number a (30 or more) of varieties has to be tested in b blocks, where each block contains a relatively small number k of varieties. In case of $a = k$ results a complete block design and usually there is only one observation for each variety $\times$ block combination. The varieties can be considered as levels of a fixed factor, whereas the blocks are usually considered as randomly selected from a population of possible blocks. One is interested in the effect of the varieties and wants to know if there is reason to suspect that an interaction of block and variety is present. The second application arises in psychology, in the calibration phase of psychologic-diagnostical instruments, where a relatively large number a (50 or more) of items are presented to a large number b (100 or more) of individuals or testees. As a result there is only one observation per item $\times$ person combination, a correct (=1) or wrong (=0) answer. The items can be considered as levels of a fixed factor, whereas the people are selected randomly from a population of possible testees. The dependent variable is binary. One is interested if a certain model - the *Rasch* model (Rasch 1960) - holds, which requires the absence of interaction between items and people.

The fixed effects model is (random variables are bold print)

$$\boldsymbol{y}_{ij} = \mu + a_i + b_j + (ab)_{ij} + \boldsymbol{e}_{ij} \quad (i = 1, \ldots, a; j = 1, \ldots, b) \tag{1}$$

with the side conditions $\sum_{i=1}^{a} a_i = 0$, $\sum_{j=1}^{b} b_j = 0$, and $\sum_i (ab)_{ij} = \sum_j (ab)_{ij} = 0$. The $(ab)_{ij}$ stand for the interactions and at this stage are deliberately not assumed to have a certain structure. The error term $\boldsymbol{e}_{ij}$ is distributed as $N(0, \sigma^2)$.

For the mixed model we are concerned with

$$\boldsymbol{z}_{ij} = \mu + a_i + \boldsymbol{b}_j + (\boldsymbol{ab})_{ij} \quad (i = 1, \ldots, a; j = 1, \ldots, b) \tag{2}$$

where $E(\boldsymbol{y}_{ij}) = \boldsymbol{z}_{ij}$ and $\boldsymbol{y}_{ij}$ are the normally distributed observations in case of variety testing and $E(\boldsymbol{y}_{ij}) = g^{-1}(\boldsymbol{z}_{ij})$, where $\boldsymbol{y}_{ij} \in \{0, 1\}$ and $g(.)$ is a known link function (e.g. the logit link) in case of binary data. Furthermore we have the side conditions $\sum_{i=1}^{a} a_i = 0$, $Var(\boldsymbol{b}_j) = \sigma_b^2$, and $Var(\boldsymbol{ab})_{ij} = \sigma_{ab}^2$. All random effects are assumed to be from an exponential family.

2. Fixed Effect Models

2.1. Additivity tests

Several tests for model (1) with normally distributed errors have been developed over the years. Five of them will be used here, namely the tests by Tukey (1949), Mandel (1961), Johnson and Graybill (1972), Tusell (1990, 1992) and Boik (1993). A review on additivity tests is given by Karabatos (2005) and Alin and Kurt (2006). Subsequently the following notation will be used: Let $\bar{y}_{..}$ denote the overall mean and $\bar{y}_{.j}$ and $\bar{y}_{i.}$ the respective j-th column and i-th row means. The matrix R will denote a residual matrix with respect to the main effects and comprises of the entries $r_{ij} = y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} - \bar{y}_{..}$. The decreasingly ordered list of eigenvalues of RR^T will be denoted by $\lambda_1 \geq \lambda_2 \geq \ldots$ and their scaled analogues as $\omega_l = \lambda_l / \sum_s \lambda_s, l = 1, 2, \ldots$

Tukey's Test (1949) Tukey suggested to estimate row and column effects first and test for interaction of the type $(ab)_{ij} = k \cdot a_i \cdot b_j$, with $k = 0$ referring to the situation of no interaction (additivity). The test statistic S_T is

$$S_T = MS_{int}/MS_{error}, \tag{3}$$

with

$$MS_{int} = \frac{\left(\sum_i \sum_j y_{ij}(\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{.j} - \bar{y}_{..})\right)^2}{\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2} \tag{4}$$

and

$$MS_{error} = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 - a\sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 - b\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 - MS_{int}}{(a-1)(b-1) - 1} \tag{5}$$

S_T is F distributed with 1 and $(a-1)(b-1)-1$ degrees of freedom under the null hypothesis of no interaction.

