Comparing two samples from an individual Likert question.

B. Derrick and P. White

Faculty of Environment and Technology, University of the West of England, Bristol, BS16 1QY (UK)

Email: ben.derrick@uwe.ac.uk; paul.white@uwe.ac.uk

ABSTRACT

For two independent samples there is much debate in the literature whether parametric or non-parametric methods should be used for the comparison of Likert question responses. The comparison of paired responses has received less attention in the literature. In this paper, parametric and non-parametric tests are assessed in the comparison of two samples from a paired design on a five point Likert question. The tests considered are the independent samples t-test, the Mann-Whitney test, the paired samples t-test and the Wilcoxon test. Pratt's modified Wilcoxon test for dealing with zero differences is also included. The Type I error rate and power of the test statistics are assessed using Monte-Carlo methods. The parameters varied are; sample size, correlation between paired observations, and the distribution of the responses. The results show that the independent samples t-test and the Mann-Whitney test are not Type I error robust when there is correlation between the two groups compared. Pratt's test more closely maintains the Type I error rate than the standard Wilcoxon test does. The paired samples t-test is Type I error robust across the simulation design. As the correlation between the paired samples increases, the power of the test statistics making use of the paired information increases. The paired samples t-test is more powerful than Pratt's test when the correlation is weak. The power differential between the test statistics is exacerbated when sample sizes are small. Assuming equally spaced categories on a five point Likert item, the paired samples t-test is not inappropriate.

Keywords: Likert item; Likert scale; Wilcoxon test; Pratt's test; Paired samples t-test

Mathematics Subject Classification: 60 62

1. INTRODUCTION

A Likert item is a forced choice ordinal question which captures the intensity of opinion or degree of assessment in survey respondents. Historically a Likert item comprises five points worded: Strongly approve, Approve, Undecided, Disapprove, Strongly Disapprove (Likert, 1932). Other alternative wording, such as "agree" or "neutral" or "neither agree nor disagree" may be used depending on the context.

The literature is sometimes confused between the comparison of samples using summed Likert scales and the comparison of samples for individual Likert items (Boone and Boone, 2012). A summed Likert scale is formed by the summation of multiple Likert items that measure similar information. This summation process necessarily requires the assignment of scores to the Likert ordinal category labels. The summation of multiple Likert items to produce Likert scales has not been without controversy but it is a well-established practice in scale construction, and is one which may

produce psychometrically robust scales with interval-like properties. Such derived scales, could potentially yield data amenable to analysis using parametric techniques (Carifo and Perla, 2007). Distinct from Likert scales, the comparison of two samples on an individual Likert question is the subject of this paper.

The response categories of a five point Likert item may be coded 1 to 5 and the item responses viewed as being ordinal under Stevens (1946) classification scheme. Extant literature acknowledges that in certain practical and methodological aspects, the Likert-item responses may approximate interval level data (Norman, 2010). The ordinal codes 1, 2, 3, 4, and 5 or alternatively -2, -1, 0, 1, 2 could be used as numerical scores in robust tests for differences. This change from codes to numeric scores is used in the creation of summated Likert scales and is at the heart of the controversy. Proponents in favour of such practice advance an argument that the Likert question is accessing some information from an underlying scale and the resultant score is a non-linear realisation from this scale (Norman, 2010). Thus, although the scored item may not perfectly have the required properties to be classed as interval level data under Stevens classification scheme, the scored item might, in practice, approximate interval level data and be amenable to analysis using parametric techniques.

When comparing two independent sets of responses from a Likert question, the independent samples t-test is frequently performed. The corresponding non-parametric test for independent samples is the Mann-Whitney-Wilcoxon test (Wilcoxon, 1945). This test may also be referred to as the Wilcoxon-Mann-Whitney test, or as is the case in this paper, simply referred to as the Mann-Whitney test.

For two independent samples, whether the correct approach for analysis should be a parametric t-test or the non-parametric Mann-Whitney test is much debated in the literature (Sullivan and Artino, 2013). The choice between parametric and non-parametric tests for the analysis of single Likert items depends on the assumptions that researchers are willing to make and the hypotheses that they are testing (Jamieson, 2004). Some practitioners are uncomfortable with a comparison of means using a parametric test, arguing that response categories cannot be justifiably assumed to be equally spaced and consequently the use of equally spaced scores is unwarranted. In contrast, Allen and Seaman (2007) suggests that Likert items measure an underlying continuous measure and suggests the use of the independent samples t-test as a pilot test, prior to obtaining a continuous measure. If the assumption that the underlying distribution is continuous can be deemed reasonable, Likert responses approximate interval data. For interval data, the use of parametric tests may not be inappropriate. When the assumption of interval data applies, consideration should be given to the sample size and distribution of the responses before applying the independent samples t-test (Jamieson, 2004).

If sample sizes are large, both parametric and non-parametric test statistics are likely to have adequate power. However, in research there is a trade-off between increasing sample size and reducing collection costs. When resource is scarce, the most powerful test statistic for small samples is of interest.

