Psychological Test and Assessment Modeling, Volume 52, 2010 (4), 343-353

Alternatives to *F*-Test in One Way ANOVA in case of heterogeneity of variances (a simulation study)

Karl Moder[1]

Abstract

Several articles deal with the effects of inhomogeneous variances in one way analysis of variance (ANOVA). A very early investigation of this topic was done by Box (1954). He supposed, that in balanced designs with moderate heterogeneity of variances deviations of the empirical type I error rate (on experiments based realized α) to the nominal one (predefined α for H_0) are small. Similar conclusions are drawn by Wellek (2003). For not so moderate heterogeneity (e.g. $\sigma_1 : \sigma_2 : \ldots = 3 : 1 : \ldots$) Moder (2007) showed, that empirical type I error rate is far beyond the nominal one, even with balanced designs. In unbalanced designs the difficulties get bigger. Several attempts were made to get over this problem. One proposal is to use a more stringent α level (e.g. 2.5% instead of 5%) (Keppel & Wickens, 2004). Another recommended remedy is to transform the original scores by square root, log, and other variance reducing functions (Keppel & Wickens, 2004, Heiberger & Holland, 2004). Some authors suggest the use of rank based alternatives to *F*-test in analysis of variance (Vargha & Delaney, 1998). Only a few articles deal with two or multi-factorial designs. There is some evidence, that in a two or multi-factorial design type I error rate is approximately met if the number of factor levels tends to infinity for a certain factor while the number of levels is fixed for the other factors (Akritas & S., 2000, Bathke, 2004).

The goal of this article is to find an appropriate location test in an oneway analysis of variance situation with inhomogeneous variances for balanced and unbalanced designs based on a simulation study.

Key words: heteroscedasticity, analysis of variance, alternatives

[1] *Correspondence concerning this article should be addressed to:* Karl Moder, PhD, University of Natural Resources and Applied Life Sciences, Institute of Applied Statistics and Computing, Peter-Jordan-Str. 82, A-1190 Vienna, Austria; email: karl.moder@boku.ac.at

1. Introduction

Analysis of variance is based on three assumptions:

- normal distributed populations,
- homogeneity of variances,
- independent samples.

In case of independent and identically normal distributed data, analyzing the effects of several levels of a factor is done by two-sample *t*-test (in the case of 2 levels) or by one-way analysis of variance (in case of 2 or more levels). These tests are uniformly most powerful tests as long as all prerequisites are met and both kinds of analysis keep type one error rate α.

Dependency is due to the experimenter and is taken into account by appropriate methods (paired *t*-test, block analysis,…).

Heteroscedasticity (random variables have different variances) is regarded to be a serious problem especially if sample sizes are unequal at different levels of the factor. In this situation both *t*-test and *F*-test in analysis of variances exceed the nominal α value depending on a number of influences like differences in variances, balanced or unbalanced design, number of observations and number of factor levels. Several solutions are recommended for this situation.

- Select a new critical value of *F* at a more stringent significance level, namely, α=0.025 to keep Type I error rate below the 5% level (Keppel, Saufley, Tokunaga, & Zedeck, 1992). This approach is not appropriate as the α-level is not kept in many situations in dependence of the underlying populations and on the other hand if heteroscedasticity is not too extreme the power of the test is low.
- Transform original scores using square roots, log-, arcsine-transformations to reduce heterogeneity and to normalize distributions (Keppel & Wickens, 2004). This does not really lead to homogeneous variances, but tests on homogeneity of variances are not powerful enough to find significant results (Moder, 2007).
- Use robust methods for analyses. For the two samples situation the problem can be handled by the use of the two-sample Welch-test (Rasch, Kubinger, & Moder, 2009) which keeps type I error rate despite heteroscedastic variances.
- Some authors recommend robust non-parametric tests (Vargha & Delaney, 1998) in case of heterogeneous variances. In several articles in the Internet (E&B, 2010; MESOSworld, 2010; Statlab, 2005) Kruskal-Wallis test is recommended in situations where homoscedasticity is violated. As this location test depends on rather equality in particular of second but also of third and fourth moments this proposal is hardly appropriate for the investigated situation.