Mandel's Test (1961) Mandel generalized the approach of Tukey (1949) and derived a test for the interaction $(ab)_{ij} = c_i \cdot b_j$ with c_i being a certain row constant. He defined the test statistic S_M to test for $c_i = 0$ as

$$S_M = \frac{\sum_i (z_i - 1)^2 \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2}{a-1} \Big/ \frac{\sum_i \sum_j ((y_{ij} - \bar{y}_{i.}) - z_i(\bar{y}_{.j} - \bar{y}_{..}))^2}{(a-1)(b-2)}, \tag{6}$$

with

$$z_i := \frac{\sum_j y_{ij}(\bar{y}_{.j} - \bar{y}_{..})}{\sum_j (\bar{y}_{.j} - \bar{y}_{..})^2}. \tag{7}$$

This statistic is F distributed with $a-1$ and $(a-1)(b-2)$ degrees of freedom under the additivity hypothesis.

Johnson & Graybill's Test (1972) These authors chose a different approach and derived a test for $(ab)_{ij} = k \cdot c_i \cdot d_j$ with c_i and d_j being a certain row or column constant and k an overall constant. They suggested the test statistic

$$S_J = \frac{\lambda_1(RR^T)}{\mathrm{tr}(RR^T)} = \omega_1. \tag{8}$$

For a type-I-error probability of α, the null hypothesis is rejected in favor of H_1 if $S_J > S_{crit}$, where S_{crit} is such that $P_{H_0}(S_J > S_{crit}) = \alpha$.

The non-standard distribution of this test statistic is derived in Johnson and Graybill (1972) and critical values are given.

Tusell's Test (Tusell (1990, 1992)) Tusell chose a similar approach to Johnson and Graybill (1972) and derived a test for $(ab)_{ij} = 0$. Without loss of generality, $a \leq b$ is assumed. The suggested test statistic is

$$S_U = (a-1)^{[(a-1)(b-1)]/2} \left(\prod_{l=1}^{a-1} \omega_l\right)^{(b-1)/2}. \tag{9}$$

Additivity is rejected if $S_U < S_{crit}$, where S_{crit} is such that $P_{H_0}(S_U < S_{crit}) = \alpha$.
Critical values for this test statistic are given e. g. in Kres (1972). Note that these tables are to be used with $(a-1) = p$ and $b = N$.

Boik's Test (1993) Boik derived a test for $(ab)_{ij} = 0$ that was designed to maximize power in the sense explained by Cox and Hinkley (1979) (therefore coined "locally best invariant test" or *LBI*). Without loss of generality, again $a \leq b$ is assumed. If the test statistic

$$S_B = (a-1)^{-1} \left(\sum_{l=1}^{a-1} \omega_l^2 \right)^{-1} . \tag{10}$$

is small, i.e. if $S_B < S_{crit}$ where $S_{crit.}$ is such that $P_{H_0}(S_B < S_{crit}) = \alpha$, H_0 is rejected.
How to compute critical values for this test statistic is explained in Boik (1993).

2.2. Modification of Tukey's test

It is known in literature (e.g. Johnson and Graybill 1972) that Tukey's test has good power for the interaction of type $(ab)_{ij} = k \cdot a_i \cdot b_j$ in model (1). We propose a modification of Tukey's test to overcome its problem with low power in more general cases.
In Tukey's test, the model

$$\boldsymbol{y}_{ij} = \mu + a_i + b_j + k \cdot a_i \cdot b_j + \boldsymbol{e}_{ij} \quad (i = 1, \ldots, a; j = 1, \ldots, b) \tag{11}$$

is tested against the submodel without interaction. The estimators of row and column effects are calculated in the same way in both models although the dependency of $\boldsymbol{y}_{ij}$ on these parameters is not linear for the full model.
The main idea behind the presented modification is that the full model (11) is fitted by non-linear regression and tested against the submodel by the likelihood ratio test. The estimators of row and column effects therefore differ for each model. The estimators of parameters in the non-linear model (11) are computed by the standard iterative procedure (e.g. Bates and Watts 1988); in one step of the iteration two parameters of the model are assumed to be fixed (equal to the last estimators) and the third parameter is estimated by the linear regression (the dependency is linear for this case).
Under the additivity hypothesis, the maximum likelihood estimators can be calculated simply as $\hat{\mu} = \bar{y}_{..}$, $\hat{a}_i^{(0)} = \bar{y}_{i.} - \bar{y}_{..}$ and $\hat{b}_j^{(0)} = \bar{y}_{.j} - \bar{y}_{..}$. The residual sum of squares equals

$$RSS^{(0)} = \sum_i \sum_j \left(y_{ij} - \hat{\mu} - \hat{a}_i^{(0)} - \hat{b}_j^{(0)} \right)^2 = \sum_i \sum_j \left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..} \right)^2 .$$

