Psychology Science, Volume 49, 2007 (1), p. 2-12

A plea for more general tests than those for location only: Further considerations on Rasch & Guiard's 'The robustness of parametric statistical methods'

WOLFGANG T. WIEDERMANN[1] & RAINER W. ALEXANDROWICZ

Abstract

Starting with the discussion between Rasch & Guiard (2004) and von Eye (2004) concerning the use of parametric and nonparametric tests for the comparison of two samples a further approach toward this question is undertaken. Student's t-test requires for its application interval scaled and normally distributed data along with homogeneous variances across groups. In case that at least one of these prerequisites is not fulfilled, common statistical textbooks for social sciences usually refer to the non-parametric Wilcoxon-Mann-Whitney test. Earlier simulation studies revealed the t-test to be rather robust concerning distributional assumptions. The current study extends these findings with respect to the simultaneous violation of distributional and homogeneity assumptions. A simulation study has shown that both tests lead to highly contradicting results, and a more general approach toward the question of whether parametric or nonparametric procedures should be used, is introduced. Results indicate that the U-Test seems to be in general a more proper instrument for psychological research.

Key words: Parametric tests, nonparametric tests, non-normality, heteroscedasticity, power

[1] Correspondence concerning this article should be addressed to Wolfgang Wiedermann, University of Klagenfurt, Department of Psychology, Universitätsstrasse 65-67, 9020 Klagenfurt, Austria; email: wwiederm@edu.uni-klu.ac.at

1. Introduction

The comprehensive article of Rasch & Guiard (2004) provides a detailed overview concerning the robustness of parametric procedures. One aspect of their work leads to the conclusion that there is no further need for the Wilcoxon-Mann-Whitney test, that is why they recommend t-test due to its robustness. In a commentary, von Eye (2004) stresses the point that the exclusive use of the t-test might be inappropriate due to its sensitivity regarding autocorrelation of values. In their reply, Guiard & Rasch (2004) clearly worked out that the same is true for the nonparametric test and therefore conclude: "*Summarising we still think there are more disadvantages than advantages in using the Wilcoxon test in place of the t-test.*" (Guiard & Rasch, 2004, p. 553). This far-reaching conclusion raises the question whether the t-test might preserve its favourable properties even for more extreme distributional aberrations, which it was not designed for. In detail, we investigate the behavior of the t-test in comparison to the U-test in the presence of simultaneous violations of both distributional and homogeneity assumptions.

Rasch & Guiard (2004; referring to Posten, 1978) address the comparison of means (*i.e.* the "*location hypothesis*" H_0: $\mu_x = \mu_y$). Their results are based on distributions with positive and negative values of skewness over $0 \leq \gamma_1 \leq 2.0$ and kurtosis over $1.4 \leq \gamma_2 \leq 7.8$. We have decided to extend these values by using the lognormal distribution ($\gamma_1 = 6.19$; $\gamma_2 = 113.93$). Our choice reflects *(i)* the fact that in the context of psychological research (heavily) skewed distributions are likely to occur and *(ii)* depicts such extreme aberrant cases as mentioned above. Examples for lognormally distributed variables are given in e.g. Sachs & Hedderich (2006) and Limpert, Stahel, & Abbt (2001).

2. The two-sample problem

Given two independent, normally distributed samples and homogeneous variances together with interval scaled values, Student's t-test is the most powerful test (Pitman, 1948, as cited in Randles & Wolfe, 1979). Given the same conditions the asymptotic relative efficiency (ARE) of the most powerful nonparametric tests is .955 compared with the t-test, which means that sample size of the nonparametric procedure must be increased by about 4.5% to achieve the same efficiency as the parametric procedure (cf. Randles & Wolfe, 1979; Nikitin, 1995). Numerous studies have dealt with the adequacy of Student's t-test if at least one assumption is violated. In case of unequal variances it has been shown that Student's t-test is only robust if sample sizes are equal (cf. Hsu, 1938; Scheffé, 1970; Posten, Yeh & Owen, 1982; Tuchscherer & Pierer, 1985; Zimmerman & Zumbo, 1993a, 1993b; Bradstreet, 1997; Zimmerman, 2004). If both sample size and variances are unequal, the Welch t-test (Welch, 1938, 1947), which does not pool the variances, is referred to as an adequate procedure.

