Testing Significance Tests: A Simulation with Cliff's Delta, *t*-tests, and Mann-Whitney *U*

Tyler Barnes M.S.

Wright State University/National Center for Organizational Development

Scott C. Moore Ph.D.

National Center for Organizational Development

Katerine Osatuke Ph.D.

National Center for Organizational Development

The National Center for Organizational Development (NCOD) at the Department of Veteran Affairs (VA) annually conducts an organizational census for VA. This survey is called the All Employee Survey (AES). The AES serves to evaluate employee engagement and satisfaction throughout VA locations and hierarchical levels, for the purposes of recognizing best practices as well as areas of concern and using these data to plan workplace improvements. As part of this process, survey ratings are compared across VA locations as well as across time. Analyses for each comparison determine 1) if the differences between the comparisons are statistically significant and 2) if the differences are large enough to be pragmatically relevant. This paper focuses on the first aspect (statistical significance). For this purpose, we use Cliff's Delta (Cliff, 1993). We will first describe the relevant aspects of the AES to explain the context of this study. Next, we will explain our reasoning for using Cliff's Delta. Then we will explore other alternatives for these pairwise comparisons and compare them to Cliff's Delta.

AES

The AES, an annual census survey at the Department of Veteran Affairs, with an approximate 60% response rate for the last decade (e.g., 59.6% at the last administration in 2017), assesses a range of workplace perceptions that collectively reflect organizational health issues, such as satisfaction, relationships between co-workers, leadership effectiveness, and so forth. Most of the survey items use a Likert-type scale rating from 1 to 5 where 1 is a strong negative response (e.g., Very Dissatisfied) and 5 is strong positive response (e.g., Strongly Agree). As part of summarizing the data yielded by the survey process, scores for VA locations at various hierarchical levels (e.g., work units; specific clinics or program offices; medical centers that include multiple clinics; integrated networks that combine several medical centers, etc.) are compared at various levels to themselves (across years) and also compared to meaningful reference groups (e.g., workgroup at a site as compared to the rest of the site; medical center compared to the other medical centers that are part of the same integrated network). Once the data are summarized in terms of these comparisons, the results are then broadly shared across

VA. This allows data users—VA leaders and employees at various levels—to compare their scores to similar groups, decide if there are strengths that can be enhanced or concerns that need to be addressed, and make plans for exchange of best practices or for making any needed improvements.

Pairwise Comparisons

Cliff's Delta

To determine if there are differences, Cliff's Delta is used to compare the scores. This statistic was chosen for multiple reasons, including that it is an effect size with a significance value, and it was built for ordinal data (same data type that is on the AES), and therefore is non-parametric.

For the AES, Cliff's Delta is used in two ways to determine significance difference between scores. First, we use the statistic to determine if the differences between the distribution of scores are statistically significant. Then we use the distribution of effect sizes for all comparisons to determine if the differences are practically relevant. Because Cliff's Delta is an effect size with a significance value, both goals can be accomplished in one analysis. This makes Cliff's Delta a practical choice.

Cliff's Delta also was built for ordinal data. This matches the type of data gathered from the AES. This makes Cliff's Delta a logical choice for its use. I will address this issue in more detail in the subsequent section.

Because Cliff's Delta is for ordinal data analysis, it is also non-parametric. There seems to be consensus that non-parametric statistics seem to outperform parametric statistics in conditions that violate the assumptions of parametric tests. However, they tend to underperform in conditions that are optimal for parametric statistics (e.g., Cliff, 1993). Often, in the AES data, we have variables that are heavily skewed, at least within the measured part of the continuum. This makes the Cliff's Delta a logical choice as well.

Alternatives

There are many pairwise independent group comparisons: several types of t – tests, Mann-Whitney U, Chi-Square Test of Homogeneity, etc. Of the many pairwise comparisons, Cliff's Delta is not used frequently. It is logical that there could be multiple possible choices suitable for the AES. However, we will focus our discussion to three other alternatives, Student's t, Welch's t, and the Mann-Whitney U. We chose Student's t because it is a very common choice, and arguably the standard, for pairwise comparisons. We chose the Mann-Whitney U because it is the non-parametric analogue to Student's t. We chose Welch's t because it is a parametric version of the t -test that helps when the data violate the assumption of homogeneity of variances.