In the full model (11) let us first take

$$\hat{k}^{(0)} = \frac{\sum_i \sum_j \left(y_{ij} - \hat{a}_i^{(0)} - \hat{b}_j^{(0)} - \hat{\mu}^{(0)} \right) \cdot \hat{a}_i^{(0)} \cdot \hat{b}_j^{(0)}}{\sum_i \sum_j \left(\hat{a}_i^{(0)} \right)^2 \cdot \left(\hat{b}_j^{(0)} \right)^2},$$

which gives the same estimator (only written in different notation) of interaction as in Tukey's test, and - to get an estimator in the non-linear model - then continue with an iterative procedure that updates the estimators based on versions of the previous step:

$$\begin{aligned}
\hat{a}_i^{(n)} &= \frac{\sum_j \left(y_{ij} - \hat{\mu} - \hat{b}_j^{(n-1)}\right) \cdot \left(1 + \hat{k}^{(n-1)} \cdot \hat{b}_j^{(n-1)}\right)}{\sum_j \left(1 + \hat{k}^{(n-1)} \cdot \hat{b}_j^{(n-1)}\right)^2} \\
\hat{b}_j^{(n)} &= \frac{\sum_i \left(y_{ij} - \hat{\mu} - \hat{a}_i^{(n-1)}\right) \cdot \left(1 + \hat{k}^{(n-1)} \cdot \hat{a}_i^{(n-1)}\right)}{\sum_i \left(1 + \hat{k}^{(n-1)} \cdot \hat{a}_i^{(n-1)}\right)^2} \\
\hat{k}^{(n)} &= \frac{\sum_i \sum_j \left(y_{ij} - \hat{a}_i^{(n-1)} - \hat{b}_j^{(n-1)} - \hat{\mu}\right) \cdot \hat{a}_i^{(n-1)} \cdot \hat{b}_j^{(n-1)}}{\sum_i \sum_j \left(\hat{a}_i^{(n-1)}\right)^2 \cdot \left(\hat{b}_j^{(n-1)}\right)^2}.
\end{aligned}$$

The iteration should be stopped if the difference $|RSS^{(n)} - RSS^{(n-1)}|$ is low; denote $RSS = RSS^{(n)}$. The iteration converges very quickly in the most cases, usually just one step is enough.

The likelihood ratio statistic, i.e. a difference of twice log-likelihoods, equals $(RSS^{(0)} - RSS)/\sigma^2$ and is asymptotically χ^2 distributed with 1 degree of freedom.

The consistent estimate of the residual variance σ^2 is $s^2 = \frac{RSS}{ab-a-b}$ and $\frac{RSS}{\sigma^2}$ is approximately χ^2 distributed with $ab - a - b$ degrees of freedom. Thus, using a linear approximation of the nonlinear model (11),

$$S_{MT} = \frac{RSS^{(0)} - RSS}{\frac{RSS}{ab-a-b}} \tag{12}$$

is F distributed with 1 and $ab-a-b$ degrees of freedom. The Modified Tukey test rejects the additivity hypothesis if and only if S_{MT} is greater than $F_{1,ab-a-b}(1-\alpha)$, where $F_{1,ab-a-b}(1-\alpha)$ stands for the $1-\alpha$ quantile of the F distribution with 1 and $ab-a-b$ degrees of freedom.

It was shown by simulations that for a small number of rows or columns the type-I-risk is around 6% instead of 5%. The reason is that the likelihood ratio test statistic converges to a χ^2 distribution rather slowly (see Bartlett 1937) and a correction for small sample sizes is needed.

One possibility to overcome this obstacle is to estimate the residual variance $s^2 = \frac{RSS}{ab-a-b}$ and then generate samples of a distribution

$$y_{ij}^{(sample)}(t) = \hat{\mu} + \hat{a}_i^{(0)} + \hat{b}_j^{(0)} + \boldsymbol{e}_{ij}^{(NEW)}(t), \quad k = 1, \ldots, N^{(sample)}$$

where $(\boldsymbol{e}_{ij}^{(NEW)})(t)$ are i.i.d. generated from a $N(0, s^2)$.