2. Methods

In a simulation study those methods are evaluated which are suggested or seem to be appropriate in case of inhomogeneous variances (in an analysis of variance model with more than 2 factor levels).

Based on these recommendations seven different kinds of analysis methods were examined:

1. *F*-test in Analysis of Variance as a kind of standard,
2. Welch-Test for more than 2 samples,
3. weighted ANOVA as available in SAS procedure Mixed based on the Satterthwaite approximation in a repeated measurement analysis,
4. Kruskal-Wallis test,
5. Permutation test using *F*- statistic as implemented in *R*-package "coin",
6. Permutation test based on Kruskal-Wallis statistic,
7. and a special kind of Hotelling's T^2 test (Hotteling, 1931; Moder, 2007).

Methods 1 - 4 were performed using SAS (SAS Institute Inc., 2008). For method 5 - 7 R (R Development Core Team, 2009) was used. All simulation runs were carried out on a PC (Intel Core i7 CPU 965 @ 3.20 GHz, Physical Memory 8192 MB, Microsoft Windows Vista Business).

F-test in analysis of variance was investigated as a comparison model. Welch-Test is a possible option in almost each major statistical package for situations when variances are inhomogeneous. Method 3 corresponds to a "Weighted Least Squares" analysis whereby the weights are determined as the inverse variances of the residuals from each group. The Satterthwaite option leads to a test which corresponds to the Welch approximation (Singer, 1998). Permutation tests are thought to be rather robust in cases of deviations from prerequisites in the analysis of variance model. A lot of test statistics are possible candidates for permutation test. As *F*-value (method 5) of the R-package "coin" is an often used statistic with permutation tests, it is therefore part in the simulation study. For some alternative to this a permutation test based on Kruskal-Wallis statistic (method 6) was evaluated. As concerns the situation of balanced designs with at least as much observations within groups as factor levels Hotelling's T^2 test was examined.

Simulation studies were performed using SAS (SAS Institute Inc., 2008) and R (R Development Core Team, 2009). Each simulation run was based on 100000 normal distributed data sets. Balanced (sample size equal at all factor levels) as well as unbalanced designs (unequal sample sizes) were studied. In case of balanced designs all 7 methods were examined otherwise only the first 6 methods were used, as Hotelling's T^2 is not applicable on unbalanced data. The number of factor levels varied between 3 and 20. The number of observations was chosen between 3 and 30. Standard deviations in populations differed between 1, 2 and 3.

3. Simulation results

In the following 2 figures some results for a specific situation of five levels of the factor and 5 observations within each sample are shown. The ratio of "true" standard deviations ($\sigma_1 : \sigma_2 : \ldots : \sigma_5$) was given by 3:1…:1. In Figure 1 the empirical type I error rate was compared to a nominal one of 0.05 for all 7 methods.