The same reactions are observable in the case of more than two groups, where it is well known that the ANOVA F-test is only usable for equal sample sizes (cf. Box, 1954). A procedure which overcomes the lack of robustness, if sample sizes differ is discussed in e.g. Welch (1951) and Brunner, Dette, & Munk (1997).

The parametric significance tests mentioned above depend on the normality assumption, which is – as Micceri (1989) impressively pointed out – rarely satisfied in practice. In this

case common textbooks (e.g. Aron, Aron, & Coups, 2006) refer to the nonparametric Wilcoxon-Mann-Whitney U-Test (Wilcoxon, 1945; Mann & Whitney, 1947), although theoretical findings by Bartlett (1935) and systematic examinations by e.g. Boneau (1960) and Posten (1978, 1984) emphasize the robustness of Student's t-test under non-normality.

Since the efficiency of parametric methods depends on the adequacy of the estimation of location parameters, Wilcox (1992, 2005a, 2005b) refers to the Yuen-Welch t-test (Yuen, 1974), which uses trimmed location and winsorized scale parameters for the computation of the test-statistic. Because trimming the samples reduces the impact of outliers, the risk of an inadequate estimation of location decreases. In order to take such corrective action into account too, this test will also be considered in the present study.

3. Method

A simulation study permits a systematic investigation of different violation conditions. For this purpose equally distributed pseudorandom numbers in the interval [0, 1] were generated by means of the *Mersenne-Twister* random number generator introduced by Matsumoto & Nishimura (1998). Standard normal variates were generated using the Box & Muller transformation (Box & Muller, 1958). These samples were used for the case of normally distributed data. For the analysis of skewed samples lognormal distributions were generated as follows: normally distributed samples were generated as described above and variables *x* and *y* underwent the following modifications: *(i)* x' = exp(x) and y' = exp(y); *(ii)* for each simulation condition k=10,000 samples of x' and y' were generated; *(iii)* the means $\hat{\mu}_{x'}$ and standard deviations $\hat{\sigma}_{x'}$ of these samples were averaged (obtaining the grand means $m^* = 1/k \sum_{i=1}^{k} \hat{\mu}_{x'}^{(i)}$ and $s^* = 1/k \sum_{i=1}^{k} \hat{\sigma}_{x'}^{(i)}$); *(iv)* each sample *x'* was standardized using these grand means, obtaining $x'' = \frac{x' - m^*}{s^*}$. The same procedure was applied to *y'*. This simulation technique is based on a contribution of Zimmerman (2004). It differs slightly from our procedure in the sense, that Zimmerman standardized *x'* and *y'* using the theoretical mean and standard deviation of the lognormal distributions instead of the empirical ones. Henceforth we use the symbols *x* and *y* for both *(a)* the original normal distributed samples and *(b)* for the lognormal samples *x"* and *y"* as described above. For a detailed description of non-normal variates and their higher moments see e.g. Evans, Hastings, & Peacock (2000).

The scores of the sample *x* were multiplied by a constant so that the ratio σ_x/σ_y had a predetermined value. After these modifications the samples were evaluated with Student's t-test, the Welch t-test, the Yuen-Welch t-test, and the Mann-Whitney U-test. Equal ($N_x = N_y$ = 10...(10)...100) and unequal (N_x [N_y] = 10 [20], 20 [50], 50 [100], 20 [10], 50 [20], 100 [50]) sample sizes were taken into account. All significance tests were nondirectional with a significance level of α = .05. For determining the robustness of the significance tests an ε-robustness of 20% was chosen. This means that a test is only robust if the relative frequency of rejecting the null hypothesis lies between .04 and .06 (cf. Rasch & Guiard, 2004). Each condition included 10,000 replications. Generation and analysis of the data was done in R 2.3.1 (R Development Core Team, 2006) and performed on a 1.60 GHz Intel Pentium M processor.

4. Results and discussion

4.1 Equal sample sizes

For normally distributed data and homogeneous variances all results were close to the nominal significance level. The Type I error rates for the Wilcoxon-Mann-Whitney-test are slightly above the significance level when sample size and the ratio σ_x/σ_y increase (Table 1, upper half).

