Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 15, Number 11, October, 2010, Addendum added October 2012 ISSN 1531-7714

Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon

Joost C. F. de Winter and Dimitra Dodou
Department of BioMechanical Engineering, Delft University of Technology

Likert questionnaires are widely used in survey research, but it is unclear whether the item data should be investigated by means of parametric or nonparametric procedures. This study compared the Type I and II error rates of the *t* test versus the Mann-Whitney-Wilcoxon (MWW) for five-point Likert items. Fourteen population distributions were defined and pairs of samples were drawn from the populations and submitted to the *t* test and the *t* test on ranks, which yields the same results as MWW. The results showed that the two tests had equivalent power for most of the pairs. MWW had a power advantage when one of the samples was drawn from a skewed or peaked distribution. Strong power differences between the *t* test and MWW occurred when one of the samples was drawn from a multimodal distribution. Notably, the Type I error rate of both methods was never more than 3% above the nominal rate of 5%, even not when sample sizes were highly unequal. In conclusion, for five-point Likert items, the *t* test and MWW generally have similar power, and researchers do not have to worry about finding a difference whilst there is none in the population.

Likert scales are widely used in various domains such as behavioral sciences, healthcare, marketing, and usability research. When responding to a Likert scale, participants specify their level of agreement to statements with typically five or seven ordered response levels. Likert item data have distinct characteristics: discrete instead of continuous values, tied numbers, and restricted range.

There exists disagreement amongst scholars about whether Likert data should be analyzed with parametric statistics such as the *t* test or nonparametric statistics such as the rank-based Mann-Whitney-Wilcoxon (MWW) (Carifio & Perla, 2008; Jamieson, 2004). Clason and Dormody (1994) investigated the use of Likert items in the *Journal of Agricultural Education* and found that out of 95 relevant articles 13% used a nonparametric test and 34% a parametric one. In a simulation study with five-point Likert items, Gregoire and Driver (1987) did not find a clear preference towards either the *t* test or nonparametric counterparts, but a reanalysis by Rasmussen (1989) pointed to flaws in that study and concluded that parametric tests are more powerful (i.e., exhibit a lower Type II error rate), except when the sample pairs are taken from the most nonnormal combination of distributions, namely from a uniform distribution and a mixed-normal one. No large differences were found between parametric and nonparametric tests regarding the occurrence of false positives (i.e., the Type I error rate). A later study of Nanna and Sawilowsky (1998) using seven-point Likert item data found greater power for MWW in almost all investigated cases. The difference between the results of Rasmussen and those of Nanna and Sawilowsky can be explained by the types of distributions analyzed. The former study investigated relatively normal distributions, whereas the latter used distributions of real data which were considerably skewed.

There exists a wealth of literature comparing the *t* test with MWW, most often focusing on relatively simple and continuous distributions and not on complex, truncated, and discrete distributions such as those of Likert data. It is well established that the *t* test has a power advantage for normal distributions with equal variances and that it is robust to modest deviations from the test assumptions (Baker, Hardyck, & Petrinovich, 1966; Glass, Peckham, & Sanders, 1972;

Heeren & D'Agostino, 1987; Posten, Yeh, & Owen, 1982; Rasch & Guiard, 2004, Sawilowsky & Blair, 1992; Sawilowsky & Hillman, 1992; Stonehouse & Forrester, 1998; Sullivan & D'Agostino, 1992; Wetherill, 1960). For highly nonnormal distributions, on the other hand, such as exponential and lognormal distributions or distributions with outliers, MWW has a power advantage (Blair & Higgins, 1980; Bridge & Sawilowsky, 1999; MacDonald, 1999; Neave & Granger, 1968). For both the *t* test and MWW, the Type I error rate deviates from the nominal value when unequal variances are combined with unequal sample sizes or when unequal variances are combined with nonnormal distributions (Fagerland & Sandvik, 2009; MacDonald, 1999; Stonehouse & Forrester, 1998; Zimmerman, 2006). In such cases, separate-variance procedures such as the Welch test are recommended as being more Type I error robust (Cribbie & Keselman, 2003; Ruxton, 2006; Zimmerman, 2006).

The effect of sample size on the performance of the *t* test and MWW is complex. Blair and Higgins (1980) found that in some conditions results for small sample sizes ($n < 10$) were opposite to those for moderate sample sizes. Many textbooks and articles mention that nonparametric tests are preferred when sample size is small and that the *t* test becomes superior when sample size increases, as a result of the Central Limit Theorem (Lumley, Diehr, Emerson, & Chen, 2002). The simulation study by Nanna and Sawilowsky (1998), however, showed that MWW achieves increased power advantages as sample size increases, indicating that it should be used not only for small but also for large sample sizes.

There exists some disagreement regarding the hypothesis being tested with MWW. MWW is often interpreted as a test of equal medians or a test of equal distributions. The correct interpretation, however, is that MWW is identical to performing a *t* test after ranking over the combined samples (Conover & Iman, 1981; Fagerland & Sandvik, 2009; Zimmerman & Zumbo, 1993). Thus, the *t* test assesses differences in means, whereas MWW assesses differences in mean ranks.

The aim of this study was to investigate whether the *t* test or MWW should be used when comparing two independent samples of five-point Likert data. A simulation study was conducted to test the hypothesis whether pairs of samples were drawn from different population distributions. We assessed the power of a test by its ability to detect whether the samples were drawn from different population distributions, regardless of the expected value of the distribution. The Type I robustness was evaluated by testing samples drawn from the same distribution. The analyses were repeated for a range of equal and unequal sample sizes.