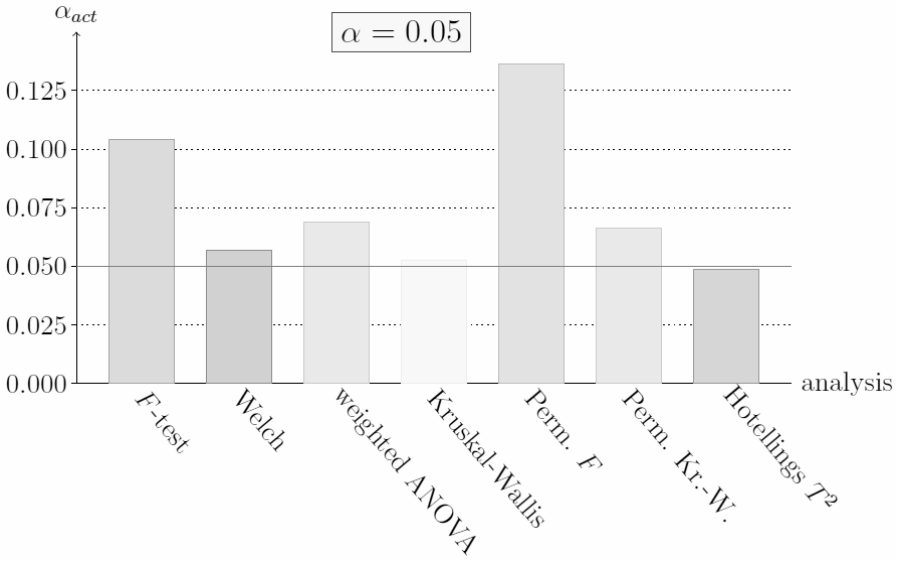

Figure 1:
Empirical significance level for 5 levels of a factor and 5 observations each with an ratio of $\sigma_1 : \sigma_2 : \ldots : \sigma_5 = 3 : 1 : \ldots : 1$ (α=0.05).

For the depicted situation there are 2 methods which keep the nominal α value, namely Kruskal-Wallis-test and Hotellings T^2. All other kinds of analyses exceed the nominal type I error rate by up to 8.6%. Especially *F*-test in analysis of variance and permutation tests based on *F*-statistic perform very badly. Welch-test is within a 20% tolerance (Rasch, Teuscher, & Guiard, 2007) and may be appropriate for this specific situation.

In Figure 2 the underlying situation is identical to that of Figure 1. In contrast to Figure 1 the nominal α value was defined as 0.01. Results are similar to that of Figure 1. Kruskal Wallis test tends to be too conservative. Hotelling's T^2 is the only method which keeps the nominal α level exactly.

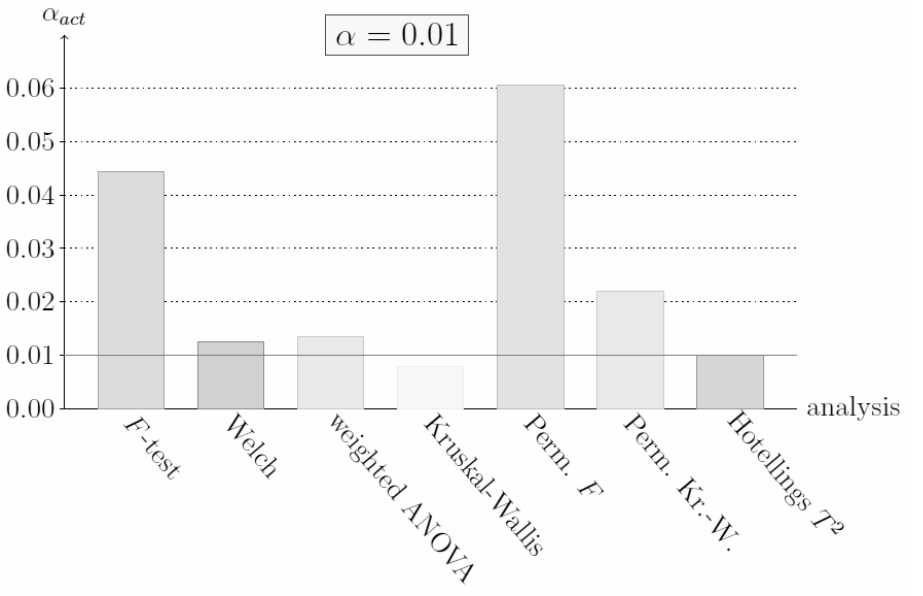

Figure 2:
Empirical significance level for 5 levels of a factor and 5 observations each with an ratio of $\sigma_1 : \sigma_2 : \ldots : \sigma_5 = 3 : 1 : \ldots : 1$ (α=0.01)

A lot of different situations in regard to factor levels, variance ratios and sample sizes were examined. Table 1 shows results for a range of factor levels between 3 and 20. The number of observations changed from 3 to 20 with equal sample sizes. The ratio of standard deviations was fixed to 2:1:…:1. Hotelling's T^2 can only be used in situations, where the number of observations per sample is at least the number of factor levels. This is caused by the fact, that no program was available for calculating the cumulative density function of Hotelling's T^2 directly. Therefore the approximation to F distribution was used. For the specific test this approximation restricts the number of factor levels to the number of observations at the most. With the availability of tables for Hotelling's T^2 this limitation will fall.