The results for lognormal samples (Table 1, lower half) and equal variances indicate that the parametric procedures are quite robust for at least moderate sample sizes, whereas the Wilcoxon-Mann-Whitney test also preserves the nominal significance level for even small samples. For the case of unequal variances, the relative frequencies of rejecting H_0 of all tests under consideration generally increase with sample size and the ratio σ_x/σ_y. But the Yuen-Welch t-test and the Wilcoxon-Mann-Whitney test do this to a much larger extent than Student's t-test and the Welch-test.

Table 1: Relative frequencies of rejecting H_0 for equal sample sizes.
(Dist=Distribution, N=Normal, LN=Lognormal, $\boldsymbol{t_{st}}$ = Student, $\boldsymbol{t_w}$ = Welch, $\boldsymbol{t_{y.1}}$ = Yuen (10% trimmed), $\boldsymbol{t_{y.2}}$ = Yuen (20% trimmed), $\boldsymbol{U}$ = Mann-Whitney, α = 5%, non-robust results are marked *italic*.)

			σ_x/σ_y = **1**					σ_x/σ_y = **2**					σ_x/σ_y = **3**				
Dist.	N_x	N_y	t_{st}	t_w	$t_{y.1}$	$t_{y.2}$	U	t_{st}	t_w	$t_{y.1}$	$t_{y.2}$	U	t_{st}	t_w	$t_{y.1}$	$t_{y.2}$	U
N	10	10	.052	.051	.050	.051	.046	.055	.049	.054	.058	.053	.060	.052	.053	.058	*.061*
	20	20	.048	.048	.050	.051	.048	.055	.052	.051	.054	.059	.054	.051	.050	.055	*.067*
	30	30	.049	.049	.049	.051	.049	.053	.052	.054	.053	*.062*	.052	.049	.051	.051	*.069*
	40	40	.050	.050	.048	.050	.049	.055	.053	.054	.053	.060	.055	.053	.051	.051	*.067*
	50	50	.048	.048	.047	.049	.049	.050	.049	.052	.051	*.061*	.049	.048	.049	.050	*.070*
	60	60	.049	.049	.050	.049	.049	.050	.050	.050	.050	.057	.053	.051	.051	.051	*.068*
	70	70	.049	.049	.049	.052	.049	.053	.052	.048	.049	.057	.050	.049	.049	.048	*.067*
	80	80	.050	.050	.050	.049	.051	.051	.050	.052	.053	*.062*	.050	.049	.050	.049	*.068*
	90	90	.051	.051	.054	.051	.051	.052	.052	.051	.049	.059	.050	.049	.048	.050	*.067*
	100	100	.045	.045	.049	.049	.048	.052	.052	.051	.051	*.063*	.049	.048	.046	.046	*.066*
LN	10	10	*.035*	*.029*	*.030*	*.033*	.045	*.084*	*.079*	*.116*	*.137*	*.168*	*.125*	*.120*	*.177*	*.216*	*.237*
	20	20	*.036*	*.034*	*.035*	*.036*	.047	*.079*	*.078*	*.145*	*.196*	*.332*	*.110*	*.107*	*.217*	*.296*	*.443*
	30	30	*.038*	*.037*	*.037*	.040	.049	*.070*	*.069*	*.160*	*.227*	*.450*	*.095*	*.093*	*.245*	*.362*	*.592*
	40	40	.042	.040	.045	.045	.052	*.070*	*.069*	*.180*	*.270*	*.560*	*.091*	*.089*	*.292*	*.433*	*.706*
	50	50	.043	.042	.042	.042	.049	*.063*	*.062*	*.199*	*.309*	*.651*	*.084*	*.083*	*.324*	*.500*	*.800*
	60	60	.043	.042	.042	.046	.053	*.064*	*.064*	*.225*	*.358*	*.729*	*.080*	*.080*	*.357*	*.553*	*.864*
	70	70	.045	.045	.041	.043	.046	*.067*	*.066*	*.250*	*.398*	*.795*	*.075*	*.074*	*.393*	*.608*	*.908*
	80	80	.047	.047	.051	.051	.055	.060	.060	*.268*	*.434*	*.841*	*.070*	*.070*	*.433*	*.665*	*.940*
	90	90	.044	.044	.045	.047	.050	.056	.056	*.292*	*.473*	*.875*	*.073*	*.073*	*.462*	*.701*	*.957*
	100	100	.050	.049	.049	.045	.049	*.062*	*.061*	*.310*	*.503*	*.906*	*.069*	*.068*	*.487*	*.747*	*.974*