The proposed statistic of interest is $\mathrm{abs}(k^{(n)})$. The additivity hypothesis is rejected if more than $(1-\alpha)\cdot 100\%$ of sampled statistics lie below the observed value of the statistic based on real data.

2.3. Evaluation of interaction in block designs

In a common fixed effects model for a block design, i.e. model (1), it is not possible to separate interaction and error term. Hence the mean squares value for interaction serves as a basis for

the test of main effects. This only makes sense in those situations where no interaction at all exists.

It is possible to gain an alternative test procedure by applying an additional restriction to the sum of interaction effects within a column. In a common block design, restrictions to factor and interaction effects refer to the sums $\sum_{i=1}^{a} a_i = 0$, $\sum_{j=1}^{b} b_j = 0$, and $\sum_i (ab)_{ij} = 0$. If we expand a block design to a Latin square design by adding an appropriate number of blocks, it seems to be reasonable to use $\sum_k (ab)_{ik} = 0$ as an additional restriction for interaction effects within each column k (as there is no real difference between blocks and columns). Applying such a restriction to a column has an interesting effect on the mean of that column. All block and factor effects are included within the column the same number of times, therefore all of these effects sum up to zero. Based on this new restriction (all interaction effects within a column sum up to 0), there is only one effect left within the mean of a column and this is the error term (besides μ). Therefore it is possible to calculate a sum of squares value for the error term by computing $SS_E = \sum_{k=1}^{a} (\bar{y}_{.k} - \bar{y}_{..})^2$. The sum of squares value for interaction can be gained as the difference $SS_{AB} = SS_T - SS_A - SS_B - SS_E$ with SS_{AB} denoting sum of squares for interaction, SS_T denoting sum of squares total, SS_A refers to sum of squares for factor A and SS_B is the sum of squares for block/factor B.

The method as such is not restricted to Latin squares but can be used in Latin rectangles as well. The essential part is that a number of columns can be found that includes all levels of blocks and factor effects. This can easily be done in Latin rectangles by combining two or more columns to a column block. Again the mean values of these column blocks include the error term only (besides μ).

Applying this method to any arbitrary block design is possible too, although some correction for the sum of column blocks has to be made.

a_1	a_2		a_3	a_4	a_5	b_1
a_5	a_3	a_4		a_1	a_2	b_2
cb_1			cb_2			

In column block 1 (cb_1) all levels of the factor A are included, but it contains two times block effect 1 (b_1) and three times block effect 2 (b_2). If we subtract the estimated effect of the second block ($\hat{b}_2$) from the sum of cb_1 and apply a corresponding correction for cb_2, an unbiased estimator for SS_E can be found. This means that there is no restriction to the use of the method in any block design.

By means of a simulation study the power of the method was evaluated. The standard deviations for interaction and factor effects varied between 0 and 4 and for the error term between 1 and 4. For each combination of factor, interaction and error effects 100 000 datasets were generated and analysed. Based on these simulations the properties of the new method are summarized in the following:

- The method keeps type-I-risk exactly.
- The power of the method is good.
- Violations of prerequisites ($\sum_k (ab)_{ik} = 0$) lead to a reduction of power.
- With a high number of plots per block ($a \geq 5$) this loss of power is small.

No restrictions to the nature of interaction as with the tests by Tukey (1949), Mandel (1961), etc. have to be made, the restriction is only related to the sum of interaction effects within a column block. The method is easy to use and common statistical packages can be applied with some additional calculations.

The following example demonstrates the method.

block	column 1	column 2	column 3	$\bar{x}_B$
1	[1] 8.452	[3] 7.880	[2] 11.374	9.235
2	[2] 8.312	[1] 11.302	[3] 11.768	10.461
3	[3] 12.600	[2] 9.192	[1] 5.122	8.971
$\bar{x}_C$	9.788	9.458	9.421	9.556

	factor 1	factor 2	factor 3
$\bar{x}_A$	8.292	9.626	10.749

Table 1: data set for a 3×3 block design with corresponding means for factor- ($\bar{x}_A$), block- ($\bar{x}_B$), and column- effects ($\bar{x}_C$) ([i] denotes i^{th} level of factor A).

The sum of squares values (SS) for block, factor and column effects are calculated as usual. The SS for columns serves as SS for the error term. The remaining SS value is a measurement for interaction (error term in Latin Squares). The following table presents the ANOVA for this kind of Analysis.