Per tradition, it seems like the t -test would be a viable alternative to Cliff's Delta. The t – test is generally robust to violations of its assumptions (Boneau, 1960). In addition, it tends to control for Type I error well in most conditions (especially in comparison to Cliff's Delta; Kromney & Hogarty, 1998). For power, it does not always perform as well as some other statistics in all

conditions (Kromney & Hogarty, 1998), especially in non-normal distributions (Blair & Higgins, 1980) like those in the AES.

Another concern is the validity of the use and interpretation of a *t* -test in relation to the ordinal scale. Although a review of this 70-year-old debate is beyond the scope of this paper, there are some aspects that need to be addressed. Previous research has mixed opinions on this issue. Baker, Hardyck, & Petrinovich (1966) claimed that *t*-tests (often assumed to be for interval/ratio data) are not greatly affected by scale of measurement barring a few precautions. Lord (1953) claimed, "Since the numbers don't remember where they came from, they always behave just the same way, regardless." (p. 21 On the other side, there are some researchers who understand this to be true, but claim the validity of the substantive inferences (not statistical inferences) is called into question (Cliff, 1993). It is important to note that a *t* – test is a test of means. Means may not be the best way to describe differences in distributions in all situations.

Another alternative to the Student's *t* – test is to use Welch's Correction to the pooled standard deviation term. This correction allows the standard *t* -test to become robust to the homogeneity of variances assumption. However, research has demonstrated that the Welch's *t* – tests become more conservative as the distributions become less normal (Reed, 2005) and when sample sizes are small (Yuen, 1974).

Another alternative is the Mann-Whitney *U*. This has the benefit of being non-parametric (and therefore fine to use for ordinal data). It has also shown that it can be more powerful than a standard *t* – test (Nanna & Sawilowsky, 1998) and especially in conditions of non-normality (Blair & Higgins, 1980). To date, we could not find an article empirically comparing Cliff's delta to the Mann-Whitney *U* in power and error rates. However, the Mann-Whitney *U* is more formally a test of if two distributions are identical (Mann & Whitney, 1947), but is constrained by not testing the shape and spread of the distribution (Cliff, 1993).

Purpose

Although there has been much empirical research and conceptualizing of these kinds of issues, there remains one problem. As posited by Cliff (1993), many of these studies tend to be narrow or not generalizable to every circumstance. Most studies have relatively small sample sizes as compared to the AES. For example, Blair & Higgins (1980) have a condition with three cases in one group, with one of the largest being $n = 100$ in Kromrey & Hogarty (1998). More recently, a study had a condition with a sample size of $n = 500$ (Poncet, Courvoisier, Combescure, & Perneger, 2016). In the AES, there are often comparisons made with thousands of cases for each group.

In addition to the sample size issue, there is also a component of unequal sample sizes in the AES. Some of the researchers incorporated unequal sample sizes (e.g., Kromrey & Hogarty, 1998), however the difference between groups does not reflect some of the situations in the AES where there could be a comparison of a group with less than 10 responses with a group of over 1,000 responses at the extreme end. Given these differences in parameters between previous research and the AES parameters, the purpose of this study is to examine which pairwise comparison statistic is the most accurate given some parameters of AES data.

Method

Design

This study is a 5x2x6x3x3 fully-crossed simulation with 100 replications each. The first condition is sample size (20,50,100, 1000,10000). The second condition is a normal distribution or a skewed normal distribution. The third is mean differences (0, 0.1, 0.2, 0.3, 0.4, 0.5). The fourth and fifth conditions are the standard deviation of both items (.5,.75,1). An additional simulation will replace condition one with unequal sample sizes (N_1= 17 and N_2 = 742 derived from common comparisons the VA encounters in a large scale organizational survey) while maintaining the remaining conditions. All of these parameters were derived from properties of the AES to make the results most applicable for that purpose.