METHOD

Fourteen diverse Likert population distributions were defined (Table 1). These distributions were considered representative for the possible distributions that may occur in real Likert item data. Ten thousand random samples were drawn for each of the 98 distribution combinations and subjected to the *t* test and MWW. MWW was conducted by first transforming the combined vector of the two samples to ranks (Conover & Iman, 1981; Fagerland & Sandvik, 2009; Zimmerman & Zumbo, 1993). An average rank was assigned in case the ranks were equal. The simulations were conducted for equal sample sizes ($n = m = 10$, $n = m = 30$, and $n = m = 200$) and for unequal sample sizes ($n = 5$, $m = 20$, and $n = 100$, $m = 10$). The computer simulation code is included in Appendix 2.

RESULTS

Table 2 shows the results of the simulation study for $n = m = 10$. Values for all investigated conditions are shown in Appendix 1. The above-diagonal numbers are the Type II error rates for the *t* test and the numbers below the diagonal are the Type II error rates for MWW minus the Type II error rates for the *t* test. A positive number below the diagonal means that the *t* test was more powerful than MWW.

For $n = m = 10$, the Type II error rate was generally high, except for the samples which were from very different distributions, such as the *(very) strongly agree* versus *(very) strongly disagree* distributions. The power differences between the *t* test and MWW were less than 5% for the majority of the pairs. MWW occasionally had a substantial power advantage when one of the samples was from the *very strongly disagree* or *very strongly agree* distributions, with a maximum of 8% for *very strongly agree* versus *disagree flat*. The *t* test had a large power advantage when *strong multimodal* was compared with *(very) strongly (dis)agree*, with a maximum difference of 21%.

For $n = m = 30$, the Type II error rate was obviously lower than for $n = m = 10$. The pattern of power differences between the *t* test and MWW was roughly equivalent. For sample pairs for which the difference in

Table 1. The population distributions of the five-point Likert data used in the simulation study

	1 (%)	**2 (%)**	**3 (%)**	**4 (%)**	**5 (%)**	**Mean**	***SD***	**Skewness**	**Kurtosis**
Very strongly agree	0	1	3	6	90	4.85	0.50	-3.70	17.03
Strongly agree	1	3	6	30	60	4.45	0.82	-1.76	6.35
Agree peak	5	10	20	45	20	3.65	1.07	-0.77	3.08
Agree flat	10	15	20	30	25	3.45	1.29	-0.46	2.12
Neutral to agree	10	20	30	25	15	3.15	1.20	-0.11	2.15
Neutral peak	0	20	50	20	10	3.20	0.88	0.51	2.68
Neutral flat	15	20	25	20	20	3.10	1.34	-0.06	1.86
Very strongly disagree	80	12	4	3	1	1.33	0.78	2.70	10.19
Strongly disagree	70	20	6	3	1	1.45	0.82	2.09	7.29
Disagree flat	25	35	20	15	5	2.40	1.16	0.53	2.37
Neutral to disagree	10	25	30	20	15	3.05	1.21	0.08	2.10
Certainly not disagree	1	4	50	30	15	3.54	0.83	0.19	2.91
Multimodal	15	5	15	25	40	3.70	1.42	-0.83	2.37
Strong multimodal	45	5	0	5	45	3.00	1.93	0	1.06

Note. The kurtosis of a normal distribution is 3. Distributions that are more outlier-prone than a normal distribution have kurtosis greater than 3; distributions that are less outlier-prone have kurtosis less than 3.

means (or the difference in mean ranks) was large, the Type II error rate was close to 0%. In that case, the null hypothesis of equal samples was almost always rejected by both the *t* test and MWW, obscuring power differences between the methods. MWW was particularly powerful when one of the pairs was from the *very strongly agree* or *neutral peak* distribution, with a power advantage up to 19%. The *t* test was more powerful when the *strong multimodal* distribution was compared with a skewed or peak distribution, with power advantages up to 26%.

For $n = m = 200$, the Type II error rate was 0% for most of the pair comparisons, as one could have expected. In the cases it was not, MWW exhibited large power advantages. For example, for *multimodal* versus *certainly not disagree*, the Type II error rate was 71% for the *t* test and only 9% for MWW. The *t* test exhibited a large power advantage again when the *strong multimodal* distribution was involved.

For unequal sample sizes it was interesting to observe that the pattern of power differences approximately corresponded to the equal sample size conditions. Also, the Type I error rate was close to the nominal value of 5%. Only in one case was the Type I error rate greater than 6% (when comparing two samples of unequal sizes from the *very strongly agree* distribution).

Table 3 shows the Cohen's *d* effect sizes of the difference in means minus the difference in mean ranks. This provides an indication of the theoretical power difference between the *t* test and MWW, independent from the ceiling effects in the Type II error rate (i.e., rates of 0% or 100%, respectively). It can be seen that the difference in means was larger than the difference in means ranks particularly for the *strong multimodal* distributions and for *(very) strongly agree* versus *(very) strongly disagree*, which corresponds to the results in Table 2.