For two independent samples, De Winter and Dodou (2010) found that both the independent samples t-test and the Mann-Whitney test are generally Type I error robust at the 5% significance level for a five point Likert item. This is true across a diverse range of distributions and sample sizes. Both tests suffer some exceptions to Type I error robustness when the distributions have extreme kurtosis and skew. The power is similar between the two tests, for both equal and unequal sample sizes. When the distribution is multimodal with responses split mainly between strongly approve and strongly disapprove, the independent samples t-test is more powerful than the Mann-Whitney test. Rasch, Teuscher and Guiard (2007) show that using the Mann-Whitney test using the Normal approximation with correction for ties is Type I error robust for two groups of independent observations on a five point Likert item.

For two independent samples, Nanna and Sawilowski (1998) found that the independent samples t-test and the Mann-Whitney test are Type I error robust for seven point Likert item responses, with the Mann-Whitney test superior in power. This is likely observed because there is more scope to apply greater skew on a higher point Likert-style scale.

The literature is much quieter on the analysis of Likert items in paired samples designs. A non-parametric test for paired samples is the Wilcoxon rank sum test (Wilcoxon, 1945). This is often referred to as the Wilcoxon signed rank test, or as is the case in this paper, simply referred to as the Wilcoxon test. When the samples are from an underlying Normal distribution, the null hypothesis is of equal distributions, but this is particularly sensitive to changes in location (Hollander, Wolfe and Chicken, 2013). Thus if samples are from a bivariate Normal distribution, assessing for a location shift is reasonable.

When comparing two groups of paired samples on a five point Likert item, the paired samples t-test is often used in preference to the Wilcoxon test (Clason and Dormody, 1994). This choice of test is not inappropriate when interval approximating data is assumed, and when the null hypothesis is one of no difference in central location (Sisson and Stocker, 1989).

The degree of correlation between two samples is likely to impact the choice of test. The correlation between two sets of responses on a Likert scale is typically hard to quantify. With respect to bivariate Normal distributions, Fradette et.al. (2003) suggest that if the correlation is small then the independent samples t-test could be used. However, under the same conditions, Zimmerman (1997) argues that using the independent samples t-test for even a small a degree of correlation violates the independence assumption and can distort the Type I error rate. For bivariate normality, Vonesh (1983) demonstrates that the paired samples t-test is more powerful than the independent samples test when $\rho \geq 0.25$.

In general, the Wilcoxon test with a correction for ties, may be used to test for a location shift between two discrete groups. The Wilcoxon test discards observations where there is a zero difference between the two groups. Given the discrete nature of Likert item data, it would not be unusual to observe a large proportion of zero differences in a sample. The discarding of many data pairs with a

zero difference may be problematic. Pratt (1959) proposed a modification of the Wilcoxon test to overcome potential problems caused by discarding zero differences. In Pratt's test, the absolute paired differences are ordered including the zero differences, ranks are applied to the non-zero differences as if the zero differences had received ranks, and these ranks used in the Wilcoxon test. Conover (1973) compared the Wilcoxon test dropping zero differences to Pratt's test incorporating zero differences and concluded that the relative performance of the two approaches depends on the underlying distribution. The comparison conducted by Conover (1973) did not include Likert items and did not extend to the inclusion of the paired samples t-test.

A further alternative method for handling zero differences suggested by Pratt (1959) is to randomly allocate zero differences to either positive or negative ranks. To achieve this for every zero difference add a random uniform deviate $\varepsilon \sim U(-0.1, 0.1)$ and then proceed with the ranking. This approach is referred to as the random epsilon method in the following.

For paired five point Likert data we seek to compare the relative behaviour of the Wilcoxon test, Pratt's test, the random epsilon method and the paired samples t-test. The comparison is undertaken by discretising realisations from bivariate Normal distributions on to a five point scale over a range of correlation coefficients, ρ, including $\rho = 0$. For this latter reason we additionally include the Mann-Whitney test and the independent samples t-test in the comparison. Mindful that differences in location are likely to be accompanied with differences in variances, we additionally include the separate variances t-test i.e. Welch's test in the comparison. It is known that for independent samples, Welch's test is Type I error robust under normality for both equal and unequal variances (Derrick, Toher and White, 2016).

Below we give the simulation study, key results and a discussion of the findings.

2. METHODOLOGY

Random Normal deviates for two groups of sample size n are generated using the Box–Muller (1958) transformation. These deviates are transformed into n pairs with Pearson's correlation coefficient ρ using methodology outlined by Kenney and Keeping (1951).

For each combination of n and ρ, correlated bivariate Normal deviates x_{ij} are generated, where *i* = {1:*n*} and *j* = {Group 1, Group 2}. The mean of the sample is varied by adding μ_j to each deviate so that x_{ij} ~N(μ_j,1). The values of each of the parameters simulated are given in Table 1.

Table 1. Summary of the simulation design.