4.2 Unequal sample sizes

Table 2 shows the relative frequencies of rejecting H_0 for unequal sample sizes. In the case of normally distributed samples with homogeneous variances, the Type I error rates are close to the nominal significance level of $\alpha = 5\%$. As the SD ratio (σ_x/σ_y) increases, two reactions of Student's t-test can be observed: The Type I error rates rise far above the nominal significance level if the larger variance is associated with the smaller sample size. If the larger variance is associated with the larger sample size, the probability of a Type I error declines below the significance level. The Wilcoxon-Mann-Whitney test suffers from the same phenomenon (although to a lower extent), while the Welch t-test overcomes this problem entirely. For skewed samples with unequal variances the U-test as well as the Yuen-Welch t-test show quite similar but less extreme results compared with the case of equal sample sizes.

The results of the different significance tests can be summarised as follows. On the one hand Student's t-test is quite robust for non-normal data with homogeneous variances. This is consistent with the theoretical findings of Bartlett (1935) and simulation results of Posten (1978, 1984) and Rasch & Guiard (2004). The same is true for the Wilcoxon-Mann-Whitney test (Zimmerman & Zumbo 1993a; Zimmerman, 1998). Furthermore our results are in accordance with the well-known fact that Student's t-test preserves the nominal significance level for normally distributed samples with heterogeneous variances if sample sizes are equal (Posten, Yeh, & Owen 1982; Tuchscherer & Pierer, 1983), while the Type I error rates of the Wilcoxon-Mann-Whitney test are slightly above the nominal significance level (Trommer, 1965).

Table 2:
Relative frequencies of rejecting H_0 for unequal sample sizes. (Dist=Distribution, N=Normal, LN=Lognormal, t_{st} = Student, t_w = Welch, $t_{y.1}$ = Yuen (10% trimmed), $t_{y.2}$ = Yuen (20% trimmed), U = Mann-Whitney, α = 5%, non-robust results are marked *italic*.)

			$\sigma_x/\sigma_y = 1$					$\sigma_x/\sigma_y = 2$					$\sigma_x/\sigma_y = 3$				
Dist.	N_x	N_y	t_{st}	t_w	$t_{y.1}$	$t_{y.2}$	U	t_{st}	t_w	$t_{y.1}$	$t_{y.2}$	U	t_{st}	t_w	$t_{y.1}$	$t_{y.2}$	U
N	10	20	.049	.049	.049	.051	.047	*.113*	.053	.055	*.062*	*.086*	*.148*	.050	.054	*.061*	*.105*
	20	50	.051	.052	.054	.055	.051	*.136*	.051	.053	.054	*.092*	*.177*	.051	.052	.054	*.117*
	50	100	.051	.051	.050	.051	.050	*.109*	.050	.049	.050	*.082*	*.134*	.048	.048	.049	*.108*
	20	10	.052	.056	.054	.055	.051	*.018*	.044	.047	.048	*.030*	*.015*	.056	.053	.057	*.032*
	50	20	.051	.051	.051	.052	.049	*.011*	.050	.048	.048	*.024*	*.006*	.047	.049	.052	*.021*
	100	50	.052	.051	.053	.053	.050	*.017*	.050	.048	.048	*.032*	*.011*	.054	.052	.053	*.031*
LN	10	20	.044	.053	.050	.052	.060	*.136*	*.121*	*.160*	*.187*	*.280*	*.210*	*.144*	*.204*	*.239*	*.357*
	20	50	.048	*.068*	.057	.058	*.065*	*.159*	*.110*	*.181*	*.231*	*.466*	*.215*	*.118*	*.228*	*.312*	.575
	50	100	.047	.053	.049	.049	.053	*.128*	*.085*	*.229*	*.353*	*.753*	*.156*	*.087*	*.337*	*.514*	*.862*
	20	10	.040	.052	.050	.050	*.062*	.045	.041	*.085*	*.115*	*.186*	.058	*.076*	*.174*	*.248*	*.275*
	50	20	.049	*.064*	.057	.058	*.065*	*.033*	.044	*.125*	*.202*	*.373*	*.031*	.058	*.258*	*.431*	*.556*
	100	50	.045	.053	.051	.053	.059	*.030*	.050	*.242*	*.414*	*.781*	*.029*	*.061*	*.452*	*.713*	*.928*