Data Generation

All samples were generated in R as a graphic rating scale ranging from 1 to 5. For the first simulation, the sample sizes for both groups were equal. For the normal distribution, one group always had a mean of 3 and the SD depended on the condition stated above (.5,.75, and 1). To simulate the mean difference conditions, the second group mean varied in .1 increments (3, 3.1, 3.2, 3.3, 3.4, and 3.5). Their standard deviations varied by condition as explained above. To make data generation more efficient, we allowed for some variability in these parameters. The means could differ from their conditions by ± .1 and the SDs could vary by ± .05. However, the mean differences between the two groups had to fall within a certain range for each condition (0 = 0-.02, 0.1 = .09-.11, 0.2 = .19-.21, 0.3 = .29-.31, 0.4 = .39-.41, 0.5 = .49-.51). For example, if the mean of group 1 was 3 and the condition was for a mean difference of 0.3, then group 2's mean had to be equal to or between 3.29 and 3.31. In all instances, the mean of group 2 was either the same or higher than group 1's mean. All rules and allowances stated above were the same for the skewed samples. The only difference was that the mean of the first group shifted to 3.5 and subsequently group 2's means increased for a maximum of 4.0 in the largest mean difference conditions.

For the second simulation, the only difference was the sample sizes of each group. These were fully-crossed. Group 1 had N = 17 and N=742 depending on the condition and the same was true for group 2.

Data Analysis

The pairwise comparisons we chose to use were Student's *t*, Welch's *t*, Cliff's Delta, and the Mann-Whitney *U*. We did not test any paired-sample statistics. Excluding Cliff's Delta, all analyses were run using base functions in R. Cliff's Delta was run using a function created, tested, and used by the VHA National Center for Organizational Development for AES. For each condition, we calculated all four statistics and assessed their significance using the standard criterion of $p < .05$ as a significant difference. To assess power, we counted the number of significant differences in the mean difference conditions for each statistic and divided the sum by the number of replications (100). To assess error, we used the same method, but error could only come from the no mean difference condition.

Results

Equal Ns

N = 20

Table 1 shows the power rates for this sample size. There were little to no detected differences until the mean difference reached 0.3. In this condition, *t* – tests were only able to detect mean differences when the SDs of both groups were .5 and in normally distributed data. Cliff's Delta performed similarly in this condition and either outperformed or performed similarly to all other statistics in all other conditions. In the mean differences of 0.4 conditions, the non-parametric statistics tended to be more robust. In the skewed conditions, the non-parametric statistics always performed the best with a clear advantage to Cliff's Delta. A similar pattern emerged in the normally distributed data, but Student's *t* out performed all other statistics in the condition with the reference group having a SD of .75 and the focal group having a SD of .5. In the mean difference of 0.5 conditions, when the standard deviations were the same in each group, the *t* -tests either outperformed or performed similarly to the non-parametric statistics. Once again, the non-parametric tests were more robust in that they detected differences across more conditions. When the standard deviations were at their peak differences, the non-parametric statistics tended to be the most powerful except in the skewed condition where the SD of group 1 was 1 and the SD of group 2 was .5. In that condition, Student's *t* was the most powerful.

N = 50

Table 2 shows the power rates for this sample size. There were no detected differences in the mean difference of 0.1 conditions. In the 0.2 conditions, Cliff's Delta tended to be the most powerful or performed similarly to the Mann-Whitney *U*. Both *t* – tests outperformed the Mann-Whitney *U* in normally distributed data, when the SDs of both groups were .5. In the mean difference of 0.3 conditions, the *t* – tests started outperforming the non-parametric statistics in the normally distributed data. Excluding the conditions with peak standard deviation differences, the *t* – tests outperformed or performed similarly to the non-parametric statistics. In the skewed data, the non-parametric statistics were more powerful. This is especially true in the conditions where the standard deviation differences were at their highest. In the skewed conditions, there is arguably not a clear winner between the Mann-Whiney *U* and Cliff's delta because differences were either small or mixed. In the mean difference of 0.4 conditions, for normally distributed data, the *t* – tests always outperformed or performed similarly to the non-parametric statistics. The only condition that did not have perfect detection rates for the *t* – test was the condition with the largest SD for both groups (SD=1 for both groups). Interestingly, in the skewed distribution, this was also the only condition in which the non-parametric statistics performed better than the *t* – test. In the mean difference of 0.5 conditions, there were perfect power rates for all statistics.