Table 2. Type II error percentage for the *t* test (above diagonal), Type II error percentage for MWW minus Type II error percentage for the *t* test (below diagonal), and Type I error percentages for the *t* test and MWW (two rightmost columns)

		Type II errors														Type I errors	
		Very strongly agree	*Strongly agree*	*Agree peak*	*Agree flat*	*Neutral to agree*	*Neutral peak*	*Neutral flat*	*Very strongly disagree*	*Strongly disagree*	*Disagree flat*	*Neutral to disagree*	*Certainly not disagree*	*Multimodal*	*Strong multimodal*		
	$n = m = 10$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	*t* test	MWW
1.	*Very strongly agree*		75	12	12	3	1	4	0	0	0	2	5	34	21	0.8	0.9
2.	*Strongly agree*	-6		54	49	25	16	28	0	0	2	20	36	75	48	4.4	5.1
3.	*Agree peak*	-5	-5		93	83	80	83	1	1	36	78	93	94	86	5.1	5.1
4.	*Agree flat*	-3	1	0		91	90	91	3	5	56	88	94	92	91	5.3	5.4
5.	*Neutral to agree*	-1	-1	0	0		95	95	6	9	72	94	88	84	94	5.0	4.9
6.	*Neutral peak*	0	-2	-4	-2	0		94	2	3	61	94	86	81	94	5.1	5.1
7.	*Neutral flat*	-1	2	1	0	0	-1		9	13	77	95	88	83	94	5.1	5.1
8.	*Very strongly disagree*	0	0	0	0	-2	-1	-2		95	35	8	0	3	36	2.2	3.3
9.	*Strongly disagree*	0	0	0	0	-1	-1	0	-2		46	11	1	4	43	4.1	5.3
10.	*Disagree flat*	0	0	0	0	0	-2	0	-8	-4		78	34	45	87	5.1	5.2
11.	*Neutral to disagree*	0	1	0	0	0	0	0	-3	-2	0		83	79	94	5.3	5.4
12.	*Certainly not disagree*	-2	-5	-2	-1	1	-1	0	0	0	1	0		91	89	5.1	5.3
13.	*Multimodal*	-6	3	-2	-1	-3	-6	-1	1	2	1	-3	-4		85	4.9	4.9
14.	*Strong multimodal*	6	21	4	0	-2	-3	-2	11	16	2	-3	1	4		5.0	5.3

Note. A positive number below the diagonal means that the *t* test was more powerful than MWW.

Table 3. Absolute of Cohen's *d* effect size of difference between untransformed data minus absolute of Cohen's *d* effect size of difference between rank-transformed data

		Very strongly agree	*Strongly agree*	*Agree peak*	*Agree flat*	*Neutral to agree*	*Neutral peak*	*Neutral flat*	*Very strongly disagree*	*Strongly disagree*	*Disagree flat*	*Neutral to disagree*	*Certainly not disagree*	*Multimodal*
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
2.	*Strongly agree*	-0.13												
3.	*Agree peak*	-0.39	-0.10											
4.	*Agree flat*	-0.27	-0.01	0.04										
5.	*Neutral to agree*	-0.33	-0.06	-0.01	-0.02									
6.	*Neutral peak*	-0.19	-0.08	-0.10	-0.07	0.04								
7.	*Neutral flat*	-0.24	0.01	0.02	0.00	0.00	0.03							
8.	*Very strongly disagree*	0.49	0.42	-0.05	-0.13	-0.20	-0.33	-0.17						
9.	*Strongly disagree*	0.53	0.40	0.00	-0.05	-0.11	-0.24	-0.07	-0.07					
10.	*Disagree flat*	-0.33	-0.03	-0.01	0.00	-0.01	-0.02	0.01	-0.19	-0.10				
11.	*Neutral to disagree*	-0.27	-0.03	-0.02	-0.02	-0.01	0.01	-0.00	-0.25	-0.15	0.00			
12.	*Certainly not disagree*	-0.22	-0.11	-0.12	0.05	0.05	-0.03	0.05	-0.09	-0.09	-0.01	0.02		
13.	*Multimodal*	-0.16	0.08	-0.14	-0.06	-0.08	-0.14	-0.04	0.04	0.09	0.00	-0.06	-0.20	
14.	*Strong multimodal*	0.09	0.30	0.24	0.11	0.04	0.04	0.01	0.16	0.24	0.16	-0.02	0.20	0.14

Note. A positive value indicates that the difference between means was larger than the difference between rank-transformed means. Contrary, a negative value indicates that the rank transformation increased the size of the difference as compared to the untransformed data. The calculations were conducted with a very large sample size ($n = m = 100{,}000$) such that the effects represent the differences at the population level.

DISCUSSION

We investigated whether the *t* test or MWW is superior for comparing two samples of Likert data. This study differed from previous ones by focusing on five-point Likert items and by simulating an extensive variety of possible distributions.

The results showed that the power differences between the *t* test and MWW were minor and exceeded 10% for only few of the 98 distribution pairs. In many cases, the Type II error rate of the *t* test and MWW was close to 0%, indicating that differences between samples were large enough to be detected at the $\alpha = .05$ level by either method. MWW excelled for skewed or peaked distributions, whereas the *t* test was superior in some cases involving multimodal distributions. The *t* test was found to be superior to MWW also for severe violations from the test assumptions (such as when comparing samples from the *strong multimodal* with *strongly agree* distribution). This can be explained by the fact that the difference in means of these populations was larger (Cohen's $d = 0.98$) than the difference in mean ranks (Cohen's $d = 0.68$).

Another noteworthy result of this study was that the Type I error rate was close to the nominal value of 5% for all sample sizes and for all combinations of distributions. These results indicate that, for both the *t* test and MWW, researchers working with Likert item data do not have to be worried about finding a difference

when there is actually none in the population. Zimmerman (2006) found that the Type I error rate can strongly deviate when samples of unequal size are tested with equal means but unequal variances. The difference with our study is that we evaluated the Type I error rate for samples from the exact same distribution.

This article extends upon the results of Nanna and Sawilowsky (1998) who found that MWW was superior in all investigated cases. The difference with our results can be explained by the fact that Nanna and Sawilowsky used seven pairs of distributions of real data which were all relatively skewed and therefore favored MWW. Our study used 98 distribution pairs which constituted a more diverse set of possible distributions. Furthermore, Nanna and Sawilosky used seven-point Likert data which allows for longer tails and more skewness than a five-point interval. Our results agree with Nanna and Sawilowsky regarding the effect of sample size. Increasing the sample size increases the power of the *t* test and MWW but does not result in the *t* test becoming the preferred method.