If sample sizes are unequal the Type I error rates of Student's t-test rise above the nominal significance level if the larger variance is associated with the smaller sample, and declines if the larger variance is associated with the larger sample (Hsu, 1938; Boneau, 1960 and Posten et al., 1982). The Wilcoxon-Mann-Whitney test shows the same distortion but to a lesser extent (Zimmerman & Zumbo 1993a), while the Welch t-test protects the nominal significance level, which is in accordance with the results of Rasch & Guiard (2004), who therefore recommend the latter one for testing H_0: $\mu_x = \mu_y$ given unequal sample sizes.

For heterogeneous variances combined with skewed samples the parametric and non-parametric procedures show entirely different reactions (Stonehouse & Forrester, 1998, and Zimmerman, 2004). The fact that the U-test rejects the H_0 without any difference in means can be explained through the specific calculation processes of the tests. Student's t-test indeed compares the sample means, i.e. the difference in location, while the Wilcoxon-Mann-Whitney test examines the samples including all ranks, i.e. the difference in shape. That is the reason why this test is more robust against violations of normality. The heavily skewed distributions combined with unequal variances lead to a considerably increased rate of rejection of H_0: F(x) = F(y), which is exactly the H_0 of the Wilcoxon-Mann-Whitney U-test and reflects the differences in shape. The larger the ratio of standard deviations (e.g. $\sigma_x/\sigma_y = 3$), the less Student's t-test and the U-test are comparable – even if there are no violations of the normality assumption.

The Yuen-Welch t-test occupies a position between Student's t-test and the U-test. This test uses trimmed means, which leads to larger differences in means. Accordingly, the relative frequencies of rejecting H_0 are between those of Student's t-test and the U-test. The effect based on sample shape is also reflected in the differences in trimmed means, but to a lesser extent. A rejection of the null hypothesis of equal trimmed means (H_0: $\mu_{tx} = \mu_{ty}$) does not imply that the untrimmed means also differ. This further indicates that the probability of rejecting H_0 is also a function of trimming. These results differ from those found by Keselman, Wilcox, Kowalchuk, & Olejnik (2002) because in their studies the trimmed population means were transformed to be equal, while in the present simulation there was no difference between the untrimmed population means.

A closer inspection of how small a difference in standard deviations has a relevant impact upon the portion of significant results given lognormal distributions was performed in a further simulation. For that purpose very small ratios σ_x/σ_y (1.05, 1.10, and 1.15) were induced for equal sample sizes of 50..(50)..200 (results are given in table 3).

Table 3:
Relative frequencies of rejecting H_0 for lognormal-distributed samples. ($\boldsymbol{t_{st}}$ = Student, $\boldsymbol{t_w}$ = Welch, $\boldsymbol{t_{y.1}}$ = Yuen (10% trimmed), $\boldsymbol{t_{y.2}}$ = Yuen (20% trimmed), $\boldsymbol{U}$ = Mann-Whitney, α = 5%, non-robust results are marked *italic*.)

		σ_x/σ_y = **1.05**					σ_x/σ_y = **1.10**					σ_x/σ_y = **1.15**				
N_x	N_y	t_{st}	t_w	$t_{y.1}$	$t_{y.2}$	U	t_{st}	t_w	$t_{y.1}$	$t_{y.2}$	U	t_{st}	t_w	$t_{y.1}$	$t_{y.2}$	U
50	50	.041	.040	.043	.047	.058	.043	.042	.046	.050	*.076*	.045	.043	.051	*.061*	*.115*
100	100	.044	.044	.051	.054	*.069*	.048	.047	.054	*.063*	*.116*	.049	.049	*.062*	*.073*	*.178*
150	150	.046	.045	.048	.052	*.077*	.047	.047	.057	*.066*	*.142*	.047	.047	*.066*	*.084*	*.245*
200	200	.047	.047	.049	.053	*.079*	.046	.046	.060	*.072*	*.176*	.050	.050	*.075*	*.100*	*.322*