In Table 1 simulation results (at a nominal α value of 0.01) for all 7 methods are presented. The number of factor levels ranged from 3 to 20. The design was a balanced one, with a number of observations which ranged from 3 to 20, too.

Table 1 shows, that even in situations with a narrow ratio of standard deviations *F*-test in analysis of variance as well as permutation tests based on *F*-statistic don't keep nominal type I error rate. Welch test performs very badly in situations when the number of factor levels is high, especially with small sample sizes. Weighted Analysis of Variance (SAS Mixed) is also not appropriate in most situations. Kruskal Wallis test is very conservative

Table 1:
Simulation results (observed type I error rates) for 7 types of analyses with a ratio of standard deviations $\sigma_1 : \sigma_2 : \ldots : \sigma_{nf} = 2:1:\ldots:1$ and a nominal α of 0.05.

						Permutation-Test		
n_f	n_{obs}	*F*-test	Welch-Test	weighted ANOVA	Kruskal-Wallis-test	*F*-statistic	Kruskal-Wallis	Hotellings T^2
3	3	0.0682	0.0439	0.0492	0.0160	0.0551	0.0527	0.0497
3	5	0.0670	0.0482	0.0580	0.0522	0.0710	0.0574	0.0505
3	10	0.0631	0.0502	0.0563	0.0539	0.0657	0.0571	0.0490
3	20	0.0618	0.0495	0.0527	0.0553	0.0646	0.0587	0.0497
5	3	0.0723	0.0651	0.0646	0.0245	0.0774	0.0598	-
5	5	0.0722	0.0555	0.0712	0.0444	0.0900	0.0602	0.0500
5	10	0.0706	0.0518	0.0618	0.0499	0.0879	0.0556	0.0506
5	20	0.0688	0.0500	0.0553	0.0495	0.0883	0.0574	0.0496
10	3	0.0741	0.1392	0.0923	0.0218	0.0927	0.0553	-
10	5	0.0492	0.0773	0.0740	0.0325	0.0999	0.0550	-
10	10	0.0733	0.0580	0.0561	0.0463	0.1049	0.0541	0.0474
10	20	0.0717	0.0536	0.0514	0.0508	0.1077	0.0553	0.0511
20	3	0.0702	0.2895	0.1107	0.0205	0.0979	0.0512	-
20	5	0.0703	0.1282	0.0762	0.0346	0.1045	0.0525	-
20	10	0.0709	0.0686	0.0428	0.0451	0.1139	0.0524	-
20	20	0.0694	0.0530	0.0390	0.0472	0.1159	0.0526	0.0451

if the sample size is smaller or equal to the number of factor levels (this situation occurs seldom in oneway analysis of variance designs but is common in block designs). A permutation test based on Kruskal Wallis test statistic exceeds nominal α but perhaps in a tolerable way. Hotelling's T^2 performs best although it seems to be conservative in some situations. Neither SAS nor R compute distribution function of Hotelling's T^2 directly but use an approximation to the *F*-distribution (Moder, 2007). Hotelling's T^2 is a uniform most powerful test for the depicted situation (Simaika, 1941; Anderson, 1958). So probably this approximation to the *F*-distribution is reason for this too low observed type I error rate with small sample sizes.