This article focused on five-point Likert items only. It is known that item reliability is affected by the number of points (depending on the distribution) and various recommendations exist ranging from 2–3 points (Matell & Jacoby, 1971) to 7–10 points (Preston & Colman, 2000). Although the analysis of single Likert items is often discouraged (e.g., Carifio & Perla, 2008), it has been shown that well-designed items can be as appropriate as multiple-item scales in terms of construct validity (Gardner, Cummings, Dunham, & Pierce, 1998). The simulation techniques used in this study can also be applied to other types of distributions, such as Likert items that are combined into a total score. Summed scores generally have a more normal distribution than single items, which may work in favor of the *t* test.

Zimmerman (2004) concluded that optimum protection of the Type I error rate is assured by using the Welch test whenever sample sizes are unequal. The present study used no corrections for heteroscedasticity, such as the Welch test. Therefore, the reported power advantages of the *t* test may be inflated due to the pairing of unequal sample sizes and variances. We verified this by repeating all *t* tests with the unequal variances option, known as the Behrens-Fisher problem. The results showed that the average Type I and Type II error rates were actually considerably higher than the corresponding error rates with the regular *t* test. Hence, for Likert data, the regular *t* test is to be recommended over the unequal variances *t* test.

In conclusion, the *t* test and MWW generally have equivalent power, except for skewed, peaked, or multimodal distributions for which strong power differences between the two tests occurred. The Type I error rate of both methods was never more than 3% above the nominal rate of 5%, even not when sample sizes were highly unequal.

REFERENCES

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement, XXVI*, 291–309.

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 nonnormal distributions. *Journal of Educational Statistics, 5*, 309–335.

Bridge, P. D., & Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: Comparative power of the t-test and Wilcoxon rank-sum test in small samples applied research. *Journal of Clinical Epidemiology, 52*, 229–235.

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education, 42*, 1150–1152.

Clason, D. L., & Dormody, T. J. (1994). Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education, 35*, 31–35.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35*, 124–133.

Cribbie, R. A., & Keselman, H. J. (2003). The effects of nonnormality on parametric, nonparametric, and model comparison approaches to pairwise comparisons. *Educational and Psychological Measurement, 63*, 615–635.

Fagerland, M. W., & Sandvik, L. (2009). The Wilcoxon-Mann-Whitney test under scrutiny. *Statistics in Medicine, 28*, 1487–1497.

Gardner, D. G., Cummings, L. L., Dunham, R. B., & Pierce, J. L. (1998). Single-item versus multiple-item measurement scales: An empirical comparison. *Educational and Psychological Measurement, 58*, 898–915.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*, 237–288.

Gregoire, T. G., & Driver, B. L. (1987). Analysis of ordinal data to detect population differences. *Psychological Bulletin, 101*, 159–165.

Heeren, T., & D'Agostino, R. (1987). Robustness of the two independent samples *t*-test when applied to ordinal scaled data. *Statistics in Medicine, 6*, 79–90.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*, 1212–1218.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23*, 151–169.

MacDonald, P. (1999). Power, Type I, and Type III error rates of parametric and nonparametric statistical tests. *The Journal of Experimental Education, 67*, 367–379.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement*, *31*, 657–674.

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods, 3*, 55–67.

Neave, H. R., & Granger, C. W. J. (1968). A Monte Carlo study comparing various two-sample tests for differences in mean. *Technometrics, 10*, 509–522.

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

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1–15.

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

Rasmussen, J. L. (1989). Analysis of Likert-scale data: A reinterpretation of Gregoire and Driver. *Psychological Bulletin, 105*, 167–170.

Ruxton, G. D. (2006). The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann-Whitney *U* test. *Behavioral Ecology, 17*, 688–690.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111*, 352–360.

Sawilowsky, S. S., & Hillman, S. B. (1992). Power of the independent samples *t* test under a prevalent psychometric measure distribution. *Journal of Consulting and Clinical Psychology, 60*, 240–243.

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

Sullivan, L. M., & D'Agostino, R. B. (1992). Robustness of the *t* test applied to data distorted from normality by floor effects. *Journal of Dental Research, 71*, 1938–1943.

Wetherill, G. B. (1960). The Wilcoxon test and non-null hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological), 22*, 402–418.

Zimmerman D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*, 173–181.

Zimmerman, D. W. (2006). Two separate effects of variance heterogeneity on the validity and power of significance tests of location. *Statistical Methodology, 3*, 351–374.

Zimmerman, D. W., & Zumbo, B. D. (1993). The relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 481–517). Hillsdale, NJ: Erlbaum.

Citation:

de Winter, J. C. F. and D. Dodou (2012). Five-Point Likert Items: *t* test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation*, 15(11). Available online: http://pareonline.net/getvn.asp?v=15&n=11.

Note:

This research was supported by the Dutch Technology Foundation STW, applied science division of NWO and the Technology Program of the Ministry of Economic Affairs.

