International Journal of Assessment Tools in Education
2023, Vol. 10, No. 1, 114–128

https://doi.org/10.21449/ijate.1183622

Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate ***Research Article***

The power and type I error of Wilcoxon-Mann-Whitney, Welch's *t*, and student's *t* tests for likert-type data

Ahmet Salih Simsek [1,*]

[1]Kırsehir Ahi Evran University, Department of Measurement and Evaluation in Education, Türkiye

ARTICLE HISTORY

Received: Oct. 03, 2022

Accepted: Mar. 05, 2023

Keywords:

Likert,
Wilcoxon-Mann-Whitney,
Student's t,
Welch's t,
Unequal group size.

Abstract: Likert-type item is the most popular response format for collecting data in social, educational, and psychological studies through scales or questionnaires. However, there is no consensus on whether parametric or non-parametric tests should be preferred when analyzing Likert-type data. This study examined the statistical power of parametric and non-parametric tests when each Likert-type item was analyzed independently in survey studies. The main purpose of the study is to examine the statistical power of Wilcoxon-Mann-Whitney, Welch's t, and Student's *t* tests for Likert-type data, which are pairwise comparison tests. For this purpose, a Monte Carlo simulation study was conducted. The statistical significance of the selected tests was examined under the conditions of sample size, group size ratio, and effect size. The results showed that the Wilcoxon-Mann-Whitney test was superior to its counterparts, especially for small samples and unequal group sizes. However, the Student's *t*-test for Likert-type data had similar statistical power to the Wilcoxon-Mann-Whitney test under conditions of equal group sizes when the sample size was 200 or more. Consistent with the empirical results, practical recommendations were provided for researchers on what to consider when collecting and analyzing Likert-type data.

1. INTRODUCTION

Likert-type item is the most preferred response format for measuring characteristics of individuals in self-report instruments such as questionnaires and scales. The results of parametric or non-parametric analyses using this item type have been reported in many studies (Schrum *et al.,* 2020). One opinion in the literature is that Likert data can be considered interval data (Norman, 2010), while another opinion is that they are ordinal (Calver & Fletcher, 2020; Carifio & Perla, 2008). However, Likert scale data (i.e., the data that are the sum or mean of Likert items) can be analyzed under the assumption that they are interval data (Boone & Boone, 2017). For interval data, it is well known that the Student's *t*-test has better statistical power than the U-test in many cases when comparing means (Boneau, 1962; Glass et.al., 1972; Zimmerman & Zumbo, 1990; Bindak, 2014). Some researchers claim that this approach is also valid for data such as the Likert scale (Norman, 2010). Discussions on this issue can only be clarified with simulation studies.

*CONTACT: Ahmet Salih Şimşek ✉ asalihsimsek@gmail.com Kirsehir Ahi Evran University, Department of Measurement and Evaluation in Education, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

The results of comparative tests performed separately for Likert items are presented in many studies. Although non-parametric tests (e.g., Wilcoxon-Mann-Whitney) should be used for each Likert item considering its measurement level, it is observed that parametric tests (e.g., Student's t) are used (Liddel & Kruschke, 2018; Schrum *et al.,* 2020; Wu & Leung, 2017). It is clear that there is still confusion about which test should be preferred for Likert items. There are few simulation studies on which test is better than its counterpart for comparing groups (de Winter & Dodou, 2010; Derrick & White, 2017). This gap in knowledge in the literature leads to debates about the analysis of Likert data.

A simulation study reported that the Student's *t*-test and the U-test generally have similar power for five-point Likert items and that strong differences in power between the Student's *t*-test and the U-test occur when one of the samples is drawn from a multimodal distribution (de Winter & Dodou, 2010). The difference in power between the two tests for the fourteen different response patterns is also presented in the same study. In another study, Derrick & White (2017) examined the power of comparison tests for paired Likert data. The study, which examined the power of tests considering sample size, the correlation between paired observations, and the distribution of responses, emphasized that the paired samples *t*-test was not appropriate for paired Likert data (Derrick & White, 2017). The conditions considered in each simulation study differ from each other. Distributional characteristics (skewed, multimodal, normal, nonnormal, homogeneous, etc.), group size (equal, not equal), and sample size (small, medium, large) are the simulation conditions commonly used for between-group comparison tests. Testing all conditions in a single simulation run makes the interpretation of results difficult. For this reason, the performance of statistical tests is usually compared under certain selected simulation conditions.

The sample size is one of the parameters needed to calculate the statistical power of a test. For parametric mean comparison tests, it is desirable that the sample size for each group is not less than 30. However, previous studies have reported that Welch's *t*-test performs better than Student's *t*-test even for extremely small sample sizes (e.g., 2, 3, 5) (de Winter, 2013). On the other hand, a recent study highlighted that the desired statistical power cannot be achieved with the Student's *t*-test for small sample sizes (e.g., 15, 30, 50), so the sample size should not be less than 100 (Sangthong, 2020). Further simulation studies are needed on this topic.

Another important parameter affecting statistical power is the ratio of group sizes. It is well known that, especially for parametric tests, maximum power is achieved for a given total sample size when the groups are of equal size (1:1) (Kim & Park, 2019). Minimum required sample size to achieve the same level of power increases when group sizes are not equal (Bulus, 2021). The Student's *t* test is robust when the groups are equal in size, but in practice, almost all studies contain unequal group sizes (Ruxton, 2006). In another study, Ahad and Yahaya (2014) showed that the statistical power of Welch's test decreases dramatically under conditions of unequal group size. In the case of unequal group size (e.g., 1:2, 1:3, 1:4), the desired statistical power may not be achieved.

The relationship between effect size and the statistical power of the test is one of the most frequently asked questions in research (Bulus, 2021; Bulus, 2022; Bulus & Dong, 2021; Dong & Maynard, 2013). In addition, the statistical power of tests that compare the means of independent groups is not independent of effect size (Wiedermann & Eye, 2013). A higher effect size leads to higher statistical power (Ahad & Yahaya, 2014; de Winter, 2013). However, power differences between statistical tests often become apparent in cases where a smaller effect size is obtained. Therefore, it is necessary to evaluate the power of statistical tests under different effect size conditions.