Probability values for Kruskal-Wallis test are calculated approximately using SAS procedure NPAR1WAY. Exact probability values are available with this procedure but their calculation is rather time consuming even for 5 levels of a factor with 3 replications each. For a smaller number of factor levels the exact calculation of probability leads to higher type one error rates than expected (For the situation of 3 levels of the factor with 3 replications each leads to an observed type I error rate of 0.0878 with an nominal α of 0.05 –

whereas the approximate type one error rate is 0.0221 which is too conservative for a ratio of $\sigma_1 : \ldots : \sigma_{nf} = 3 : \ldots : 1$).

Using permutation tests with small sample sizes and a high number of permutations (e.g. 10000) is problematic. Because of the rather low number of possible rearrangements of data the probability to create all possible datasets for several times is high. So it would be better to use exact probability values, but this is not available in the coin package for more than 2 factor levels.

In Table 2 an equivalent situation as in Table 1 is presented. In difference to Table 1 the ratio of standard deviations was fixed to 3:1:…:1.

Based on the results of Table 2 none of the presented methods except Hotelling's T^2 test can be recommended.

As already Box (1954) pointed out, the influence of heteroscedasticity is much higher in unbalanced designs than in balanced ones. In the following some results for unbalanced designs are presented. Table 3 illustrates the situation if all standard deviations are 1 except for the first one which is 3 and all sample sizes are 3 except for the second one which is 5.

Table 2:
Simulation results (observed α-values) for 7 types of analyses with a ratio of standard deviations $\sigma_1 : \sigma_2 : \ldots : \sigma_{nf} = 3 : 1 : \ldots : 1$ and a nominal α of 0.05.

						Permutation-Test		
n_f	n_{obs}	*F*-test	Welch-Test	weighted ANOVA	Kruskal-Wallis-test	*F*-statistic	Kruskal-Wallis	Hotellings T^2
3	3	0.0920	0.0510	0.0548	0.0221	0.0864	0.0686	0.0497
3	5	0.0836	0.0501	0.0561	0.0598	0.0962	0.0634	0.0500
3	10	0.0786	0.0510	0.0525	0.0629	0.0885	0.0682	0.0490
3	20	0.0728	0.0495	0.0521	0.0651	0.0795	0.0666	0.0497
5	3	0.1083	0.0691	0.0643	0.0333	0.1231	0.0741	-
5	5	0.1028	0.0573	0.0688	0.0526	0.1361	0.0658	0.0500
5	10	0.0969	0.0510	0.0612	0.0576	0.1362	0.0635	0.0506
5	20	0.0927	0.0518	0.0527	0.0607	0.1324	0.0642	0.0496
10	3	0.1175	0.1433	0.0869	0.0258	0.1543	0.0620	-
10	5	0.1126	0.0826	0.0793	0.0434	0.1684	0.0611	-
10	10	0.1124	0.0583	0.0578	0.0523	0.1775	0.0577	0.0474
10	20	0.1069	0.0520	0.0520	0.0554	0.1861	0.0603	0.0511
20	3	0.1157	0.2913	0.1057	0.0223	0.1731	0.0562	-
20	5	0.1138	0.1292	0.0747	0.0378	0.1861	0.0576	-
20	10	0.1099	0.0695	0.0442	0.0471	0.1997	0.0549	-
20	20	0.1075	0.0545	0.0408	0.0517	0.2137	0.0547	0.0431

Hotelling's T^2 cannot handle such situations as it is restricted to balanced designs and therefore it is not entered in Table 3, 4 and 5.

Table 3:
Simulation results (observed α-values) for 6 types of analyses with a ratio of standard deviations $\sigma_1 : \sigma_2 : \ldots : \sigma_{nf} = 3:1:\ldots:1$ and a nominal α of 0.05 and sample size is 5 for the second sample otherwise it is 3.