Corresponding Author:

Joost de Winter,
Department of BioMechanical Engineering,
Faculty of Mechanical, Maritime and Materials Engineering,
Delft University of Technology, Mekelweg 2,
2628 CD Delft, The Netherlands
E-mail: j.c.f.dewinter [at] tudelft.nl

Appendix 1

Type II error percentage for the *t* test (above diagonal), Type II error percentage for MWW minus Type II error percentage for the *t* test (below diagonal), and Type I error percentages for the *t* test and MWW (two rightmost columns)

		Type II errors														Type I errors	
		Very strongly agree	*Strongly agree*	*Agree peak*	*Agree flat*	*Neutral to agree*	*Neutral peak*	*Neutral flat*	*Very strongly disagree*	*Strongly disagree*	*Disagree flat*	*Neutral to disagree*	*Certainly not disagree*	*Multimodal*	*Strong multimodal*		
	$n = m = 10$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	*t* test	MWW
1.	*Very strongly agree*		75	12	12	3	1	4	0	0	0	2	5	34	21	0.8	0.9
2.	*Strongly agree*	-6		54	49	25	16	28	0	0	2	20	36	75	48	4.4	5.1
3.	*Agree peak*	-5	-5		93	83	80	83	1	1	36	78	93	94	86	5.1	5.1
4.	*Agree flat*	-3	1	0		91	90	91	3	5	56	88	94	92	91	5.3	5.4
5.	*Neutral to agree*	-1	-1	0	0		95	95	6	9	72	94	88	84	94	5.0	4.9
6.	*Neutral peak*	0	-2	-4	-2	0		94	2	3	61	94	86	81	94	5.1	5.1
7.	*Neutral flat*	-1	2	1	0	0	-1		9	13	77	95	88	83	94	5.1	5.1
8.	*Very strongly disagree*	0	0	0	0	-2	-1	-2		95	35	8	0	3	36	2.2	3.3
9.	*Strongly disagree*	0	0	0	0	-1	-1	0	-2		46	11	1	4	43	4.1	5.3
10.	*Disagree flat*	0	0	0	0	0	-2	0	-8	-4		78	34	45	87	5.1	5.2
11.	*Neutral to disagree*	0	1	0	0	0	0	0	-3	-2	0		83	79	94	5.3	5.4
12.	*Certainly not disagree*	-2	-5	-2	-1	1	-1	0	0	0	1	0		91	89	5.1	5.3
13.	*Multimodal*	-6	3	-2	-1	-3	-6	-1	1	2	1	-3	-4		85	4.9	4.9
14.	*Strong multimodal*	6	21	4	0	-2	-3	-2	11	16	2	-3	1	4		5.0	5.3
	$n = m = 30$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	*t* test	MWW
1.	*Very strongly agree*		35	0	0	0	0	0	0	0	0	0	0	1	0	3.9	5.2
2.	*Strongly agree*	-13		10	6	0	0	1	0	0	0	0	3	29	4	4.9	4.9
3.	*Agree peak*	0	-4		90	61	57	59	0	0	2	48	92	94	65	5.1	5.1
4.	*Agree flat*	0	1	2		85	85	82	0	0	10	76	94	89	82	5.1	5.1
5.	*Neutral to agree*	0	0	-2	-2		95	95	0	0	33	94	70	63	94	4.9	4.8
6.	*Neutral peak*	0	0	-14	-7	0		94	0	0	16	91	67	63	92	5.0	5.0
7.	*Neutral flat*	0	0	3	0	0			0	0	43	95	68	62	95	5.3	5.4
8.	*Very strongly disagree*	0	0	0	0	0	0	0		91	2	0	0	0	1	4.5	5.1
9.	*Strongly disagree*	0	0	0	0	0	0	0	-4		6	0	0	0	2	4.8	5.0
10.	*Disagree flat*	0	0	0	1	0	-1	2	-1	-2		45	1	4	70	5.4	5.4
11.	*Neutral to disagree*	0	0	-2	-2	0	0	0	0	0	-1		56	53	95	4.9	5.0
12.	*Certainly not disagree*	0	-2	-7	0	5	-4	6	0	0	1	2		90	73	5.1	5.0
13.	*Multimodal*	0	13	-6	-5	-11	-19	-5	0	0	0	-8	-16		65	4.9	4.9
14.	*Strong multimodal*	1	26	21	6	-2	-4	-2	4	13	14	-3	13	16		5.3	5.4

Appendix 1 (Continued)

Type II error percentage for the *t* test (above diagonal), Type II error percentage for MWW minus Type II error percentage for the *t* test (below diagonal), and Type I error percentages for the *t* test and MWW (two rightmost columns)