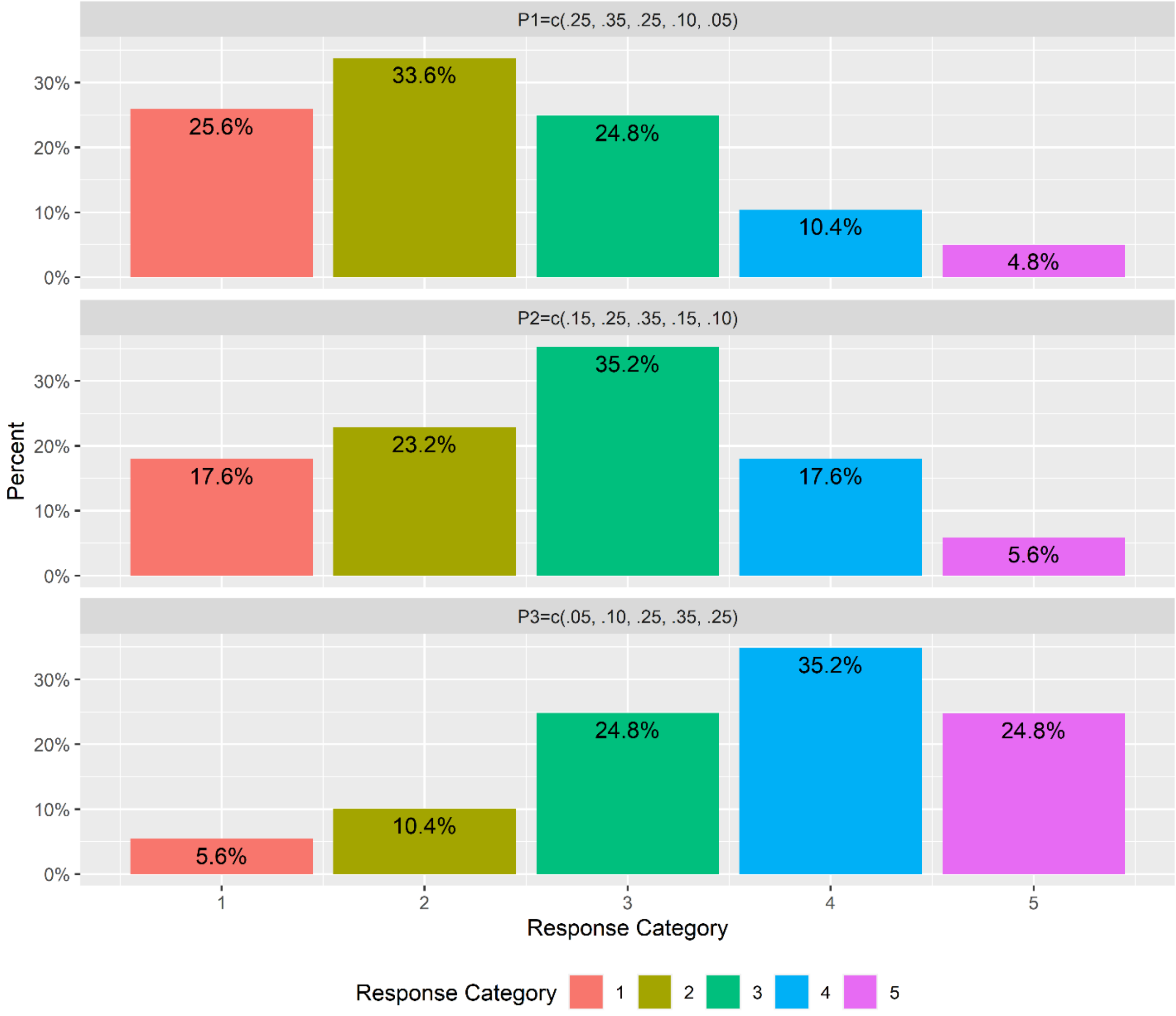
The current literature on the statistical power of tests comparing means between two groups is reviewed in this study. Although there is some disagreement about the effectiveness of parametric and nonparametric tests under different conditions, there is a lack of research on the use of parametric tests to compare the means of two groups for Likert scale data. Few studies have addressed this issue (de Winter & Dodou, 2010; Derrick & White, 2017). The gap in the literature regarding the lack of empirical studies examining different sample sizes, group size ratios, and effect sizes for Likert-type data is significant. This study is unique in that it aims to examine experimental conditions that have not previously been examined in selected comparison tests. Empirical evidence of the performance of parametric and nonparametric tests in analyzing Likert-type data under various conditions, such as sample size, group size ratio, and effect size, can help reduce uncertainties in the literature. Therefore, the objective of this study is to investigate the performance of Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests for Likert-type data using the Monte Carlo simulation method under different sample characteristics. The study evaluates the statistical power of the selected parametric and nonparametric tests under simulated conditions with different sample sizes, group size ratios, and effect sizes. Based on the research findings, this study provides some guidelines for researchers interested in analyzing Likert-type data.

2. METHOD

The Monte Carlo simulation study was conducted with the data generated for a 5-point Likert item. During data generation, three different populations of respondents were designed in which the responses "disagree," "neutral," and "agree" predominated. The probability distribution of the responses designed for the three respondent populations according to the categories from 1 to 5 is P1=c(.25, .35, .25, .10, .05), P2=c(.15, .25, .35, .15, .10), and P3=c(.05, .10, .25, .35, .25), respectively. The probabilities determined for the response patterns are consistent with research outcomes in the literature (see de Winter & Dodou, 2010). The distribution of the categories of the generated data according to the population is shown in Figure 1. The data were generated with a total of 15000 observations, 5000 for each group.

In order to examine the statistical power of the selected tests at different effect sizes, samples were drawn from the P1-P2, P2-P3, and P1-P3 population pairs, and the statistical power of the selected tests was examined. In this way, the statistical power for the selected tests was calculated in the samples of populations with small (.20<d≤.50), medium (.50<d≤1.00), and large (d>1.00) effect sizes. The responses for populations P1, P2, and P3 were distributed with a mean of 2.34, 2.71, and 3.63, respectively. The effect size of the difference between populations in the generated data was determined to be 0.32 for P1-P2, 0.82 for P2-P3, and 1.15 for P1-P3. With three different effect sizes (small, medium, large), five different sample sizes (N=30, N=50, N=100, N=200, N=400), and three different group sizes ratios (1:1, 1:2, 1:4), the analyzes were performed with 5000 replicates for each condition. Table 1 shows the group sizes for the sample size and ratio of group size conditions. The results for the selected tests were obtained by analyzing the results of 225000 samples.