Effect	*df*	*SS*	*MS*	*F*	*Prob*
factor	2	9.0799	4.5400	37.11	0.02624
block	2	3.7893	1.8946	15.49	0.06066
Interaction	2	32.7668	16.3834	133.91	0.00741
Error	2	0.2447	0.1223		
Total	8	45.8807			

Table 2: ANOVA for the data set of table 1 corresponding to the new methode

Analysing data from table 1 as an ordinary Latin Square leads to F-values all smaller then 1 and to no significant result.

3. Mixed Effect Models

3.1. Type-I-risk and power of additivity tests

The tests described in Section 2.1 were designed for the ANOVA model with both factors fixed. In this section their type-I-risk will be verified for the mixed ANOVA model by simulation. Only the most common nominal significance level 5% was assumed.

The number of levels of the fixed factor were $a = 3, 4, \ldots, 10$, the number of levels of the

random factor b was chosen to lie between 4 and 50 (by 2 between 4 and 20, by 5 between 20 and 50), the variance of the random factor was $\sigma_b^2 = 2, 5, 10$ and the variance of the random error $\sigma^2 = 1$. Thus we have 360 combinations in total. In case of other values of σ^2, the model can be scaled.

In each step of the simulation a data set was generated based on the model without interaction. Then the test of no interaction was applied. The percentage of significant test statistics after all steps was used as the actual level (α_{act}) of the test.

For each combination of parameter values, 10 000 simulations were repeated 10 times and the standard error of the estimation of the mean actual level was computed based on these 10 repetitions. Then a one-sided one sample t-test of $\alpha_{act} \leq .05$ was performed (on the 5% level, without correction for multiple testing). The results of these tests are summarized in Table 3.

Test	$\alpha_{act} \leq .05$		$\alpha_{act} > .05$	
Tukey test	349	(96.94)	11	(3.06)
Mandel test	348	(96.67)	12	(3.33)
Johnson Graybill test	339	(94.17)	21	(5.83)
LBI test	336	(93.33)	24	(6.67)
Tusell test	337	(93.61)	23	(6.39)

Table 3: Absolute frequencies and percentage of non-significant (second and third column) and significant (fourth and fifth column) t-tests of $\alpha_{act} \leq .05$ for all 360 combinations of parameter values.

For the tests by Tukey (1949) and Mandel (1961) results that in the vast majority (>95%) of cases the actual level is not significantly above the .05 level. In less than 4% the type-I-risk is higher than the nominal .05. For the other tests, the estimated level is greater than .05 in slightly more cases. However, this may also be false positives caused by multiple testing.

In ANOVA models with both factors fixed, there is the important assumption of $\sum_{i=1}^{a} a_i = \sum_{j=1}^{b} b_j = 0$. In the case of a mixed model that is not valid. It is assumed that the expected value of the random term $E(\boldsymbol{b}_j)$ equals zero, but in one particular case the sum was not zero (almost surely). It might have caused the inaccuracy of the tests. However, for high numbers of levels of the random factor, the sum converges to zero (law of large numbers) and this problem disappears.

To evaluate the power of the additivity tests, a simulation study was performed. Assume model (2) and two types of interaction:

(i) $(\boldsymbol{ab})_{ij} = k \cdot a_i \cdot \boldsymbol{b}_j$ where k is a real constant.

(ii) $(\boldsymbol{ab})_{ij} = k \cdot a_i \cdot \boldsymbol{c}_j$ where $\boldsymbol{c}_j$ are independent normally distributed random variables with zero mean and variance σ_b^2, independent of $\boldsymbol{b}_j$ and $\boldsymbol{e}_{ij}$, and a real constant k.

The other parameters were set to $\mu = 0$, $\sigma_B^2 = 2$, $\sigma^2 = 1$, $a = 10$, $(a_1, \ldots, a_{10}) = (-2.03, -1.92, -1.27, -0.70, 0.46, 0.61, 0.84, 0.94, 1.07, 2.00)$. Two possibilities were considered for b, either $b = 10$ or $b = 50$. For the interaction parameter k, 10 different values between 0 and 12 were considered.