Even such negligible ratios of standard deviations lead to a systematic increase of rejections of a "false" or, as the case may be "another" null hypothesis by the U-test. The relative frequency of rejecting H_0 is a function of both the ratio of standard deviations and the sample size, which is in accordance with e.g. Zimmerman (2003). Furthermore even in case of moderate sample sizes and irrelevant SD ratios (such as $N_x = N_y = 100$ and $\sigma_x/\sigma_y = 1.05$) the proportion of significant results rises markedly above the significance level of $\alpha = .05$. Even for small samples ($N_x = N_y = 25$) the probability of rejecting H_0 increases far above the nominal significance level (*cf.* figure 1).

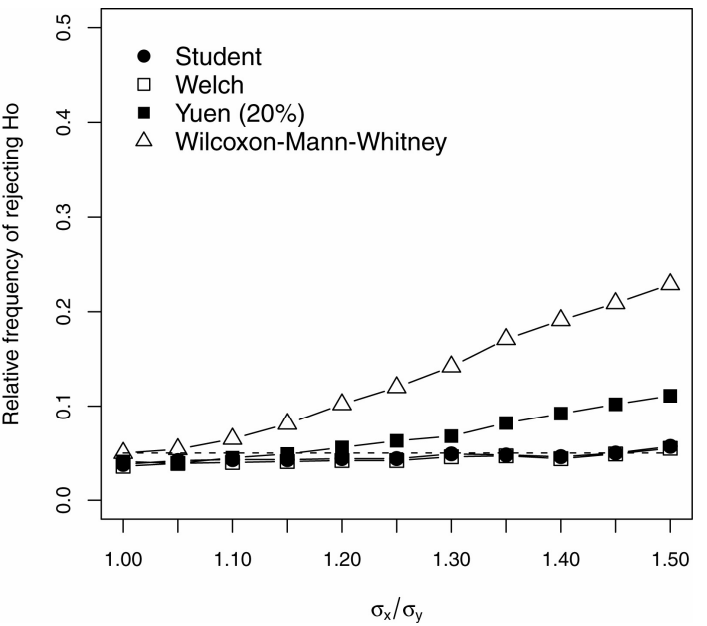

Figure 1:
Relative frequencies of rejecting H_0 as a function of the ratio of standard deviations for lognormal samples ($N_x = N_y = 25$)

5. Conclusion

Numerous previous studies have analyzed the robustness of Student's t-test and the Wilcoxon-Mann-Whitney test. Rasch & Guiard (2004) mention that the Wilcoxon-Mann-Whitney test can be used for testing H_0: $\mu_x = \mu_y$ if samples do not differ in higher moments and they conclude that there is no need to use the U-test because Student's t-test is robust for non-normal distributions.

The aim of the present study was to test whether the conclusion of Rasch & Guiard (2004) that the t-test is robust under non–normality conditions can be generalized to the case of extreme non–normality along with heterogeneity of variances. The results clearly show

that parametric and nonparametric tests lead to different decisions: e.g. given a σ_x/σ_y of 1.1 and a sample of $N_x = N_y = 100$, Student's t-test has 4.8% significant results while the Wilcoxon-Mann-Whitney U-test rejects H_0 more than twice that often (11.6%). In terms of robustness Student's t-test can be regarded as robust, while the Wilcoxon-Mann-Whitney U-test cannot. But the non-robustness of the latter one could be seen from another direction: we want to raise the question whether the location hypothesis (H_0: $\mu_x = \mu_y$) of Student's t-test is what psychologists really are interested in? Maybe a psychological investigation is being considered successful when researchers are able to find a more general kind of difference. Consider the case that an intervention study using a two-group design leads to a higher proportion of patients showing an improvement but the distribution becomes skewed. The difference in means might underestimate the true effect of this intervention due to a remaining portion of patients showing no improvement or even a worsening. Similarly, a training program can be effective for some of the students (in this case an effect is given from a global point of view), but for others not. (Of course, more complex analyses might be indicated, e.g. estimation of a mixture model, but we would like to restrict our considerations to the amply used two-sample comparison). From a technical perspective psychologists are rather interested in the H_0: $F(x) = F(y)$ than H_0: $\mu_x = \mu_y$. From that point of view the higher proportion of significant results of the Wilcoxon-Mann-Whitney U-test mentioned above could also be seen as the higher power of the test to find differences aside of location.