Sample size, n	10, 20, 30, 50
Correlation coefficient, ρ	0.00, 0.25, 0.50, 0.75
Scenarios	$\begin{array}{llllll} & \mu_1 & \mu_2 & \eta_1 & \eta_2 & \\ A) & 0 & 0 & 0 & 0 & \\ B) & 0.5244 & 0.5244 & 1 & 1 & H_0 \\ C) & 1.2816 & 1.2816 & 2 & 2 & \end{array}$ $\begin{array}{llllll} & \mu_1 & \mu_2 & \eta_1 & \eta_2 & \\ D) & 0 & 0.5244 & 0 & 1 & \\ E) & 0 & 1.2816 & 0 & 2 & \\ F) & 0.5244 & 1.2816 & 1 & 2 & H_1 \\ G) & -0.5244 & 0.5244 & -1 & 1 & \\ H) & -0.5244 & 1.2816 & -1 & 2 & \\ I) & -1.2816 & 1.2816 & -2 & 2 & \end{array}$
Test Statistics	T_1 Paired samples t-test T_2 Independent samples t-test T_3 Welch's t-test W_1 Wilcoxon test (Traditional method, discarding zeroes) W_2 Pratt's test (Wilcoxon test, Pratt's zeroes modification) W_3 Random ε (Wilcoxon test, $\varepsilon \sim U(-0.1,0.1)$ added to zeroes) MW Mann-Whitney test.
Number of iterations	10,000
Nominal significance level	5% (two-sided test)
Programming language	R version 3.1.3
	Complete tables of all results available on request.

Without loss of generality the five points on the Likert scale are numbered from -2 to 2, the "neutral" response is 0. The Likert-style responses y_{ij} are calculated using the cut-points as follows:

$$y_{ij} = \begin{cases} 2 & \text{if} \quad x_{ij} > 0.8416 \\ 1 & \text{if} \quad 0.2533 \le x_{ij} \le 0.8416 \\ 0 & \text{if} \quad -0.2533 \le x_{ij} \le 0.2533 \\ -1 & \text{if} \quad -0.8416 \le x_{ij} \le -0.2533 \\ -2 & \text{if} \quad x_{ij} < -0.8416 \end{cases}$$

The cut-points are calculated so that under N(0,1) the theoretical distribution of the Likert-style responses is uniform. The median of Group 1 and the median of Group 2 are represented by η_1 and η_2 respectively. Scenarios A) to I) in Table 1 give an example of each of the possible bivariate

pairings of η_1 and η_2 within a five point Likert design. For example, scenario D) $\eta_1 = 0, \eta_2 = 1$, is equivalent to $\eta_1 = 1, \eta_2 = 0$; $\eta_1 = 0, \eta_2 = -1$; and $\eta_1 = -1, \eta_2 = 0$.

For selected parameter combinations within the factorial simulation design, theoretical observed proportions of y_{ij} are illustrated in Figure 1. These showcase the range of distributions in the simulation design.

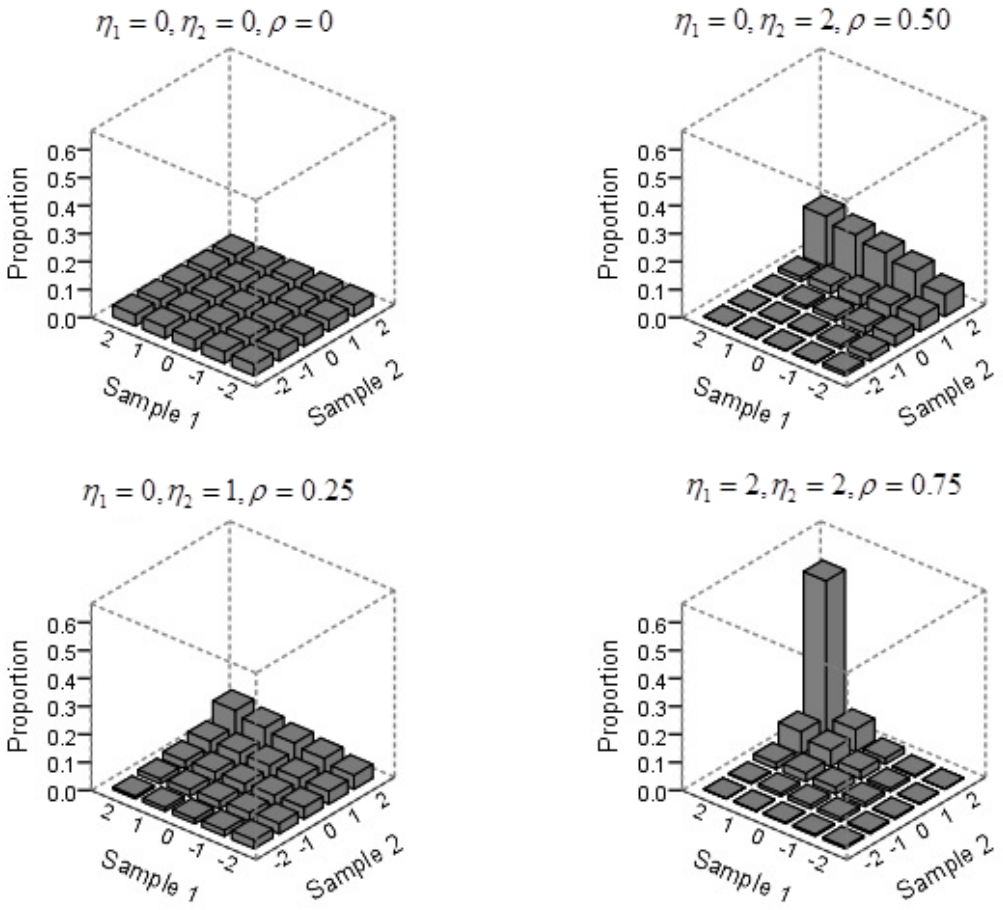

Figure 1. Theoretical distributions of the proportion of observed responses, for selected parameter combinations.