Figure 1. *Distribution of the response by Likert response category.*



Note. 5-point Likert response categories were expressed as 1 - disagree, 3 - neutral, and 5 - agree

Table 1. *The sample sizes and group sizes in the simulations*

	Sample size									
	N=30		N=50		N=100		N=200		N=400	
Group size ratio	n_1	n_2	n_1	n_2	n_1	n_2	n_1	n_2	n_1	n_2
1:1	15	15	25	25	50	50	100	100	200	200
1:2	10	20	17	34	33	67	66	134	133	267
1:4	6	24	10	40	20	80	40	160	80	320

Note. n_1 and n_2 indicate the group size for independent samples

Understanding the distributional properties of the data generated is a critical factor to consider when evaluating the results of comparative tests. With this in mind, the assumption of homogeneity of variance, a crucial assumption for parametric tests such as Student's *t*-test (Field, 2009), has been reviewed. The assumption of homogeneity of variances for the 225.000 samples drawn was tested using Levene's test. It was found that the assumption of homogeneity of variances was met for the majority of the samples (96% of all replications). In those cases where the homogeneity assumption was not met, a chi-square analysis was performed to identify if this was due to sample size and group size ratio bias. The results showed that there was no bias in either sample size (chi-square=0.2597, p=.878) or group size ratio (chi-square= 4.0562, p=.398). Likert-type data, by their nature, are interval data and are not expected to be normally distributed. However, Student's *t*-test, which is known to be robust to violations of the normality assumption (Bridge & Sawilowsky, 1999; Zimmerman, 1985). Heeren and D'Agostino (1999), show the robustness of the two independent samples *t*-tests to type I errors for Likert-type data when the sample size is small. One of the main objectives of the study is to investigate the statistical power and type I error rate of Student's *t*-test when the required assumptions are not met. As in similar simulation studies, the normality assumption for Likert-type data was not considered in this study. Data generation and analysis were performed using R software. Part of the simulation study codes can be found in the Appendix. Statistical power calculations were performed using the WMWssp (v.0.4.0) and MESS (v.0.5.7) packages. The statistical power of the Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests for the simulation conditions were compared using statistical and graphical methods.

3. RESULTS

The study first examined the statistical power and type I error rate of the selected tests based on sample size and group size ratio. The mean values of the type I error rate and statistical power of the Wilcoxon-Mann-Whitney (WMW), Student's t, and Welch's *t*-tests for all effect size conditions were presented in Table 2 without being reported separately by effect size classification. The selected tests were evaluated by comparison with reference values, using an alpha level of 0.05 for the two-way hypothesis and a minimum statistical power of 0.80 suggested by Cohen (1988).

Table 2. *The type I error and the power of Wilcoxon-Mann-Whitney, Student's t, and Welch's t by sample size and group size ratio.*

Sample size	Group size ratio	Method	Power				Type I Error			
			Mean	sd	se	ci	Mean	sd	se	ci
N=30	1:1	WMW	0.540	0.33	0.003	0.005	0.051	0.22	0.003	0.006
N=30	1:1	Student's t	0.532	0.34	0.003	0.005	0.052	0.22	0.003	0.006
N=30	1:1	Welch's t	0.532	0.34	0.003	0.005	0.051	0.22	0.003	0.006
N=30	1:2	WMW	0.541	0.34	0.003	0.005	0.047	0.21	0.003	0.006
N=30	1:2	Student's t	0.505	0.33	0.003	0.005	0.052	0.22	0.003	0.006
N=30	1:2	Welch's t	0.496	0.33	0.003	0.005	0.051	0.22	0.003	0.006
N=30	1:4	WMW	0.550	0.34	0.003	0.005	0.044	0.21	0.003	0.006
N=30	1:4	Student's t	0.435	0.31	0.003	0.005	0.048	0.21	0.003	0.006
N=30	1:4	Welch's t	0.392	0.29	0.002	0.005	0.056	0.23	0.003	0.006
N=50	1:1	WMW	0.654	0.34	0.003	0.005	0.050	0.22	0.003	0.006
N=50	1:1	Student's t	0.645	0.34	0.003	0.005	0.050	0.22	0.003	0.006
N=50	1:1	Welch's t	0.645	0.34	0.003	0.005	0.050	0.22	0.003	0.006
N=50	1:2	WMW	0.658	0.34	0.003	0.005	0.048	0.21	0.003	0.006
N=50	1:2	Student's t	0.614	0.34	0.003	0.005	0.049	0.22	0.003	0.006
N=50	1:2	Welch's t	0.609	0.34	0.003	0.005	0.052	0.22	0.003	0.006
N=50	1:4	WMW	0.657	0.34	0.003	0.005	0.052	0.22	0.003	0.006
N=50	1:4	Student's t	0.551	0.34	0.003	0.005	0.053	0.22	0.003	0.006
N=50	1:4	Welch's t	0.526	0.33	0.003	0.005	0.054	0.23	0.003	0.006
N=100	1:1	WMW	0.784	0.31	0.003	0.005	0.051	0.22	0.003	0.006
N=100	1:1	Student's t	0.773	0.32	0.003	0.005	0.050	0.22	0.003	0.006
N=100	1:1	Welch's t	0.773	0.32	0.003	0.005	0.050	0.22	0.003	0.006
N=100	1:2	WMW	0.787	0.31	0.003	0.005	0.048	0.21	0.003	0.006