					Permutation-Test	
n_f	*F*-test	Welch-Test	weighted ANOVA	Kruskal-Wallis-test	*F*-statistic	Kruskal-Wallis
3	0.1307	0.0586	0.0563	0.0970	0.1376	0.1092
5	0.1309	0.0721	0.0603	0.0417	0.1485	0.0805
10	0.1270	0.1366	0.0789	0.0289	0.1647	0.0646
20	0.1174	0.2829	0.0974	0.0225	0.1745	0.0563

For a small number of factor levels no method keeps nominal α value. Welch-Test and weighted ANOVA perform best in this situation, but with a higher number of factor levels, the results get worse. Kruskal-Wallis test cannot be recommended, because it gets more conservative with increasing number of factor levels. Permutation variant of this test seems to become better with increasing factor levels, but with a higher number of observations (e.g. sample sizes: 5,25,10 observed α=0.118 at the 5%-level) the problem get worse. Overall no method is appropriate for the situation of inhomogeneous variances as soon as the biggest sample size does not correspond to the highest variance.

Situation for Table 4 is contrary to that of Table 3 in that sense, as the biggest sample size (5) corresponds with the highest standard deviations (3).

Table 4:
Simulation results (observed α-values) for 6 types of analyses with a ratio of standard deviations $\sigma_1 : \sigma_2 : \ldots : \sigma_{nf} = 3:1:\ldots:1$ and a nominal α of 0.05 and sample size is 5 for the first sample otherwise it is 3.

					Permutation-Test	
n_f	*F*-test	Welch-Test	weighted ANOVA	Kruskal-Wallis-test	*F*-statistic	Kruskal-Wallis
3	0.0379	0.0428	0.0463	0.0371	0.0429	0.0452
5	0.0457	0.0606	0.0639	0.0302	0.0678	0.0491
10	0.0552	0.1326	0.0973	0.0222	0.0951	0.0471
20	0.0598	0.2793	0.1174	0.0188	0.1211	0.0440

In situations, where the highest standard deviation corresponds to a high number of observations (Table 4) and the number of factor levels is low (3) all examined methods work well. As soon as n_f increases there is only one method which seems to be appropriate namely permutation test based on Kruskal-Wallis test statistic, but the test becomes very conservative if the number of observations increases (e.g. sample sizes: 25,10,5 observed α=0.012 at the 5%-level).

Table 5 reflects situations where highest sample sizes come with highest standard deviations, highest sample sizes come with lowest standard deviations and situations with no relation between height of standard deviation and sample size.

From Table 5 it can be seen, that type I error rate depends to a high degree on how sample sizes and standard deviations are connected. If a small number of observations comes with a high standard deviation, then type I error rate exceeds the nominal α level. If small sample sizes correlate with small standard deviations, then Welch test and weighted ANOVA exceed type I error rate, all others are too conservative. In situations where the height of standard deviations is not bound to high or low sample sizes all methods are more or less inappropriate.

Table 5:
Simulation results (observed α-values) for 5 levels of a factor with a ratio of sample sizes of 5:5:10:15:15 and different order of standard deviations.

ratio of standard deviations	*F*-test	Welch-Test	weighted ANOVA	Kruskal-Wallis-test	Permutation-Test *F*-statistic	Permutation-Test Kruskal-Wallis
2:2:1:1:1	0.1835	0.0647	0.0688	0.0907	0.1828	0.1039
1:1:1:2:2	0.0231	0.0564	0.0656	0.0251	0.0401	0.0314
1:1:2:1:1	0.0718	0.0601	0.0770	0.0503	0.0853	0.0568
2:1:1:1:2	0.0688	0.0627	0.0709	0.0510	0.0831	0.0598
3:3:1:1:1	0.2775	0.0683	0.0597	0.1221	0.2851	0.1334
1:1:1:3:3	0.0233	0.0504	0.0610	0.0256	0.0453	0.0305
1:1:3:1:1	0.0953	0.0594	0.0717	0.0569	0.1277	0.0634
3:1:1:1:3	0.0837	0.0617	0.0640	0.0613	0.1119	0.0718

3.1 Conclusions

In case of heteroscedasticity the following conclusions can be drawn from simulation study:

- *F*-test in analysis of variance is unsuitable for analysis.
- Hotelling's T^2 test keeps nominal type one error rate exactly in all situations of balanced designs. Some conservative results seem to be based on the approximation to the *F*-statistic.