		Type II errors														Type I errors	
		Very strongly agree	*Strongly agree*	*Agree peak*	*Agree flat*	*Neutral to agree*	*Neutral peak*	*Neutral flat*	*Very strongly disagree*	*Strongly disagree*	*Disagree flat*	*Neutral to disagree*	*Certainly not disagree*	*Multimodal*	*Strong multimodal*		
	$n = m = 200$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	*t* test	MWW
1.	*Very strongly agree*		0	0	0	0	0	0	0	0	0	0	0	0	0	5.1	5.0
2.	*Strongly agree*	0		0	0	0	0	0	0	0	0	0	0	0	0	4.6	4.8
3.	*Agree peak*	0	0		60	1	1	1	0	0	0	0	79	93	1	5.1	5.2
4.	*Agree flat*	0	0	16		32	38	24	0	0	0	11	87	53	22	5.1	5.1
5.	*Neutral to agree*	0	0	0	-7		92	93	0	0	0	87	3	2	85	5.3	5.4
6.	*Neutral peak*	0	0	-1	-22	2		86	0	0	0	71	2	2	74	5.1	5.1
7.	*Neutral flat*	0	0	1	0	0	4		0	0	0	93	2	1	91	5.0	5.0
8.	*Very strongly disagree*	0	0	0	0	0	0	0		67	0	0	0	0	0	5.1	5.0
9.	*Strongly disagree*	0	0	0	0	0	0	0	-26		0	0	0	0	0	5.1	5.1
10.	*Disagree flat*	0	0	0	0	0	0	0	0	0		0	0	0	3	5.0	5.0
11.	*Neutral to disagree*	0	0	0	-3	-2	2	-1	0	0	0		0	0	94	4.8	4.9
12.	*Certainly not disagree*	0	0	-42	6	6	-1	6	0	0	0	0		71	5	4.8	5.1
13.	*Multimodal*	0	0	-36	-23	-1	-2	-1	0	0	0	0	-62		2	5.3	5.3
14.	*Strong multimodal*	0	0	54	40	4	6	-1	0	0	40	-4	57	22		4.9	4.8
	$n = 5, m = 20$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	*t* test	MWW
1.	*Very strongly agree*		79	18	19	6	2	8	0	0	0	4	7	48	32	7.4	7.7
2.	*Strongly agree*	-7		61	57	32	22	36	0	0	5	27	42	79	57	4.4	4.5
3.	*Agree peak*	-7	-5		94	86	82	85	1	2	42	82	93	94	85	4.7	4.9
4.	*Agree flat*	-5	1	0		92	90	92	5	8	62	89	92	93	90	5.0	5.3
5.	*Neutral to agree*	-1	1	0	0		94	95	9	13	76	95	88	85	92	4.7	4.7
6.	*Neutral peak*	0	0	-3	0	0		93	3	5	68	93	88	82	87	5.1	5.1
7.	*Neutral flat*	-1	4	1	0	0	1		14	20	81	95	86	85	93	4.6	4.8
8.	*Very strongly disagree*	0	0	1	0	-1	0	-1		94	43	12	1	5	44	4.8	4.4
9.	*Strongly disagree*	0	0	1	2	0	-1	1	0		55	17	1	7	51	4.5	4.1
10.	*Disagree flat*	0	1	3	2	1	0	1	-8	-4		81	44	51	86	4.6	4.8
11.	*Neutral to disagree*	0	2	0	0	0	1	0	-2	-1	0		84	82	92	4.7	4.7
12.	*Certainly not disagree*	-1	-2	-1	1	2	0	2	1	1	1	1		89	86	4.9	4.5
13.	*Multimodal*	-11	3	-1	-1	-2	-3	-1	2	4	3	-1	-1		87	4.6	5.0
14.	*Strong multimodal*	6	20	4	1	-1	-2	0	11	17	2	-1	1	3		5.1	5.0

Appendix 1 (Continued)

Type II error percentage for the *t* test (above diagonal), Type II error percentage for MWW minus Type II error percentage for the *t* test (below diagonal), and Type I error percentages for the *t* test and MWW (two rightmost columns)

		Type II errors														Type I errors	
		Very strongly agree	*Strongly agree*	*Agree peak*	*Agree flat*	*Neutral to agree*	*Neutral peak*	*Neutral flat*	*Very strongly disagree*	*Strongly disagree*	*Disagree flat*	*Neutral to disagree*	*Certainly not disagree*	*Multimodal*	*Strong multimodal*		
	$n = 100$, $m = 10$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	*t* test	MWW
1.	*Very strongly agree*		57	1	1	0	0	0	0	0	0	0	0	7	1	4.9	4.2
2.	*Strongly agree*	-15		28	20	4	1	5	0	0	0	2	10	53	16	4.4	4.9
3.	*Agree peak*	-1	-6		92	73	70	72	0	0	9	64	92	93	75	5.0	5.2
4.	*Agree flat*	0	2	1		88	87	87	0	0	28	84	91	91	84	5.1	5.1
5.	*Neutral to agree*	0	0	-1	-1		93	94	0	0	51	94	79	75	89	4.9	4.8
6.	*Neutral peak*	0	0	-10	-2	1		91	0	0	36	90	77	73	85	4.9	4.8
7.	*Neutral flat*	0	2	3	0	0	2		0	1	61	95	77	73	92	4.9	4.8
8.	*Very strongly disagree*	0	0	0	0	0	0	0		94	10	0	0	0	6	4.6	4.1
9.	*Strongly disagree*	0	0	0	0	0	0	0	-4		18	1	0	0	11	4.4	5.0
10.	*Disagree flat*	0	0	2	3	1	-1	3	-5	-4		62	8	16	78	4.9	5.0
11.	*Neutral to disagree*	0	1	-1	-1	0	1	0	0	0	1		70	68	91	5.6	5.4
12.	*Certainly not disagree*	0	-3	-4	2	5	-1	6	0	0	2	3		87	76	5.0	5.0
13.	*Multimodal*	-2	9	-2	-3	-6	-13	-3	0	0	2	-6	-5		76	4.6	4.6
14.	*Strong multimodal*	3	34	13	5	1	3	0	11	21	10	0	12	11		4.9	4.8

Note. A positive number below the diagonal means that the *t* test was more powerful than MWW.