Moreover, as Micceri (1989) convincingly argued the normality assumption is seldomly realized. In addition, slightest departures from $\sigma_x/\sigma_y = 1$ are likely to occur, so the results presented here seem to apply more often than one might expect. Obviously, this effect increases with both sample size and skewness. Regarding the interests of psychological research our conclusion would be that the U-test seems superior for general usage. Here we are in line with Siegel (1956), who wrote "*I believe that the nonparametric techniques of hypothesis testing are uniquely suited to the data of the behavioral sciences*" (p. vii).

Another important fact is that the t-test – although robust – loses power for various non-normal distributions. The ARE of the Wilcoxon-Mann-Whitney test with respect to Student's t-test is 1 for uniformly distributed samples, 1.097 for logistic distributions and 1.5 for Laplace (double exponential) distributions (cf. Randles & Wolfe, 1979). This power advantage has repeatedly been shown by various simulation studies (e.g. Blair & Higgins, 1980; Posten, 1982; Zimmerman & Zumbo, 1990; Zimmerman & Zumbo, 1993a). And even if the seldom case of normal distributions along with homogeneous variances should occur and the t-test would be indicated then, the U-test used instead still exhibits an ARE of 95.5%. Considering the modification of the U-Test proposed by Berchtold (1979), an ARE of 99.2% is possible, which seems rather useful for psychological research.

Acknowledgement

We would like to thank the anonymous reviewers for their valuable comments on an earlier version of this article.

References

Aron, A., Aron, E. N., & Coups, E. J. (2006). Statistics for Psychology (4th ed.). New Jersey: Pearson Education Inc.

Bartlett, M. S. (1935). The effect of non-normality on the t-distribution. Proceedings of the Cambridge Philosophical Society 31, 223 - 231.

Berchtold, H. (1979). A modified Mann-Whitney test with improved asymptotic relative efficiency. Biometrical Journal 21, 649-655.

Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's *t* statistic under various non-normal distributions. Journal of Educational Statistics 5, 309-335.

Boneau, C. A. (1960). The effects of violation of assumptions underlying the t-test. Psychological Bulletin 57, 49-64.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variances. The annals of mathematical statistics 25, 290-302.

Box, G. E. P., & Muller, M. (1958). A note on the generation of normal deviates. Annals of Mathematical Statistics 29, 610-611.

Bradstreet, T. E. (1997). A Monte Carlo Study of Type I error rates for the two-sample Behrens-Fisher problem with and without rank transformation. Computational Statistics & Data Analysis 25, 167-179.

Brunner, E., Dette, H., & Munk, A. (1997). Box-type approximations in nonparametric factorial designs. Journal of the American Statistical Association 92, 1494-1502.

Evans, M., Hastings, N., & Peacock, B. (2000). Statistical Distributions (3rd ed.). New York: Wiley.

Guiard, V., & Rasch, D. (2004). The robustness of two sample tests for means. A reply on von Eye's comment. Psychology Science 46 (4), 549-554.

Hsu, P. L. (1938). Contributions to the theory of Student's t-Test as applied to the problem of two samples. Statistical Research Memoirs 2, 1-24.

Keselman, H. J., Wilcox, R. R., Kowalchuk, R. K., & Olejnik, S. (2002). Comparing trimmed or least square means of two independent skewed populations. Biometrical Journal 44, 478-489.

Limpert, E., Stahel W.A., & Abbt M. (2001). Log-normal Distributions across the Sciences: Keys and Clues. BioScience 51 (5), 341–352.

Mann, H.B., & Whitney, D.R. (1947). On a Test whether one of two Random Variables is Stochastically Larger than the Other. Annals of Mathematical Statistics 18, 50-60.

Matsumoto, M., & Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Transactions on Modeling and Computer Simulation 8, 3-30.

Micceri, T. (1989). The Unicorn, The Normal Curve and Other Improbable Creatures. Psychological Bulletin 105 (1), 156-166.

Nikitin, Y. (1995). Asymptotic efficiency of nonparametric tests. Cambridge: Cambridge University Press.

Posten, H. O. (1978). The Robustness of the two sample t-Test over the Pearson System. Journal of Statist. Comput. and Simulation 6, 295-311.