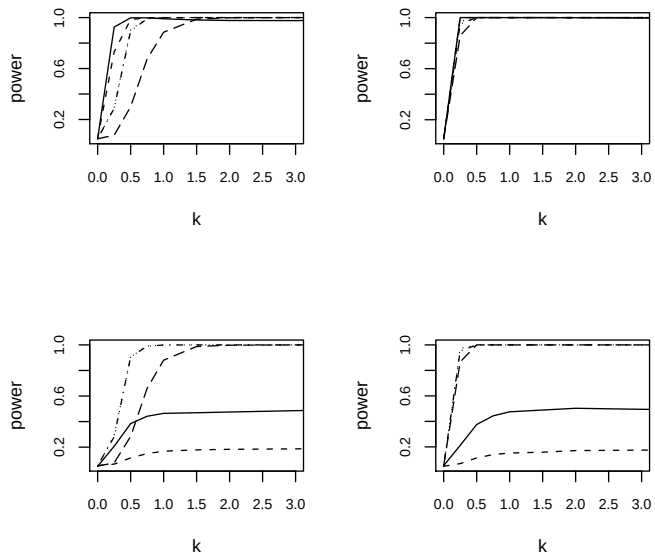

Figure 1: Power dependence on k, b ($b = 10$ left, $b = 50$ right) and interaction type ((i) upper panel, (ii) lower panel). Tukey's test: solid line, Mandel's test: dashed line, Johnson & Graybill's test: dotted line, LBI test: dot-dash line (these two lines overlapped), Tusell's test: long dash line.

For each combination of parameter values a dataset was generated based on model (2). The tests of additivity were carried out with $\alpha = 5\%$ and the decision was noted. This step was repeated 10 000 times. The percentage of positive results was used as an estimate of the power of the test. Tukey's and Mandel's tests outperformed the omnibus tests for interaction (i) but they completely failed to detect the interaction (ii), even for large values of k. Therefore, it is desirable to have a test which is able to detect a spectrum of practically relevant alternatives, while still having the power comparable to the tests by Tukey (1949) and Mandel (1961) for the most common interaction scheme (i). The Modified Tukey test from Section 2.2 might serve this purpose, which should be investigated in further research.

3.2. Robustness with binary data

Another interesting application of additivity tests arises in the calibration of instruments for psychological assessment. The appropriate model is again a mixed model as in model (2), but this time an additional complication appears, namely that only binary responses ("item solved" or "item not solved") are observed. If no interaction is present and a logit link is used for $g(.)$, model (2) reduces to - if formulated differently as $E(\boldsymbol{y}_{ij}) = g^{-1}(\boldsymbol{b}_j - \mu_i)$ - the *Rasch* model (Rasch 1960) and plays an important role in measurement theory of psychological traits. Here, a_i denotes the "easiness" of item i (item parameter) and $\boldsymbol{b}_j$ denotes the ability of person j (person parameter).

The *Rasch* model has a number of desirable properties (see, e.g., Fischer and Molenaar 1995) and a necessary condition for the *Rasch* model to hold is the absence of any kind of interaction between the person parameter and the item parameter. It therefore assumes strict additivity. This assumption can be tested with additivity tests.

One possible way of being able to test the additivity assumption would be to use one of the parametric tests presented above. Clearly, these tests have not been developed for the binary case, so we investigated their behaviour in case of binary data to see if they are robust against the violation of the assumption of a normal distribution.

To assess the type-I-risk, the number of levels for the fixed factor (items) were $a = 10, 15, \ldots, 40, 50, \ldots, 100, 150$ and 200 and the number of levels of the random factor (the testees) were $b = 50, 100, \ldots, 400, 500, \ldots$ and 1000. The levels a_i of the fixed factor were set as equally spaced out on the interval $[-5, 5]$, which basically corresponds to the whole spectrum of item difficulties that arise in practice. The levels of the fixed factor $\boldsymbol{b}_j$ were drawn randomly from a $N(0, 1)$, again corresponding to the values of person parameters that are likely to occur in practice. In each step of the simulation a data set was generated by calculating the solving probability p_{ij} for each person j on item i by $g^{-1}(a_i + \boldsymbol{b}_j) = exp(a_i + \boldsymbol{b}_j)/1 + exp(a_i + \boldsymbol{b}_j)$. Then a Bernoulli trial has been carried out with the parameter p_{ij}, which led to a matrix that conforms to the *Rasch* model. 10 000 data matrices were generated for each person $\times$ item combination.

Unfortunately, none of the tests showed robustness against the violation of the normality assumption and could be used to test if interactions are present.