For non-parametric tests, exact p-values are difficult to obtain due to the frequent occurrence of ties for Likert data. When there are ties, the Normal approximation corrected for ties can be used to calculate p-values (Hollander, Wolfe and Chicken, 2013). The Normal approximations for both the Mann-Whitney test and the Wilcoxon test are very accurate even for small sample sizes (Bellera, Julien and Hanley, 2010). The continuity correction factor is often used when approximating discrete distributions using the Normal distribution. The correction factor has little impact when $n \geq 10$ (Emerson and Moses, 1985). The non-parametric tests are performed using the Normal approximation with correction for ties. A continuity correction factor is also applied. Two-sided tests are performed at the nominal 5% significance level.

For each of the parameter combinations within the simulation design the process outlined above is repeated 10,000 times. The proportion of the 10,000 iterations where H_0 is rejected is calculated. For the three scenarios in the simulation design where H_0 is true, the proportion of iterations where H_0 is rejected represents the Type I error rate. For the six scenarios in the simulation design where H_1 is true, the proportion of iterations where H_0 is rejected represents the power of the test.

3. RESULTS

Type I error rates for each of the test statistics are considered. This is followed by a summary of the power of the test statistics.

When the null hypothesis is true, H_0 rejection rates within the interval [0.025 , 0.075] are within Bradley's (1978) liberal limits. This Type I error robustness criteria is often used by researchers, although there is no consensus on the most appropriate criteria (Serlin, 2000). For the three scenarios in the simulation design where H_0 is true, the Type I error rates for each of the test statistics is given in Table 2.

Table 2. Type I error rates for selected combinations. For each parameter combination the test statistics within Bradley's (1978) liberal robustness criteria is highlighted in bold.

ρ	η_1	η_2	n	T_1	T_2	T_3	W_1	W_2	W_3	MW
0	0	0	10	**0.0528**	**0.0510**	**0.0498**	**0.0375**	**0.0487**	**0.0385**	**0.0441**
0	0	0	20	**0.0523**	**0.0513**	**0.0511**	**0.0466**	**0.0484**	**0.0464**	**0.0486**
0	0	0	30	**0.0494**	**0.0508**	**0.0508**	**0.0464**	**0.0501**	**0.0463**	**0.0494**
0	1	1	10	**0.0484**	**0.0498**	**0.0472**	**0.0344**	**0.0466**	**0.0356**	**0.0426**
0	1	1	20	**0.0549**	**0.0527**	**0.0524**	**0.0489**	**0.0506**	**0.0488**	**0.0509**
0	1	1	30	**0.0471**	**0.0486**	**0.0481**	**0.0447**	**0.0473**	**0.0446**	**0.0455**
0	2	2	10	**0.0352**	**0.0461**	**0.0313**	0.0168	**0.0570**	**0.0185**	**0.0441**
0	2	2	20	**0.0450**	**0.0460**	**0.0450**	**0.0350**	**0.0520**	**0.0400**	**0.0480**
0	2	2	30	**0.0410**	**0.0500**	**0.0500**	**0.0400**	**0.0440**	**0.0410**	**0.0490**
0.75	0	0	10	**0.0438**	0.0018	0.0018	0.0243	**0.0546**	**0.0268**	0.0014
0.75	0	0	20	**0.0498**	0.0005	0.0005	**0.0392**	**0.0482**	**0.0405**	0.0007
0.75	0	0	30	**0.0463**	0.0006	0.0006	**0.0432**	**0.0459**	**0.0438**	0.0005
0.75	1	1	10	**0.0381**	0.0014	0.0012	0.0207	**0.0514**	**0.0219**	0.0012
0.75	1	1	20	**0.0514**	0.0006	0.0006	**0.0398**	**0.0485**	**0.0406**	0.0004
0.75	1	1	30	**0.0468**	0.0009	0.0009	**0.0404**	**0.0439**	**0.0410**	0.0008
0.75	2	2	10	0.0221	0.0036	0.0025	0.0077	**0.0402**	0.0103	0.0036
0.75	2	2	20	**0.0460**	0.0050	0.0050	**0.0270**	**0.0470**	**0.0310**	0.0080
0.75	2	2	30	**0.0470**	0.0040	0.0040	**0.0380**	**0.0520**	**0.0400**	0.0050

Table 2 shows that all of the test statistics under consideration fulfil Bradley's Type I error robustness criteria when $\rho = 0$. As the correlation increases, test statistics assuming independent samples (T_2, T_3 and MW) do not maintain Type I error robustness. Test statistics assuming independent samples are valid when the structure of the data is unpaired, but appear biased when the structure of the data is paired.