Appendix 2

Computer simulation code (Matlab 7.7.0.471, R2008b).

```
clc;clear all;close all

D= [.00 .01 .03 .06 .90   % 1. Very strongly agree
    .01 .03 .06 .30 .60   % 2. Strongly agree
    .05 .10 .20 .45 .20   % 3. Agree peak
    .10 .15 .20 .30 .25   % 4. Agree flat
    .10 .20 .30 .25 .15   % 5. Neutral to agree
    .00 .20 .50 .20 .10   % 6. Neutral peak
    .15 .20 .25 .20 .20   % 7. Neutral flat
    .80 .12 .04 .03 .01   % 8. Very strongly disagree
    .70 .20 .06 .03 .01   % 9. Strongly disagree
    .25 .35 .20 .15 .05   % 10. Disagree flat
    .10 .25 .30 .20 .15   % 11. Neutral to disagree
    .01 .04 .50 .30 .15   % 12. Certainly not disagree
    .15 .05 .15 .25 .40   % 13. Multimodal
    .45 .05 .00 .05 .45]; % 14. Strong multimodal

nm=[10 10
    30 30
    200 200
    5 20
    100 10];

reps=10000;
Dcum=cumsum(D,2);

for i1=1:size(nm,1)   % loop across the different sample sizes
    pT=NaN(size(Dcum,2),size(Dcum,2),reps);
    pW=NaN(size(Dcum,2),size(Dcum,2),reps);
    n = nm(i1,1); m = nm(i1,2);
    disp([m n])
    for i2=1:size(Dcum,1);  % loop across the 14 distributions
        for i3=1:size(Dcum,1) % loop across the 14 distributions
            for i4=1:reps;
                L1=NaN(n,1);   L2=NaN(m,1);
                X1=rand(n,1);  X2=rand(m,1);  % draw random variables
                                              % between 0 and 1
               for i5=1:n   % generate Likert item data for sample 1
                    L1(i5)= find(X1(i5)<Dcum(i2,:),1);
                end
                for i5=1:m  % generate Likert item data for sample 2
                    L2(i5)= find(X2(i5)<Dcum(i3,:),1);
                end
                [h,pT(i2,i3,i4)]=ttest2(L1,L2);  % conduct t test
                V=tiedrank([L1;L2]);L1=V(1:n);L2=V(n+1:end); % transform to ranks
                [h,pW(i2,i3,i4)]=ttest2(L1,L2);  % conduct t test on
                                                 % rank-transformed data
            end
        end
    end

    Tsuccess=squeeze(sum(pT<.05,3));   % determine if null hypothesis is rejected
    Wsuccess=squeeze(sum(pW<.05,3));
```

```
    % Generating the layout of Table 2 & Appendix 1
    temp1=1-(triu(Tsuccess)+tril(Tsuccess)')./(2*reps); % Type II errors for t test.
                                          % average of above and below diagonal
    temp1=temp1-tril(temp1);
    temp2=(tril(Tsuccess-Wsuccess)+triu(Tsuccess-Wsuccess)')./(2*reps);
                % Difference between type II errors of MWW minus t test
                % average of above and below diagonal
    temp2=temp2-diag(diag(temp2));
    temp3=temp1+temp2;
    temp3=temp3+diag(NaN(size(Dcum,1),1));
    Type2=round(100*temp3);
    Type1=round(1000*[diag(Tsuccess)/reps diag(Wsuccess)/reps])/10;  % Type I errors
    disp([Type2 Type1])  % Display output
end
```

Addendum to
De Winter, J. C. F., & Dodou, D. (2010). Five-Point Likert Items: *t* test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation, 15*, 11.

J. C. F. de Winter and D. Dodou
Department of BioMechanical Engineering, Delft University of Technology

In the published paper, the Type II error rates for unequal sample sizes (see final two matrices of Appendix 1) were averaged for $n = m$ and $m = n$. Below, the complete matrices are reported (based 10,000 repetitions). It can be seen that the matrices are not symmetric. For example, when comparing Very strongly agree (n = 5) with Multimodal (m = 20), MWW has a power advantage of 23%. Oppositely, when comparing Multimodal (n = 5) with Very strongly agree (m = 20), the *t* test has a power advantage of 2%.

The text of the published paper remains valid. However, it should be mentioned that when sample sizes are unequal, power differences between the *t* test and MWW are more complex than when sample sizes are equal.

		Type II error percentage for the *t* test														Type I error percentage	
		Very strongly agree	*Strongly agree*	*Agree peak*	*Agree flat*	*Neutral to agree*	*Neutral peak*	*Neutral flat*	*Very strongly disagree*	*Strongly disagree*	*Disagree flat*	*Neutral to disagree*	*Certainly not disagree*	*Multimodal*	*Strong multimodal*		
	n = 5, m = 20	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	*t* test	MWW
1.	*Very strongly agree*		96	23	28	8	3	12	0	0	0	6	9	72	48	7.1	7.2
2.	*Strongly agree*	62		69	69	37	22	43	0	0	5	32	40	93	75	4.4	4.4
3.	*Agree peak*	12	57		97	88	79	90	1	2	44	85	89	98	96	5.0	5.2
4.	*Agree flat*	11	47	90		91	83	93	4	6	59	89	87	94	97	5.0	5.2
5.	*Neutral to agree*	3	27	83	93		89	97	7	11	76	95	81	89	99	5.2	5.3
6.	*Neutral peak*	2	21	87	97	98		99	2	4	75	98	86	93	99	5.1	5.3
7.	*Neutral flat*	5	28	80	91	93	86		11	17	77	93	77	87	98	5.1	5.0
8.	*Very strongly disagree*	0	0	2	7	11	3	17		98	49	15	1	6	62	4.8	4.2
9.	*Strongly disagree*	0	0	3	9	16	5	24	90		62	20	1	8	69	4.6	4.3
10.	*Disagree flat*	0	4	41	64	77	59	84	37	47		82	35	56	97	4.6	4.7
11.	*Neutral to disagree*	3	23	79	91	94	88	97	9	13	80		76	86	99	5.0	5.1
12.	*Certainly not disagree*	5	44	96	99	95	88	96	0	1	49	92		98	98	5.4	5.2
13.	*Multimodal*	25	65	89	92	81	73	84	4	5	46	78	80		93	4.9	5.1
14.	*Strong multimodal*	16	39	75	82	85	76	88	28	34	77	86	74	81		5.2	5.1