4. Discussion

Of the six tests (including the Modified Tukey test) for model (1) which were developed in the literature or by us, five have been tested by simulation for quantitative and binary data. Unfortunately, for the binary case, no test can be recommended. For quantitative data, the actual type-I-risk was pretty near to the nominal one for all tests, with the LBI and Tusell's being the worst.

Concerning power, the tests behaved quite differently for different types of interaction. For interaction type (i), the tests by Tukey and Mandel were the best but they failed with interaction of type (ii). The others performed well in case of interaction type (ii). For both cases, the Modified Tukey test provided satisfactory results.

To summarize: For model (1), the experimental design approach using no structure for the interaction term can be recommended but its behaviour for the mixed model is still an open question. For the mixed model we recommend to use the tests by Mandel or Tukey or its modified version as long as we are interested in the error of the first kind only. If we are interested in high power and no specific interaction can be assumed, the omnibus tests or the Modified Tukey test would be methods of choice.

5. Computational Details

All calculations have been carried out with R (R Core Team 2008). Implementations of the mentioned additivity tests can be found in the R package **additivityTests** (Simeckova, Simecek, and Rusch 2007).

References

Alin A, Kurt S (2006). "Testing non-additivity (interaction) in two-way ANOVA tables with no replication." *Statistical methods in medical research*, **15**(1), 63–85.

Bartlett MS (1937). "Properties of sufficiency and statistical tests." *Royal Society of London Proceedings Series A*, **160**, 268–282.

Bates DM, Watts DG (1988). *Nonlinear regression analysis and its applications*. John Wiley & Sons, New York.

Boik RJ (1993). "Testing additivity in two-way classifications with no replications: the locally best invariant test." *Journal of Applied Statistics*, **20**, 41–55.

Cox DR, Hinkley DV (1979). *Theoretical Statistics*. Chapman and Hall, London.

Fischer GH, Molenaar IW (eds.) (1995). *Rasch Models: Foundations, Recent Developments and Applications*. Springer, New York.

Johnson DE, Graybill FA (1972). "An analysis of a two-way model with interaction and no replication." *Journal of the American Statistical Association*, **67**, 862–869.

Karabatos G (2005). "Additivity Test." In BS Everitt, DC Howell (eds.), *Encyclopedia of Statistics in Behavioral Science*, pp. 25–29. Wiley.

Kres H (1972). *Statistical Tables for Multivariate Analysis*. Springer, New York.

Mandel J (1961). "Non-additivity in two-way analysis of variance." *Journal of the American Statistical Association*, **56**, 878–888.

R Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rasch G (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.

Simeckova M, Simecek P, Rusch T (2007). *additivityTests: Additivity tests in the two way ANOVA with single sub-class numbers*. R package version 0.3, URL https://github.com/rakosnicek/additivityTests.

Tukey JW (1949). "One degree of freedom for non-additivity." *Biometrics*, **5**, 232–242.

Tusell F (1990). "Testing for interaction in two-way ANOVA tables with no replication." *Computational Statistics and Data Analysis*, **10**, 29–45.

Tusell F (1992). "Corrigendum: Testing for interaction in two-way ANOVA tables with no replication." *Computational Statistics and Data Analysis*, **13**, 121.

Affiliation:

Dieter Rasch
Institute of Applied Statistics
University of Natural Resources and Applied Life Sciences
Vienna, Austria
Tel.: +43-1-47654-5061
Fax: +43-1-47654-5069
E-mail: dieter.rasch@boku.ac.at

Thomas Rusch
Department of Statistics and Mathematics
WU Vienna
Vienna, Austria
Tel.: +43-1-31336-4338
Fax: +43-1-31336-774
E-mail: thomas.rusch@wu-wien.ac.at

Marie Šimečková
Institute of Animal Science
Biometric Unit
Prague, Czech Republic
Tel:: +420-267-009-527
E-mail: simeckova.marie@vuzv.cz

Klaus D. Kubinger
Division for Psychological Assessment and Applied Psychometrics
Faculty of Psychology
University of Vienna, Austria
Tel.: +43-1-4277-47850
E-mail: klaus.kubinger@univie.ac.at

Karl Moder
Institute of Applied Statistics
University of Natural Resources and Applied Life Sciences
Vienna, Austria
Tel.: +43-1-47654-5062
E-mail:karl.moder@boku.ac.at

Petr Šimeček
Institute of Animal Science
Biometric Unit
Prague, Czech Republic
E-mail: simecek.petr@vuzv.cz