Test statistics making use of paired information are robust across the range of ρ. Pratt's test (W_2) is Type I error robust for every combination of parameters under the simulation design. T_1, W_1 and W_3 are also generally Type I error robust, with minor deviations when the sample size is small ($n = 10$) and both samples are heavily skewed ($\eta_1 = 2, \eta_2 = 2$).

Figure 2 summarises for each test statistic the Type I error rates for all of the sample size and correlation coefficient combinations within the design. It can be seen from Figure 2 that the paired samples t-test (T_1) and Pratt's test (W_2) perform closest to the nominal Type I error rate of 5% across the simulation design.

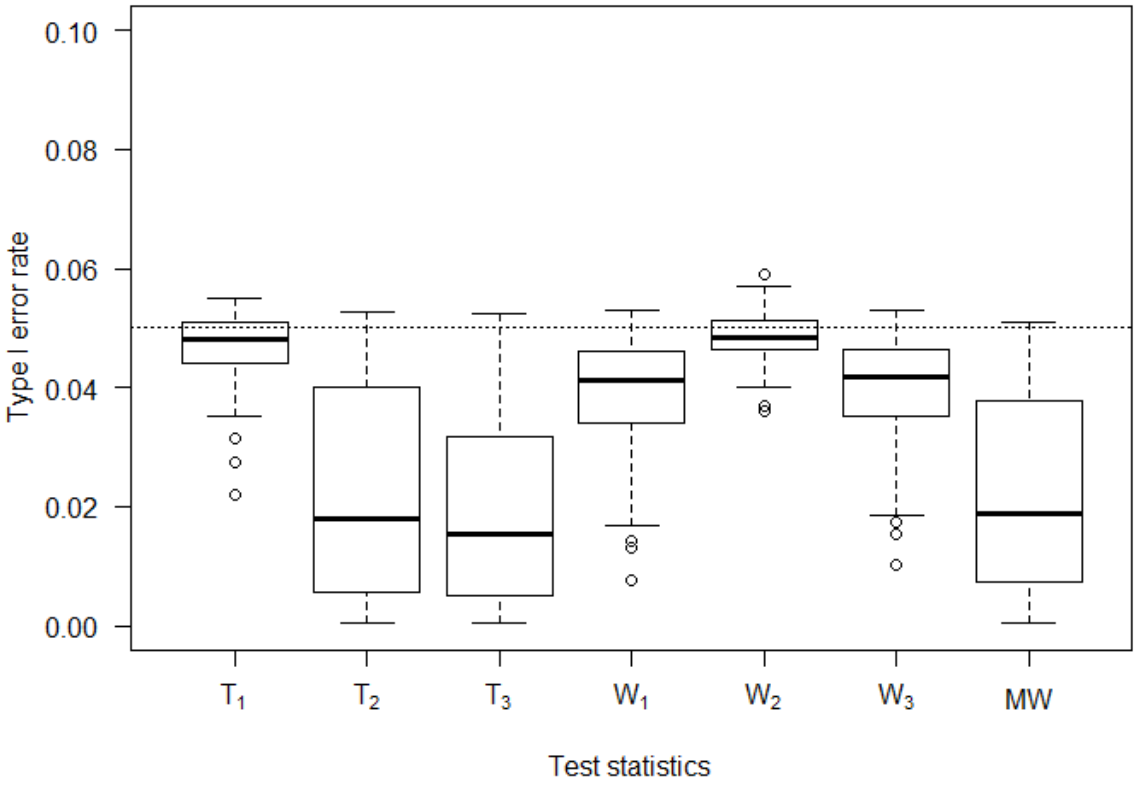

Figure 2. Type I error rates for each test statistic, averaged over each combination of parameters. The dotted line represents significance level of 5%.

Figure 2 demonstrates that each of the test statistics are generally conservative. A conservative test statistic is of less concern than a liberal test statistic (Mehta and Srinivasan, 1970). Alternative Type I error robustness criteria states that Type I error rates within the interval [0 , 0.055] are acceptable (Sullivan and D'Agostino, 1996). For all of the test statistics, each of the parameter combinations within the factorial design are within at least one of the mentioned Type I error robustness criteria. Hence, the power of each of the test statistics can be reasonably compared.

The power for each of the test statistics is given in Table 3, for the six scenarios in the simulation design where H_1 is true.

Table 3. Power for selected conditions. For each parameter combination the most powerful test is highlighted in bold.