N=100	1:2	Student's t	0.758	0.32	0.003	0.005	0.050	0.22	0.003	0.006
N=100	1:2	Welch's t	0.756	0.32	0.003	0.005	0.052	0.22	0.003	0.006
N=100	1:4	WMW	0.782	0.31	0.003	0.005	0.046	0.21	0.003	0.006
N=100	1:4	Student's t	0.702	0.34	0.003	0.005	0.048	0.21	0.003	0.006
N=100	1:4	Welch's t	0.693	0.34	0.003	0.005	0.051	0.22	0.003	0.006
N=200	1:1	WMW	0.874	0.24	0.002	0.004	0.052	0.22	0.003	0.006
N=200	1:1	Student's t	0.860	0.25	0.002	0.004	0.052	0.22	0.003	0.006
N=200	1:1	Welch's t	0.860	0.25	0.002	0.004	0.052	0.22	0.003	0.006
N=200	1:2	WMW	0.873	0.24	0.002	0.004	0.048	0.21	0.003	0.006
N=200	1:2	Student's t	0.847	0.27	0.002	0.004	0.048	0.21	0.003	0.006
N=200	1:2	Welch's t	0.846	0.27	0.002	0.004	0.049	0.22	0.003	0.006
N=200	1:4	WMW	0.867	0.26	0.002	0.004	0.051	0.22	0.003	0.006
N=200	1:4	Student's t	0.808	0.30	0.002	0.005	0.053	0.22	0.003	0.006
N=200	1:4	Welch's t	0.805	0.30	0.002	0.005	0.055	0.23	0.003	0.006
N=400	1:1	WMW	0.950	0.13	0.001	0.002	0.045	0.21	0.003	0.006
N=400	1:1	Student's t	0.937	0.15	0.001	0.002	0.049	0.22	0.003	0.006
N=400	1:1	Welch's t	0.937	0.15	0.001	0.002	0.049	0.22	0.003	0.006
N=400	1:2	WMW	0.947	0.14	0.001	0.002	0.045	0.21	0.003	0.006
N=400	1:2	Student's t	0.924	0.17	0.001	0.003	0.045	0.21	0.003	0.006
N=400	1:2	Welch's t	0.923	0.17	0.001	0.003	0.047	0.21	0.003	0.006
N=400	1:4	WMW	0.940	0.16	0.001	0.003	0.052	0.22	0.003	0.006
N=400	1:4	Student's t	0.888	0.22	0.002	0.004	0.052	0.22	0.003	0.006
N=400	1:4	Welch's t	0.887	0.22	0.002	0.004	0.051	0.22	0.003	0.006

Note: se: Standard error; ci: %95 confidence interval

The results for sample size show that statistical power of .80 and above is achieved for samples of N=200 and above. Regardless of the method chosen, statistical power increases in parallel with the increase in sample size. The Wilcoxon-Mann-Whitney method provides higher statistical power than the other methods, and the difference in power increases for smaller samples. Student's t and Welch's *t*-tests showed similar power for the sample size condition. The results of the group size ratio show that the Wilcoxon-Mann-Whitney method has a relatively small advantage over the other methods for the same group size, while the Wilcoxon-Mann-Whitney method has a significant power advantage over the other methods as the difference between group sizes increases. Depending on the group size ratio, higher statistical power was obtained for Wilcoxon-Mann-Whitney, Student's t, and Welch's t.

When evaluating the Type I error rate, a lower Type I error was found for the Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests when the sample size was 400. In this context, our analyses have revealed that the sample size leads to a lower Type I error rate than expected. Compared to a group size ratio of 1:4, the type I error rate has proved to be more conservative and stable under the conditions of 1:1 and 1:2. The results show that in the case where the group size ratio is the highest, the type I error rate is not consistent, and the type I error rate may exceed the expected value of 5% depending on the sample size. The results indicate that the sample size and group size ratio for Likert data have a lower Type I error than the Wilcoxon-Mann-Whitney test, Student's *t*-test, and Welch's test.

Figure 2 and Figure 3 indicate how the empirical power and Type I error rate of the Wilcoxon-Mann-Whitney, Student's t, and Welch's t methods change as a function of the interaction between sample size and group size ratio.

Figure 2. *Empirical power for Wilcoxon-Mann-Whitney, Student's t, and Welch's t in terms of sample size and group size ratio*

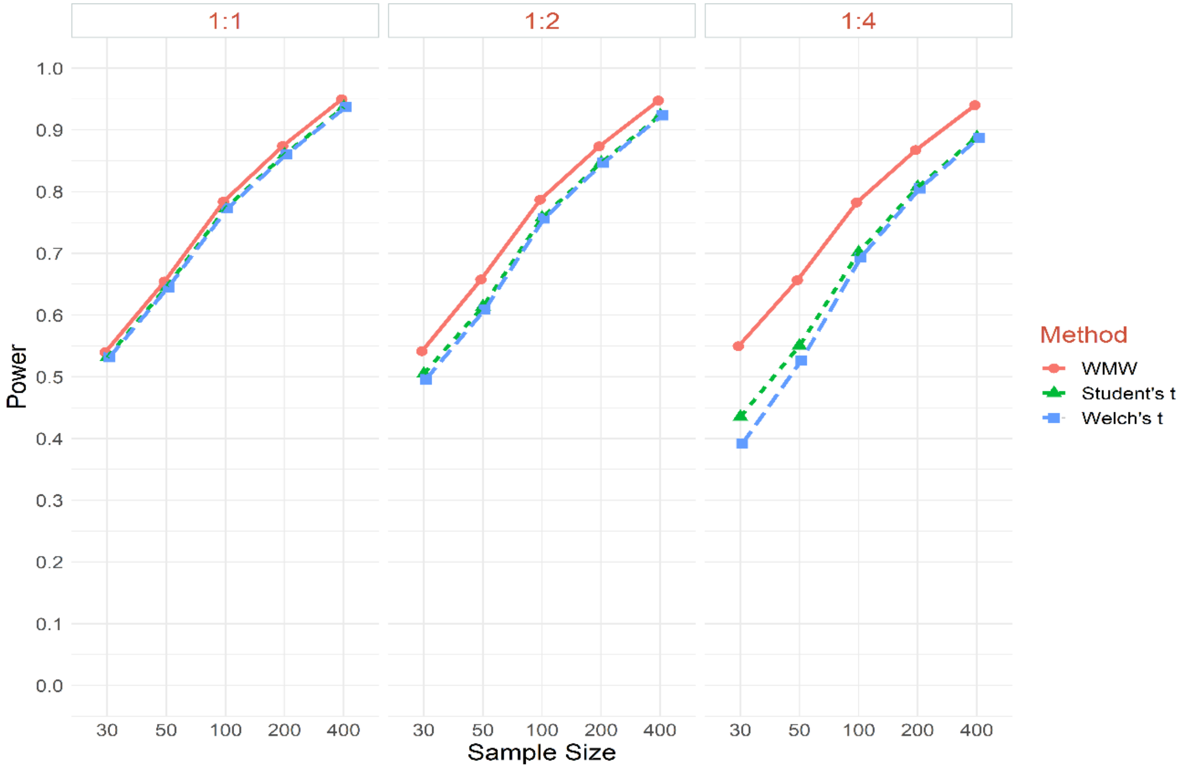


Figure 3. *Type I error rates for Wilcoxon-Mann-Whitney, Student's t, and Welch's t in terms of sample size and group size ratio*