- Kruskal-Wallis test does not exceed nominal α in some situations (if one uses an approximative test) but is very conservative.
- Permutation tests are no solution to the problem of heteroscedasticity. Especially *F*-statistic performs very badly.
- Welch test may be useful for a small number of factor levels (Rasch et al.., 2009), but is unsuitably for more than 2 or 3 levels.
- Weighted ANOVA of SAS's Procedure Mixed cannot be recommended.
- In case of unbalanced designs none of the examined methods performs well in all situations.

So as an overall conclusion there is only one method which is applicable in case of heteroscedasticity namely Hotelling's T^2. A test strategy based on pretests is of no use because of the low power of these tests (Rasch et al.., 2009). So if there are any doubts about homogeneity of variances it is better to apply Hotelling's T^2 for balanced designs[2]. In cases of unequal sample sizes one should search for the method which works best for a specific situation.

References

Anderson, T. W. (1958). *An introduction to multivariate analysis*. Wiley, New York.

Akritas, M., & S., A. (2000). Asymptotics for Analysis of Variance When the Number of Levels is Large. *Journal of the American Statistical Association*, *95*(449), 212-226.

Bathke, A. (2004). The ANOVA F test can still be used in some balanced designs with unequal variances and nonnormal data. *Journal of Statistical Planning and Inference*, *126*(449), 413-422.

Box, G. E. P. (1954b). Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, II. Effect of Inequality of Variances and of Correlation of Errors in the Two-Way Classification. *Annals of Mathematical Statistics*, *25*(3), 484-498.

E&B (2010). URL: http://www.hku.hk/ecology/staffhp/kl/Biom08.ppt [August 24, 2010], University of Hong Kong.

Heiberger, R. M., & Holland, B. (2004). *Statistical Analysis and Data Display*. Springer Sience+Business Media Inc.

Hotelling, H. (1931). The Generalization of Student's Ratio. *Annals of Mathematical Statistics*, *2*(3), 360-378.

Keppel, G., Saufley, W. H., Tokunaga, H., & Zedeck, S. (1992). *Introduction to design and analysis: A students handbook.* (second ed.). New York: W. H. Freemann.

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

[2] Script files for SPSS, SAS and R are available from the author

MESOSworld (2010). URL: http://www.mesosworld.ch/syllabi/zh-methpsy/uebungen/Uebungsserie9_Loesungen.pdf, [August 24, 2010], Methodological Education for the Social Sciences.

Moder, K. (2007). How to keep the Type I Error Rate in ANOVA if Variances are Heteroscedastic. *Austrian Journal of Statistics*, *36*(3), 179-188.

Rasch, D., Kubinger, K. D., & Moder, K. (2009). The two-sample t-test: pre-testing its assumptions does not pay. *Statistical Papers*. (ISSN 0932-5026 (Print) 1613-9798 (Online))

Rasch, D., Teuscher, F., & Guiard, V. (2007). How robust are tests for two independent samples in case of ordered categorical data? *Journal of Statistical Planning and Inference*, *133*, 2706-2720.

R Development Core Team. (2009). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria:. Available from http://www.R-project.org

SAS Institute Inc. (2008). SAS/STAT 9.2 User's Guide (Version 9.2 ed.) [Computer software manual]. Cary, NC, USA.: . (ISBN 1-58025-494-2)

Smaika, J. B. (1941). On an optimum property of two important statistical tests. *Biometrika 32*(1), 70-80

Singer, J. D. (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Educational and Behavioral Statistics*, *23*(4), 323-355.

Statlab (2005). URL: http://www.stat.ualberta.ca/statslabs/stat252/files/ins2.pdf [August 24, 2010], University of Alberta.

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis Test and Stochastic Homogeneity. *Journal of Educational and Behavioral Statistics*, *23*(2), 170-192.

Wellek, S. (2003). *Testing statistical hpotheses of equivalence*. Chapman & Hall/CRC Press LLC.