ρ	η_1	η_2	n	T_1	T_2	T_3	W_1	W_2	W_3	MW
0	-1	1	10	0.5261	**0.5682**	0.5651	0.4587	0.4943	0.4594	0.5366
0	-1	1	20	0.8496	**0.8640**	0.8637	0.8283	0.8350	0.8292	0.8593
0	-1	1	30	0.9594	**0.9650**	**0.9650**	0.9534	0.9544	0.9535	0.9629
0	0	1	10	0.1742	**0.1905**	0.1873	0.1381	0.1631	0.1386	0.1707
0	0	1	20	0.3200	**0.3300**	0.3292	0.2956	0.3057	0.2965	0.3202
0	0	1	30	0.4522	**0.4631**	0.4627	0.4354	0.4404	0.4354	0.4573
0	0	2	10	0.6502	**0.7044**	0.6970	0.5745	0.6275	0.5791	0.6805
0	0	2	20	0.9415	**0.9497**	0.9494	0.9291	0.9338	0.9292	0.9488
0	0	2	30	0.9913	**0.9935**	0.9934	0.9897	0.9910	0.9900	0.9929
0.75	-1	1	10	0.9299	0.5935	0.5878	0.8741	**0.9353**	0.8808	0.5618
0.75	-1	1	20	**0.9989**	0.9508	0.9508	0.9988	0.9989	0.9988	0.9501
0.75	-1	1	30	**1.0000**	0.9967	0.9967	**1.0000**	**1.0000**	1.0000	0.9970
0.75	0	1	10	0.3974	0.0842	0.0819	0.3025	**0.4186**	0.3112	0.0764
0.75	0	1	20	**0.7496**	0.2342	0.2328	0.7119	0.7381	0.7164	0.2341
0.75	0	1	30	**0.9079**	0.4336	0.4334	0.8976	0.9000	0.8990	0.4292
0.75	0	2	10	0.9585	0.7548	0.7413	0.8986	**0.9706**	0.9114	0.7345
0.75	0	2	20	**0.9998**	0.9890	0.9888	0.9997	**0.9998**	0.9997	0.9896
0.75	0	2	30	**1.0000**	0.9996	0.9996	**1.0000**	**1.0000**	1.0000	0.9997

Table 3 shows that the power difference between the independent samples t-test (T_2) and Welch's test (T_3) is negligible. Additionally, there is little power differential between the traditional Wilcoxon test (W_1) and the Random ε method (W_3).

To summarise the power across the parameters within the simulation design, Figure 3 depicts how the test statistics T_1, T_2, W_1, W_2 and MW perform with increasing ρ for a small sample size of $n = 10$. Figure 4 depicts how the test statistics T_1, T_2, W_1, W_2 and MW perform with increasing ρ for a larger sample size of $n = 20$.

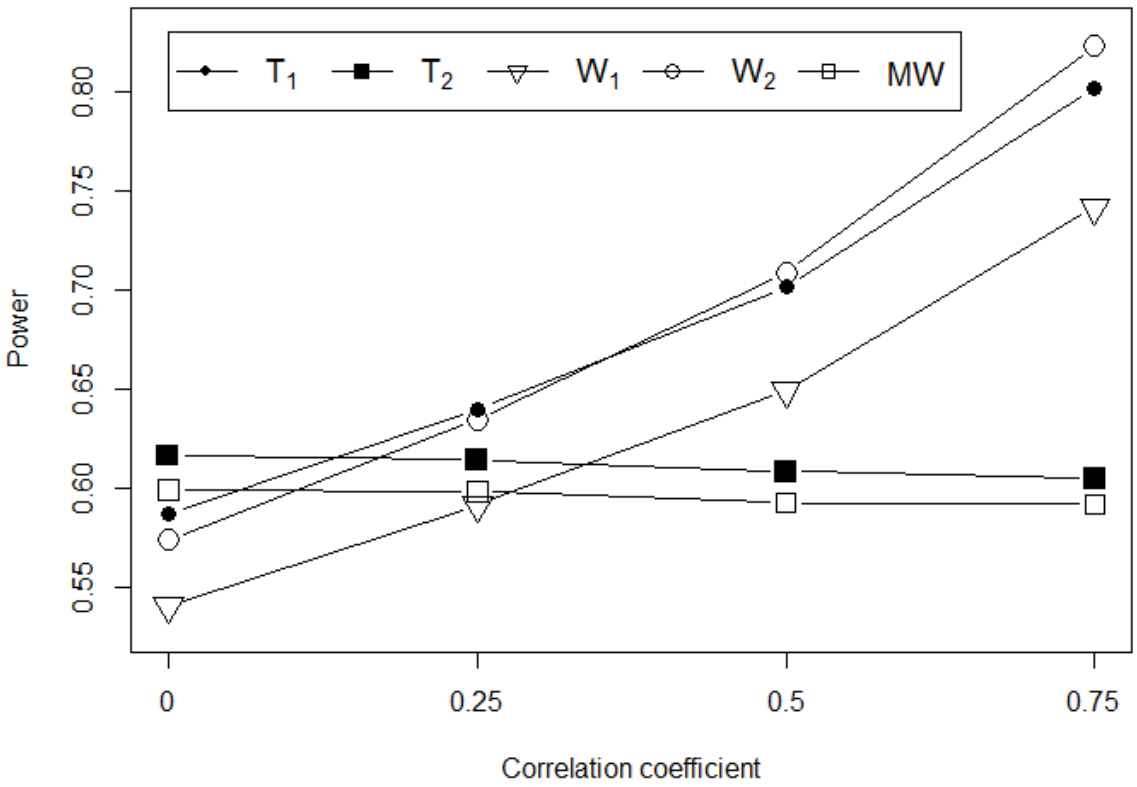

Figure 3. Power of the test statistics T_1, T_2, W_1, W_2 and MW where $n = 10$, averaged across each scenario within the simulation design.