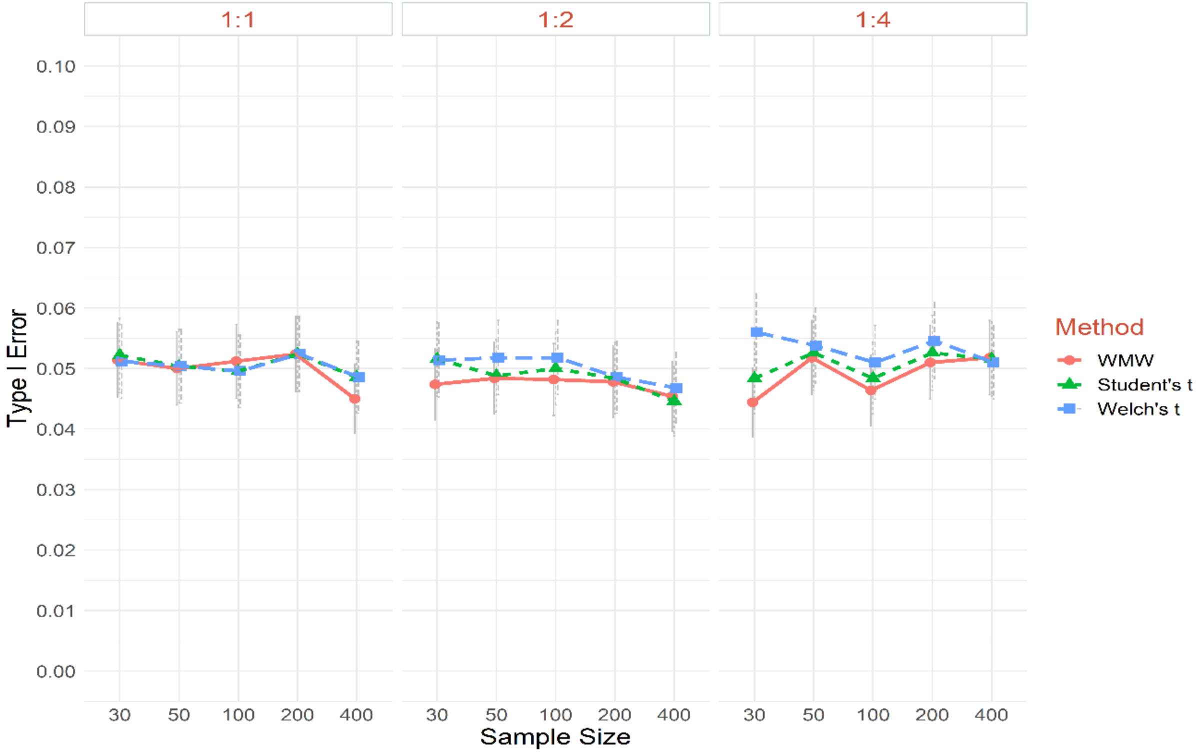
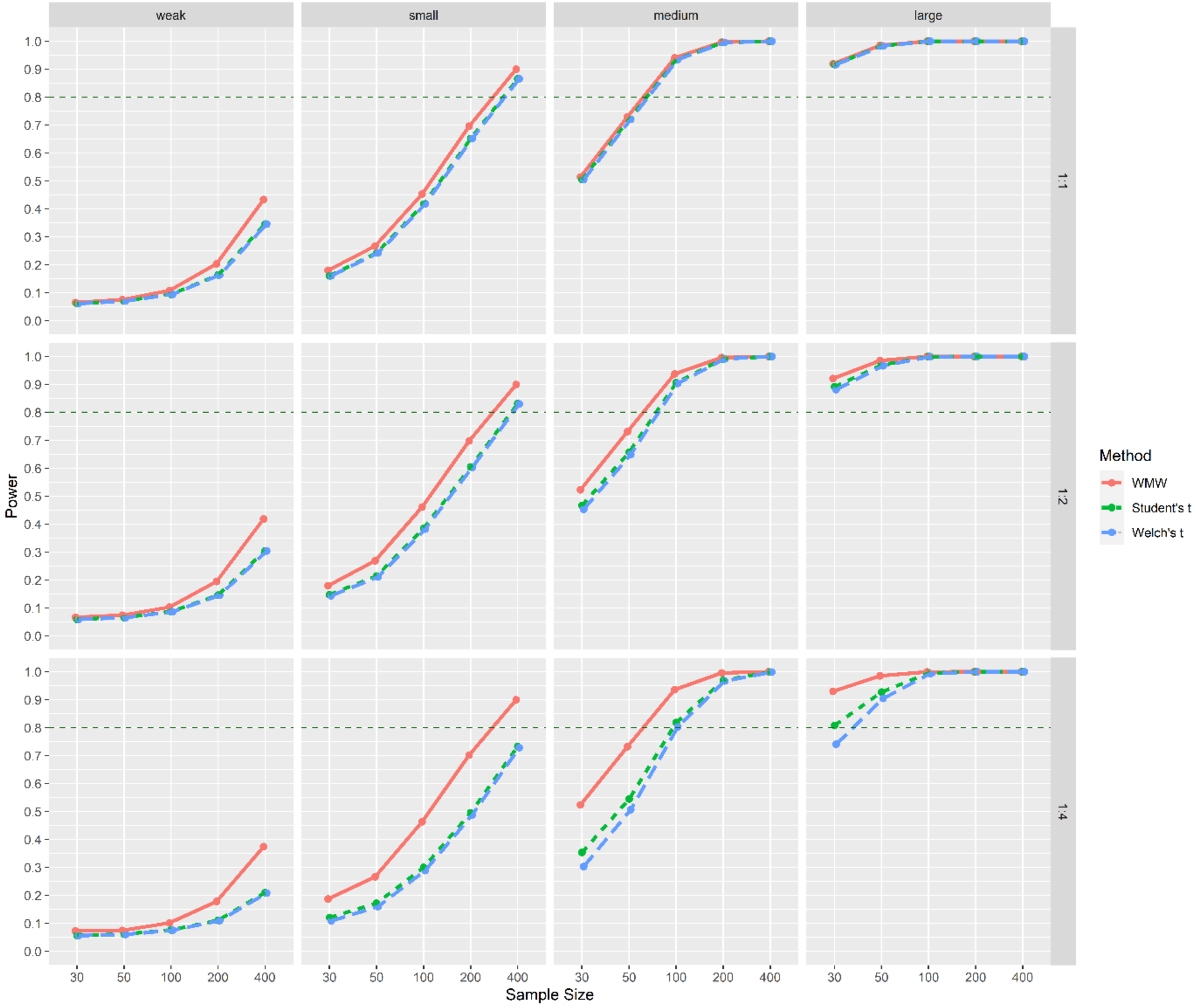


Figure 2 indicates that Wilcoxon-Mann-Whitney outperforms the other methods in terms of both sample size and group size ratio. In particular, when group sizes are unequal, Wilcoxon-Mann-Whitney seems to offer a more significant performance advantage than other methods.