Posten, H. O. (1982). Two-sample Wilcoxon power over the Pearson system and comparison with the t-test. Journal of Statist. Comput. and Simulation 16, 1-18.

Posten, H. O., Yeh, H.C., & Owen, D.B. (1982). Robustness of the two-sample t-test under violations of the homogeneity of variance assumptions. Communications in Statistics: Theory and Methods 11, 109-126.

Posten, H. O. (1984). Robustness of the two-sample t-Test. In Rasch, D. & Tiku, M. L. (Ed.), Robustness of statistical methods and nonparametric statistics (S. 92-99). Dordrecht: D. Reidel Publishing Company

R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (URL http://www.R-project.org.)

Randles, R. H., & Wolfe, D. A. (1979). Introduction to the theory of nonparametric statistics. New York: Wiley.

Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods. Psychology Science 46 (2), 175-208.

Sachs L., & Hedderich J. (2006). Angewandte Statistik. [Applied Statistics]. 12th ed. Berlin: Springer.

Scheffé, H. (1970). Practical Solutions of the Behrens-Fisher Problem. Journal of the American Statistical Association 65 (332), 1501-1508.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill.

Stonehouse, J. M., & Forrester, G. J. (1998). Robustness of the t and U tests under combined assumption violations. Journal of Applied Statistics 25 (1), 63-74.

Trommer, R. (1965). Untersuchungen zur Robustheit des Wilcoxon-Tests gegenüber Streuungsungleichheit. [Investigations on robustness of the Wilcoxon-test against inequality of variances]. Biometr. Z. 9 (1), 14-21.

Tuchscherer, A., & Pierer, H. (1985). Simulationsuntersuchungen zur Robustheit verschiedener Verfahren zum Mittelwertsvergleich im Zweistichprobenproblem (Simulationsergebnisse). [Simulation studies on robustness of several methods for the comparison of means in the two-sample problem]. In Rudolph, P. E. (Hrsg.), Robustheit V – Arbeitsmaterial zum Forschungsthema Robustheit. Probleme der angewandten Statistik, Heft 15, S. 1-42, Dummersdorf-Rostock.

von Eye, A. (2004). Robustness is parameter-specific. A comment on Rasch and Guiard's robustness study. Psychology Science 46 (4), 544-548.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. Biometrika 29, 350-362.

Welch, B. L. (1947). The generalisation of Student's problem when several different population variances are involved. Biometrika 34, 28-35.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. Biometrika 38 (3/4), 330-336.

Wilcox, R. (1992). Why can methods for comparing means have relatively low power, and what can you do to correct the problem. Current Direction in Psychological Science 1 (3), 101-105.

Wilcox, R. (2005a). Introduction to Robust Estimation and Hypothesis Testing (2nd ed.). San Diego: Elsevier Academic Press.

Wilcox, R. (2005b). New methods for comparing groups. Current Directions in Psychological Science 14 (5), 272-275.

Wilcoxon, F. (1945). Individual comparison by ranking methods. Biometrics 3, 80-82.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. Biometrika 61 (1), 165-170.

Zimmerman, D. W., & Zumbo, B. (1990). The relative power of the Wilcoxon-Mann-Whitney test and Student *t* test under simple bounded transformations. The Journal of General Psychology 117 (4), 425-436.

Zimmerman, D. W., & Zumbo, B. (1993a). The relative power of parametric and nonparametric statistical methods. In Kernen, G., Lewis, C. (Ed.), A Handbook for data analysis in the

behavioral sciences: methodological issues (S. 481-517). New Jersey: Lawrence Erlbaum Associates.

Zimmerman, D. W., & Zumbo B. (1993b). Rank transformations and the power of the Student t-test and Welch t-test for non-normal populations with unequal variances. Canadian Journal of Experimental Psychology 47 (3), 523-539.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions, Journal of Experimental Education 67 (1), 55-68.

Zimmerman, D. W. (2003). A warning about the large sample Wilcoxon-Mann-Whitney test. Understanding Statistics 2 (4), 267-280.

Zimmerman, D.W. (2004). Inflation of Type I error rates by unequal variances associated with parametric, nonparametric, and rank transformation tests. Psicológica 25, 103-133.