Figure 3 shows that the independent samples t-test consistently out performs the Mann-Whitney test. When $\rho = 0$ the independent samples t-test is the recommended test of choice. When ρ >0 the paired samples t-test is more powerful than the independent samples t-test. These findings are consistent with the paired samples t-test and the independent samples t-test for continuous data (Fradette et. al. 2003; Zimmerman, 1997; Vonesh, 1983).

It can also be seen from Figure 3 that the standard Wilcoxon test consistently lacks power compared to Pratt's test and the paired samples t-test. When $\rho = 0.25$ the paired samples t-test is the most powerful test. As the correlation increases, Pratt's method becomes the test of choice.

As $\rho \to 1$ the power of both T_1 and W_2 increases. Given that both the paired samples t-test and Pratt's test have high power when the correlation is strong, the decision between the two tests is not of any major practical consequence in these circumstances.

Figure 4 shows that as sample size increases, the choice between the Wilcoxon test, Pratt's test and the paired samples t-test becomes less important. The sample size is large enough to compensate for discarded zeroes in the Wilcoxon test for $n \geq 20$.

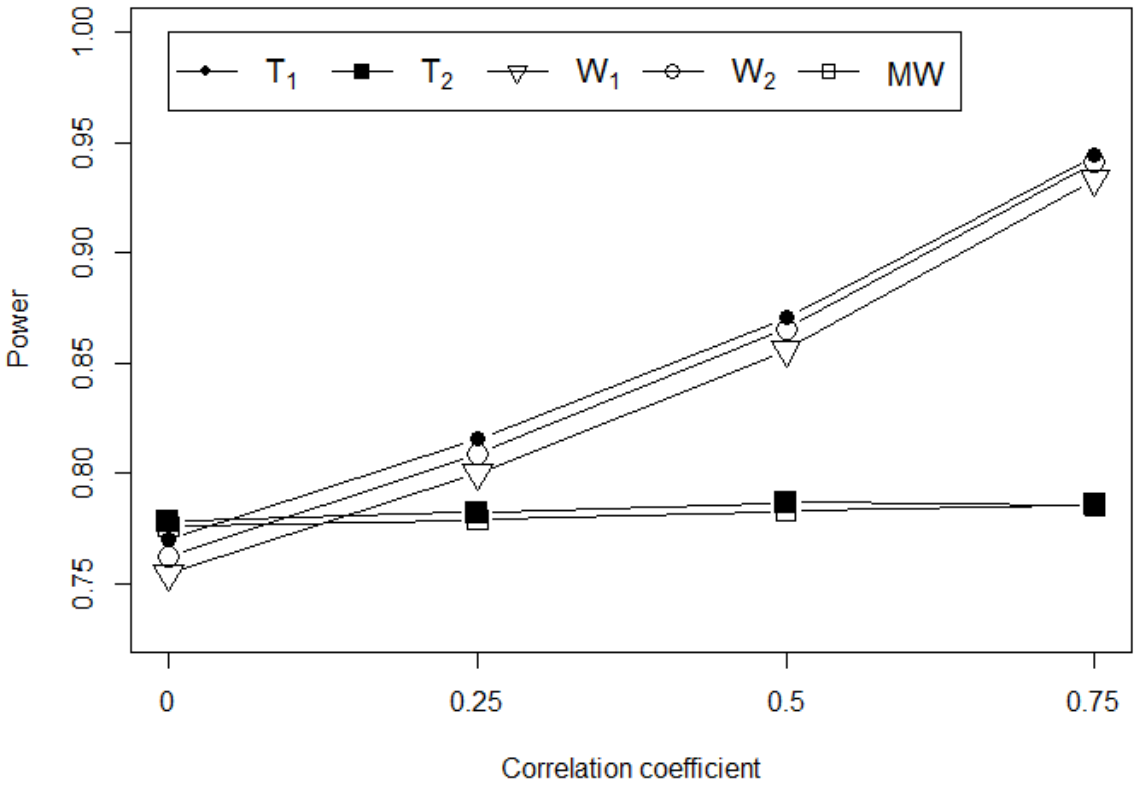

Figure 4. Power of the test statistics T_1, T_2, W_1, W_2 and MW where $n = 20$, averaged across each scenario within the simulation design.

4. CONCLUSION

Simulations have been performed based on an underlying continuum with a nonlinear transformation mapping to a five point equally spaced scoring scheme. The results indicate that parametric statistical procedures maintain good statistical properties for these data, i.e. the scores seemingly have interval like properties. This tends to suggest that if any real world application has a five point Likert scale designed to have perceived equally spaced categories, then the analyst may proceed with parametric approaches.

When comparing two independent samples on a five point Likert question, the independent samples t-test, Welch's test and the Mann-Whitney test are Type I error robust. There is little practical difference between the power of these three tests. These findings support those in the literature (De Winter and Dodou, 2010; Rasch, Teuscher and Guiard, 2007).