Figure 4 indicates how the Wilcoxon-Mann-Whitney, Student's t, and Welch's t power functions change for effect size, sample size, and group size ratio. Looking at the line plots for the statistical power functions at weak and small effect sizes, the statistical power for the selected tests increases as the sample size increases and the group size ratio approaches 1:1. Cohen (1988) suggests that at least .80 should be used as a threshold for statistical power. For this reason, statistical power of .80 was accepted as the lower limit for evaluating the performance of the compared tests. However, it was found that the statistical power of .80 could not be achieved under any weak effect size condition, while it was achieved at N=400 for the small effect size condition. Moreover, the advantage of the power function of Wilcoxon-Mann-Whitney came into play when the difference between weak and small effect sizes, sample size, and group sizes increased. Under the conditions of medium and large effect size, Wilcoxon-Mann-Whitney responded more conservatively to changes in both sample size and group size ratio, while other methods were affected by the changes. In particular, under the condition that the group size ratio was 1:4 and the sample size was 30, Wilcoxon-Mann-Whitney had a remarkable advantage over other methods. However, it was found that although the group size ratio varied with a large effect size for samples of 50 and above, other methods offered statistical power that was close to Wilcoxon-Mann-Whitney.

Figure 4. *Line plot of the power function for effect size x sample size x group size ratio.*



4. DISCUSSION and CONCLUSION

Survey studies using Likert-type items are very common in the social sciences, e.g., education, psychology, and health. When each Likert item must be analyzed independently, researchers are faced with the question of which parametric or nonparametric tests to use. This study examined the performance of Student's t, Welch's t, and Wilcoxon-Mann-Whitney tests in analyzing Likert items for two independent groups. Under all conditions examined in the simulation studies, the Wilcoxon-Mann-Whitney test performed similarly well or better than its counterparts, Student's t and Welch's *t*-tests. The Type I error rate was found to be lower for Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests when the sample size was 400. The results revealed that a lower Type I error rate was achieved as the sample size increased, and the type I error rate was more conservative and stable under the conditions of 1:1 and 1:2 group size ratios. On the other hand, when the group size ratio was the highest, the type I error rate was found to be inconsistent and may exceed the expected value of 5% depending on the sample size. One reason why the Mann-Whitney U test may be a better choice for analyzing Likert-type data is that it measures the median difference between two groups, as opposed to the mean difference measured by the Student's *t*-test. In addition, the Mann-Whitney U test is also known to be more informative than Student's *t*-test when the modal value deviates from the mean, it is less sensitive to outliers, and it does not require a normality assumption (Wilcox, 2012; Field, 2009).

In contrast to previous studies, this study examined the statistical power of the selected tests in terms of sample size, group size ratio, and effect size for Likert-item analysis. The results obtained for sample size show that the statistical power of all tests is affected by sample size. These results are similar to those obtained in previous studies for continuous data (Dwivedi *et al.,* 2017; Ma *et al.,* 2021; Sangthong, 2020; Wiedermann & Eye, 2013). The results suggest that the findings on sample size for continuous data also apply to Likert-type data. The results also showed that Wilcoxon-Mann-Whitney was stronger than its counterparts under all sample size conditions. Previous studies support these empirical findings (de Winter & Dodou, 2010; Nanna & Sawilowsky, 1998). The second major finding of the study relates to the group size ratio. The results show that Student's t and Welch's t are sensitive to changes in the group size ratio. When the group size ratio is 1:1 for both tests, the statistical power is similar to Wilcoxon-Mann-Whitney, but when the ratio changes, the statistical power decreases. Wilcoxon-Mann-Whitney provides better power for Likert data than its counterparts when the group size ratio changes. These results are confirmed by some studies in the literature (Ahad & Yahaya, 2014; de Winter & Dodou, 2010; Dwivedi *et al.,* 2017; Zimmerman, 2004).

When the statistical power of the tests was examined considering the interaction of sample size and group size ratio, Student's t and Welch's t achieved the desired statistical power when the sample size was 200 or more. Likert-type data are defined as ordinal or interval values. Since the data structure is suitable for the structure of Wilcoxon-Mann-Whitney, it can be stated that it performs better than its counterparts. However, an important finding of the study is that Student's t and Welch's *t*-tests perform similarly well as Wilcoxon-Mann-Whitney for large samples (N > 200) and the same group ratio (1:1). This result implies that parametric tests such as Student's t can be used for the analysis of Likert-type data under certain conditions. Finally, the statistical power of the tests selected by effect size showed that high statistical power was obtained even for small samples for medium and large effect sizes. However, for the weak effect size, the desired statistical power was not achieved even with a sample size of 400. For the small effect size, it was found that the selected statistical tests could achieve the desired statistical power with a sample size of 400.

4.1. Implications

This study examined the performance of pairwise comparison tests for independent groups under conditions of sample size, group size ratio, and effect size. Considering the research findings and limitations, some suggestions for future research and researchers were developed.

4.1.1. *Methodological implications*

The results of this study have significant methodological implications for the appropriate selection of statistical tests when analyzing data collected using Likert-type scales. The simulation studies demonstrated that the Wilcoxon-Mann-Whitney test was either comparable or superior to Student's t and Welch's t-tests in terms of Type I error rate and statistical power. Additionally, the Mann-Whitney U test emerged as a suitable alternative for the Student's *t*-test when analyzing Likert-type data, as it measures the median difference between two groups and is less sensitive to outliers, and does not require a normal distribution assumption. The study results also indicated that the statistical power of the selected tests was dependent on both sample size and the ratio of group sizes, and the Wilcoxon-Mann-Whitney test was found to be more robust under all conditions. However, it is worth noting that the simulation studies were based on a distribution commonly found in Likert items while actual response distributions may vary due to human subjectivity. To validate the findings, further research using real-world survey data with high participation rates (such as PISA and TIMSS) is recommended. The impact of missing data, which is frequently encountered in real-world applications, should also be explored in future studies.