Type II error percentage for MWW minus Type II error percentage for the *t* test

		Very strongly agree	*Strongly agree*	*Agree peak*	*Agree flat*	*Neutral to agree*	*Neutral peak*	*Neutral flat*	*Very strongly disagree*	*Strongly disagree*	*Disagree flat*	*Neutral to disagree*	*Certainly not disagree*	*Multimodal*	*Strong multimodal*
	$n = 5, m = 20$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1.	*Very strongly agree*		-9	-9	-8	-1	1	-3	0	0	0	-1	0	-23	-2
2.	*Strongly agree*	-6		-9	-7	-3	1	-3	0	0	1	-1	0	-6	7
3.	*Agree peak*	-5	0		-1	-2	-1	-1	0	1	1	-2	0	-1	0
4.	*Agree flat*	-1	7	1		-1	2	-1	1	3	4	0	3	-1	0
5.	*Neutral to agree*	0	5	1	0		1	0	-2	0	1	0	4	-3	-1
6.	*Neutral peak*	0	-1	-5	-2	0		0	-2	-3	-3	-1	0	-7	-1
7.	*Neutral flat*	1	10	3	0	0	2		0	3	3	0	5	-1	-1
8.	*Very strongly disagree*	0	0	1	1	0	1	-2		0	-9	-2	2	1	1
9.	*Strongly disagree*	0	0	2	1	0	1	-2	-1		-8	-2	2	1	4
10.	*Disagree flat*	0	2	4	1	0	4	-1	-7	0		-1	7	-2	0
11.	*Neutral to disagree*	0	6	1	-1	0	1	0	-2	0	2		4	-3	-1
12.	*Certainly not disagree*	-2	-5	-3	-1	-1	1	-1	0	0	-5	-1		-4	-1
13.	*Multimodal*	2	12	0	-1	-1	1	-1	4	8	7	0	3		2
14.	*Strong multimodal*	14	33	6	2	-1	-2	0	20	29	3	-1	2	5	

		Type II error percentage for the *t* test														Type I error percentage	
		Very strongly agree	*Strongly agree*	*Agree peak*	*Agree flat*	*Neutral to agree*	*Neutral peak*	*Neutral flat*	*Very strongly disagree*	*Strongly disagree*	*Disagree flat*	*Neutral to disagree*	*Certainly not disagree*	*Multimodal*	*Strong multimodal*		
	$n = 100$, $m = 10$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	*t* test	MWW
1.	*Very strongly agree*		44	0	0	0	0	0	0	0	0	0	0	3	1	4.4	4.2
2.	*Strongly agree*	71		25	15	3	1	4	0	0	0	2	10	36	9	4.5	4.6
3.	*Agree peak*	1	31		87	69	78	65	0	0	9	60	97	87	58	5.1	5.2
4.	*Agree flat*	1	25	96		91	97	86	0	0	30	86	99	89	74	4.8	4.9
5.	*Neutral to agree*	0	5	76	86		98	93	0	0	52	94	90	68	80	5.1	5.2
6.	*Neutral peak*	0	2	63	77	87		82	0	0	29	84	79	56	69	4.9	4.9
7.	*Neutral flat*	0	6	80	88	97	99		0	1	66	97	92	71	84	5.0	4.9
8.	*Very strongly disagree*	0	0	0	0	0	0	0		91	8	0	0	0	4	4.6	3.9
9.	*Strongly disagree*	0	0	0	0	0	0	1	97		15	0	0	0	6	4.3	4.9
10.	*Disagree flat*	0	0	9	26	51	44	57	13	23		60	9	14	62	4.5	4.7
11.	*Neutral to disagree*	0	3	69	81	94	98	92	0	1	63		84	60	81	4.8	4.8
12.	*Certainly not disagree*	0	11	87	82	67	75	62	0	0	7	57		76	54	5.3	5.1
13.	*Multimodal*	11	67	98	93	82	89	75	0	0	18	74	99		66	4.8	4.9
14.	*Strong multimodal*	1	23	93	96	99	100	99	9	15	94	100	98	86		4.7	4.7

Type II error percentage for MWW minus Type II error percentage for the *t* test

		Very strongly agree	*Strongly agree*	*Agree peak*	*Agree flat*	*Neutral to agree*	*Neutral peak*	*Neutral flat*	*Very strongly disagree*	*Strongly disagree*	*Disagree flat*	*Neutral to disagree*	*Certainly not disagree*	*Multimodal*	*Strong multimodal*
	$n = 100$, $m = 10$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1.	*Very strongly agree*		-13	0	0	0	0	0	0	0	0	0	0	0	4
2.	*Strongly agree*	-16		-4	7	2	0	5	0	0	0	2	-3	22	40
3.	*Agree peak*	-1	-8		2	1	-16	6	0	0	3	1	-6	-2	21
4.	*Agree flat*	0	-3	0		-1	-5	1	0	0	0	-2	-1	-2	7
5.	*Neutral to agree*	0	-1	-2	0		0	0	0	0	0	0	0	-3	2
6.	*Neutral peak*	0	0	-5	1	2		4	0	0	8	3	-1	0	7
7.	*Neutral flat*	0	0	-1	0	0	0		0	0	-1	0	-1	-1	1
8.	*Very strongly disagree*	0	0	0	0	0	0	0		-4	-5	0	0	0	15
9.	*Strongly disagree*	0	0	0	0	0	0	1	-4		-3	0	0	0	28
10.	*Disagree flat*	0	0	1	5	1	-11	5	-5	-5		2	-4	8	16
11.	*Neutral to disagree*	0	0	-3	0	0	0	0	0	0	0		-3	-1	1
12.	*Certainly not disagree*	0	-2	-1	6	10	-2	13	0	0	8	9		-1	23
13.	*Multimodal*	-5	-2	-3	-3	-10	-24	-4	0	0	-4	-11	-10		15
14.	*Strong multimodal*	1	26	5	2	0	0	0	6	15	3	0	1	6	

Note. A positive number means that the *t* test was more powerful than MWW.