When the structure of the experimental design includes paired observations, the independent samples t-test, Welch's test and the Mann-Whitney test do not fulfil all Type I error robustness

definitions. Nevertheless, these tests are conservative in nature and so their use may not be completely unjustified. However, these tests lack power in a paired design and are therefore not recommended, unless it is considered that the relationship between the two groups being compared is extremely small.

When sample sizes are large, there is little practical difference in the conclusions made from the paired samples t-test, the Wilcoxon test, or Pratt's test. When the sample size is large the choice becomes a more theoretical question about the exact form of the hypothesis being tested and the assumptions made.

When sample sizes are small and the correlation between two paired groups is strong, Pratt's test outperforms the paired samples t-test and the Wilcoxon test. When the correlation between the two groups is weak, the paired samples t-test outperforms the Wilcoxon test and Pratt's test.

5. REFERENCES

Allen, I. E., Seaman, C. A. 2007. Likert scales and data analyses. Quality Progress, **40(7)**, 64.

Bellera, C. A., Julien, M., Hanley, J. A. 2010. Normal approximations to the distributions of the wilcoxon statistics: Accurate to what N? graphical insights. Journal of Statistics Education, **18(2)**, 1-17.

Boone, H. N., Boone, D. A. 2012. Analyzing likert data. Journal of Extension, **50(2)**, 1-5.

Box, G. E., Muller, M. E. 1958. A note on the generation of random normal deviates. The Annals of Mathematical Statistics, 29(2), 610-611.

Bradley, J. V. 1978. Robustness? British Journal of Mathematical and Statistical Psychology, **31(2)**, 144-152.

Carifio, J., Perla, R. J. 2007. Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. Journal of Social Sciences, **3(3)**, 106-116.

Clason, D. L., Dormody, T. J. 1994. Analyzing data measured by individual likert-type items. Journal of Agricultural Education, **35**, 4.

Conover, W. J. 1973. On methods of handling ties in the Wilcoxon signed-rank test. *Journal of the* American Statistical Association, **68(344)**, 985-988.

De Winter, J. C., Dodou, D. 2010. Five-point likert items: T test versus mann-whitney-wilcoxon. Practical Assessment, Research Evaluation, **15(11)**, 1-12.

Derrick, B., Toher, D., White, P. 2016. Why Welch's test is Type I error robust. The Quantitative Methods in Psychology, **12(1)**, 30-38.

Emerson, J. D., Moses, L. E. 1985. A note on the wilcoxon-mann-whitney test for 2 xk ordered tables. Biometrics, **41(1)**, 303-309.

Fradette, K., Keselman, H., Lix, L., Algina, J., Wilcox, R. R. 2003. Conventional and robust paired and independent-samples t tests: Type I error and power rates. Journal of Modern Applied Statistical Methods, **2(2)**, 22.

Hollander, M., Wolfe, D. A., Chicken, E. 2013. Nonparametric statistical methods. John Wiley Sons.

Jamieson, S. 2004. Likert scales: How to (ab) use them. Medical Education, **38(12)**, 1217-1218.

Kenney, J. F., Keeping, E. S. 1951. Mathematics of Statistics; Part Two, Princeton, NJ: Van Nostrand.

Likert, R. 1932. A technique for the measurement of attitudes. Archives of Psychology.

Mehta, J., Srinivasan, R. 1970. On the Behrens—Fisher problem. Biometrika, **57(3)**, 649-655.

Nanna, M. J., Sawilowsky, S. S. 1998. Analysis of likert scale data in disability and medical rehabilitation research. Psychological Methods, **3(1)**, 55.

Norman, G. 2010. Likert scales, levels of measurement and the “laws” of statistics. Advances in Health Sciences Education, **15(5)**, 625-632.

Pratt, J. W. 1959. Remarks on zeros and ties in the wilcoxon signed rank procedures. Journal of the American Statistical Association, **54(287)**, 655-667.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org. version 3.1.3.

Rasch, D., Teuscher, F., Guiard, V. 2007. How robust are tests for two independent samples? Journal of Statistical Planning and Inference, **137(8)**, 2706-2720.

Serlin, R. C., 2000. Testing for robustness in monte carlo studies. Psychological Methods, **5(2)**, 230.

Sisson, D. V., Stocker, H. R. 1989. Research corner: Analyzing and interpreting likert-type survey data. Delta Pi Epsilon Journal, 31(2), 81.

Stevens, S. S. 1946. On the theory of scales of measurement. American Association for the Advancement of Science. **103(2684)**, 667-680.

Sullivan, G. M., Artino Jr, A. R. 2013. Analyzing and interpreting data from likert-type scales. Journal of Graduate Medical Education, **5(4)**, 541-542.

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(12)**, 1938-1943.

Vonesh, E. F. 1983. Efficiency of repeated measures designs versus completely randomized designs based on multiple comparisons. Communications in Statistics-Theory and Methods, **12(3)**, 289-301.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. Biometrics Bulletin, **1(6)**, 80-83.

Zimmerman, D. W. 1997. Teacher’s corner: A note on interpretation of the paired-samples t test. Journal of Educational and Behavioral Statistics, **22(3)**, 349-360.