4.1.2. *Practical implications*

The study provides practical recommendations for researchers using Likert-type data in their studies. Based on the results, the Wilcoxon-Mann-Whitney test is recommended as the preferred choice over Student's *t*-test and Welch's *t*-test when analyzing Likert-type data because it provides higher statistical power. In addition, the Student's t and Welch's *t*-tests are as robust as the Wilcoxon-Mann-Whitney test when the sample size is 200 or more and the group size is the same. However, when the sample size is less than 100 and the ratio of group sizes is unequal, the Wilcoxon-Mann-Whitney test should be preferred because it provides higher power. It is emphasized that sample size, group size ratio, and effect size should be considered when selecting a pairwise comparison test for Likert-type data to achieve the desired level of statistical power.

4.2. Limitations

The present study has some limitations regarding the generalizability of the results. Specifically, the results on statistical power and Type I error rates for the tests are limited to the specific conditions examined in this study. These conditions include the use of Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests for Likert-type data in independent sample environments and sample sizes of 30, 50, 100, 200, and 400, which are commonly used in survey studies. However, the applicability of these results to smaller or larger sample sizes needs to be investigated in future studies. In addition, the power of the tests examined in this study is limited to group size ratios of 1:1, 1:2, and 1:4, which may not cover all possible group size ratios in survey studies. Therefore, further research is recommended to investigate the performance of these tests at different group sizes. Another limitation of this study is the use of only three response patterns (disagree, neutral, and agree) as a reference for the generated data. The scatter properties of the data were not considered when determining the effect size. Therefore, future studies should investigate the performance of the selected tests under different response patterns and distributional properties. In summary, although the present study provides valuable insights into the performance of Wilcoxon-Mann-Whitney, Student's t, and Welch's *t*-tests for Likert data under certain conditions, the generalizability of the results is limited with the simulation conditions. In order to improve the external validity of these results, future studies should consider a wider range of sample sizes, group sizes, response patterns, and distributional characteristics.

Acknowledgments

Acknowledgments of people, grants, and funds should be placed in a separate section before the References. If the study has been previously presented at a conference or a scholarly meeting, it should be mentioned here.

Declaration of Conflicting Interests and Ethics

The author declares no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Ahmet Salih Şimşek https://orcid.org/0000-0002-9764-3285

REFERENCES

Ahad, N.A., & Yahaya, S.S.S. (2014). Sensitivity analysis of Welch's *t*-test. AIP Conference Proceedings, 1605(February 2015), 888–893. https://doi.org/10.1063/1.4887707

Bindak, R. (2014). Comparison Mann-Whitney U Test and Students' *t* Test in Terms of Type I Error Rate and Test Power: A Monte Carlo Sımulation Study. *Afyon Kocatepe University Journal of Sciences and Engineering, 14*, 5-11. https://doi.org/10.5578/fmbd.7380

Boneau, C.A. (1962). A comparison of the power of the U and *t*-tests. Psychological Review, 69, 246-256. https://doi.org/10.1037/h0047269

Boone, H.N., Boone, D.A. 2012. Analyzing Likert data. *Journal of Extension, 50*(2), 1-5. Retrieved February 20, 2023, from https://eric.ed.gov/?id=EJ1042448

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*(3), 229-35. https://doi.org/10.1016/S0895-4356(98)00168-1

Bulus, M. (2021). Sample size determination and optimal design of randomized/non-equivalent pretest-posttest control-group designs. *Adiyaman Univesity Journal of Educational Sciences, 11*(1), 48-69. https://doi.org/10.17984/adyuebd.941434

Bulus, M. (2022). Minimum detectable effect size computations for cluster-level regression discontinuity: Specifications beyond the linear functional form. *Journal of Research on Education Effectiveness, 15*(1), 151-177. https://doi.org/10.1080/19345747.2021.1947425

Bulus, M., & Dong, N. (2021). Bound-constrained optimization of sample sizes subject to monetary restrictions in planning multilevel randomized trials and regression discontinuity studies. *The Journal of Experimental Education, 89*(2), 379-401. https://doi.org/10.1080/00220973.2019.1636197

Calver, M., & Fletcher, D. (2020). When ANOVA isn't ideal: Analyzing ordinal data from practical work in biology. *The American Biology Teacher, 82*(5), 289-294. https://doi.org/10.1525/abt.2020.82.5.289

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical education, 42*(12), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

Champagne, C.A., & Curran, P.J. (2017). Using Monte Carlo simulations to demonstrate the importance of statistical power. *The Journal of Educational Research, 110*(6), 524-532. https://doi.org/10.1080/00220671.2015.1079697

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

de Winter, J.F., & Dodou, D. (2010). Five-point Likert items: *t*-test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research, and Evaluation, 15*(1), 11. https://doi.org/10.7275/bj1p-ts64

de Winter, J.F. (2013) Using the Student's *t*-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, 18, 10. https://doi.org/10.7275/e4r6-dj05

Delacre, M., Lakens, D., & Leys, C. (2017). Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test. *International Review of Social Psychology, 30*(1), 92. https://www.rips-irsp.com/articles/10.5334/irsp.661/

Derrick, B., & White, P. (2017). Comparing two samples from an individual Likert question. *International Journal of Mathematics and Statistics, 18*(3). Retrieved February 20, 2023, from http://www.ceser.in/ceserp/index.php/ijms/article/view/4997

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24-67. https://doi.org/10.1080/19345747.2012.673143

Dwivedi, A.K., Mallawaarachchi, I., & Alvarado, L.A. (2017). Analysis of small sample size studies using non-parametric bootstrap test with pooled sampling method. *Statistics in Medicine, 36*, 2187 - 2205. https://doi.org/10.1002/sim.7263

Field, A. (2009). Discovering statistics using SPSS (3rd ed.). Sage publications.

Glass, G., Peckham, P., & Sanders, J. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42*, 237-288. https://doi.org/10.3102/00346543042003237

Harpe, S.E. (2015). How to analyze Likert and other rating scale data. Currents in Pharmacy Teaching and Learning, 7, 836-850. https://doi.org/10.1016/j.cptl.2015.08.001

Heeren, T., & D'Agostino, R.B. (1987). Robustness of the two independent samples *t*-test when applied to ordinal scaled data. *Statistics in Medicine, 6*(1), 79-90. https://doi.org/10.1002/sim.4780060110

Jamieson S. (2004). Likert scales: how to (ab)use them. *Medical education, 38*(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Kim, T.K., & Park, J.H. (2019). More about the basic assumptions of *t*-test: normality and sample size. *Korean Journal of Anesthesiology, 72*(4), 331-335. https://doi.org/10.4097/kja.d.18.00292

Liddell, T.M., & Kruschke, J.K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong?. *Journal of Experimental Social Psychology, 79*, 328-348. https://doi.org/10.1016/j.jesp.2018.08.009

Ma, C., Wang, X., Xia, L., Cheng, X., & Qiu, L. (2021). Effect of sample size and the traditional parametric, non-parametric, and robust methods on the establishment of reference intervals: Evidence from real-world data. *Clinical Biochemistry, 92*, 67–70. https://doi.org/10.1016/j.clinbiochem.2021.03.006

Nanna, M.J., & Sawilowsky, S.S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods, 3*(1), 55-67. https://doi.org/10.1037/1082-989X.3.1.55

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*, 625-632. https://doi.org/10.1007/s10459-010-9222-y

Ruxton, G.D. (2006). The unequal variance Student's *t* testis an underused alternative to Student's *t* test and the Mann–Whitney U test. *Behavioral Ecology, 17*(4), 688–690. https://doi.org/10.1093/beheco/ark016

Sangthong, M. (2020). The Effect of the Likert Point Scale and Sample Size on the Efficiency of Parametric and Non-parametric Tests. *Thailand Statistician, 18*(1), 55–64.

Schrum, M.L., Johnson, M., Ghuy, M., & Gombolay, M.C. (2020). *Four years in review: Statistical practices of Likert scales in human-robot interaction studies*. In Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (pp. 43-52). https://doi.org/10.1145/3371382.3380739

Wiedermann, W., & von Eye, A. (2013). Robustness and power of the parametric *t*-test and the non-parametric Wilcoxon test under non-independence of observations. *Psychological Test and Assessment Modeling, 55*(1), 39-61.

Wilcox, R.R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Academic Press.

Wu, H., & Leung, S.O. (2017). Can Likert scales be treated as interval scales? Simulation study. *Journal of Social Service Research, 43*(4), 527-532. https://doi.org/10.1080/01488376.2017.1329775

Zimmerman D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*, 173-181. https://doi.org/10.1348/000711004849222

Zimmerman, D.W. & Zumbo, B.D. (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, https://doi.org/10.1080/00221309.1990.9921148

Zimmerman, D.W. (1985). Power Functions of the *t*-test and Mann-Whitney U Test Under Violation of Parametric Assumptions. *Perceptual and Motor Skills, 61*, 467 - 470. https://doi.org/10.2466/pms.1985.61.2.467

APPENDIX

```
for (i in 1:5) {  #SS
  for (j in 1:3) {  #Dist
    for (k in 1:100) {  #iterid
iterid<-k
Nsize<-Ssize[i]  #Nsizeal data size
N1ratio<-Sprop[j]  #Group distribution proportion
#-- GROUP STAT
n1<-round(Nsize*N1ratio,0)
n2<-Nsize-n1
nmin<-min(n1,n2)
Nratio<-round((1-N1ratio)/N1ratio,0)
sG1<-sample(dG1,n1)
sG2<-sample(dG2,n2)
Diff.p = abs(mean(sG1)-mean(sG2))/min(mean(sG1),mean(sG2))
#-- cohen's d effect size
SD1<-(n1-1)*((sdG1)^2)
SD2<-(n2-1)*((sdG2)^2)
sdpooled<-sqrt((SD1+SD2)/(n1+n2-2))
d<-abs((mG1-mG2)/sdpooled)
#--- POWER TEST for WMW
#library(WMWssp)
p_WMW<-as.numeric(WMWssp_maximize(sG1, sG2, alpha = alpha, N=Nsize)$power)
#--- POWER TEST for t-test
#library("MESS") -> unequal group size
p_ttest<-as.numeric(power_t_test(n=nmin, delta=d, sig.level = alpha, ratio=Nratio, sd.ratio=SDratio,
df.method ="classical", type = "two.sample", alternative="two.sided")$power)
p_welch<-as.numeric(power_t_test(n=nmin, delta=d, sig.level = alpha, ratio=Nratio, sd.ratio=SDratio,
df.method ="welch", type = "two.sample", alternative="two.sided")$power)
#--- p value for WMW
#library(rcompanion)
WMW.z<-as.numeric(wilcoxonZ(sG1,sG2))
WMW.d<-as.numeric(wilcoxonRG(WMW_Y,WMW_G))
WMW.w<-as.numeric(wilcox.test(sG1,sG2, alternative = "two.sided", exact = FALSE)$statistic)
WMW.p<-as.numeric(wilcox.test(sG1,sG2, alternative = "two.sided", exact = FALSE)$p.value)
#--- p value for t-test
ttest.t<-as.numeric(t.test(sG1,sG2, alternative = "two.sided", var.equal = TRUE)$statistic)
ttest.p<-as.numeric(t.test(sG1,sG2, alternative = "two.sided", var.equal = TRUE)$p.value)
#--- p value for welch
welch.t<-as.numeric(t.test(sG1,sG2, alternative = "two.sided", var.equal = FALSE)$statistic)
welch.p<-as.numeric(t.test(sG1,sG2, alternative = "two.sided", var.equal = FALSE)$p.value)
#--- Homogeniy of Variance Analysis
#library(DescTools)
WMW_Y<-c(sG1,sG2)
WMW_G<-factor(c(rep("1", length(sG1)),  rep("2", length(sG2))))
levene.p<-as.numeric(LeveneTest(WMW_Y,WMW_G)$`Pr(>F)`[1])
levene.factor<-ifelse(levene.p>.05, 1, 0)
#--- Normality
#library(DescTools)
G1.p.norm<-shapiro.test(sG1)$p.value
G2.p.norm<-shapiro.test(sG2)$p.value
G1.factor.norm<-ifelse(G1.p.norm>.05, 1, 0)
G2.factor.norm<-ifelse(G2.p.norm>.05, 1, 0)
interimdata<-as.vector(cbind(simid,iterid,iNsize,iNratio,
                    n1,n2,d,d.cat,
                    p_Wilcoxon-Mann-Whitney,p_ttest,p_welch,
                    mG1,mG2,sdG1,sdG2,
                    WMW.z,WMW.d,
                    WMW.w,WMW.p,
                    ttest.t,ttest.p,
                    welch.t,welch.p,
                    levene.p,levene.factor,
                    G1.p.norm,G2.p.norm,
                    G1.factor.norm,G2.factor.norm))
simdata<-rbind(simdata,interimdata)
simid<-simid+1
iterid<-k+1
} j<-j+1} i<-i+1}
```