N = 100

Table 3 shows the power rates for this sample size. In the mean differences of 0.3 and larger, the *t* -tests either performed similarly or outperformed the non-parametric statistics. Also, in the mean difference of 0.1 conditions, there were no detected differences. In the mean difference of 0.2 conditions, the non-parametric statistics were more robust in that they detected differences in more conditions. In the conditions that included a SD of .5 and did not have a SD of 1, the *t* – tests performed the best or equal to the non-parametric statistics. In all other conditions within the mean difference of 0.2 conditions, the non-parametric statistics performed better with a slight advantage to the Mann-Whitney *U*.

N > 100

In the N=1,000 conditions (Table 4), the *t*-tests always had perfect detection rates. For the non-parametric statistics, they seemed to be less powerful at detecting differences when the mean differences were 0.1 and the SD of group 1 was 1. This is also the first time in which they consistently performed worse in the skewed conditions. For N=10,000, there were perfect detection rates for every condition and for every statistic.

Error

Table 5 shows the error rates for N= 10,000. There were perfect error rates for all other conditions. In the normally distributed data, error rates were generally small. However, the *t* – test had the smallest amount of error. In the non-parametric statistics, Cliff's Delta had a smaller amount of error than the Mann-Whitney *U*. In the skewed conditions, when the *t* – test had errors, the rates were slightly higher than the non-parametric statistics, but overall, the error rates were small. For the condition with equal SDs, the non-parametric statistics had zero error. In all other conditions not already mentioned, the non-parametric statistics had a considerable amount of error with the Mann-Whitney *U* having slightly more. The error rates for all conditions might be explained by the data generation process. Because it is challenging to generate data with an exactly zero mean difference for vary large sample sizes, these statistics might be able to detect even negligible differences. For example, in the condition where the reference group has an SD of 1 and the SD of the reference group is .5, the non-parametric statistics could detect a difference between the distributions with a mean difference of 0.0002. The *t* – tests might not detect a difference because the mean differences are negligible. However, the distributions themselves might have been differently shaped enough for the non-parametric statistics to detect a difference.

Different Ns

In all conditions, there were perfect error rates for every statistic. This is unsurprising given the error rates from the equal Ns described above.

Normally distributed data

The results for this condition are in Table 6. For distributions with a mean difference of 0.1 and 0.2 there are few detected differences and only by the Mann-Whitney *U*. When the mean differences become more pronounced, an interesting pattern emerges. Whenever Student's *t* and the Mann-Whitney *U* are the most powerful, Welch's *t* and Cliff's Delta tend to be considerably weaker. Occasionally there were no detected differences in one of those two statistics, but relatively high detection rates with Student's *t* and the Mann-Whitney *U*. The reverse is also true to an even greater extent. In the mean difference of 0.3 conditions, the standard deviation of the large group was small (.5), then Student's *t* is preferred. If the standard deviation of the larger group is .75 or above, then Cliff's Delta would be preferred slightly over Welch's *t* because it could detect some distributional differences where the Welch test could not. In the mean difference of 0.4 conditions, if the SD of the larger group is .75 or .5, then Student's *t* is the most powerful. If the SD of the larger group is 1, then Cliff's Delta would be preferred. For the mean difference of 0.5 conditions, in general, Student's *t* is the most powerful yielding perfect power rates in most conditions along with having acceptable power rates in the others. The Mann-Whitney *U* also performed well in all conditions except when both groups have a SD of 1.

Skewed data

The results for this condition are in Table 7. There were no differences detected when the mean difference between distributions was 0.1. In the mean difference of 0.2 in the normally distributed data conditions, the Mann-Whiney *U* detected some differences. Cliff's Delta also detected some differences. In the condition where the small group had a SD of .5 and the large group had a SD of 1, it had a perfect detection rate. In the mean difference of 0.3 and larger conditions, there was a similar pattern to the normally distributed data. If the standard deviation of the large group was small, then Student's *t* is preferred. If not, then either Cliff's Delta or the Mann-Whitney *U* performed the best with an arguable slight advantage to Cliff's Delta. In the 0.4 mean difference conditions, if the standard deviations of the large group were not 1, then Student's *t* is preferred. In those conditions where the mean of the large group is 1, Cliff's Delta would be preferred. In the mean difference of 0.5 conditions, the Mann-Whitney *U* seemed to perform the best with Student's *t* as a viable alternative. The only statistic that we would not suggest using is Welch's *t* because when the SD of the small group was 1, it did not perform well.

Discussion

Pairwise comparisons are routine statistical procedures that can supplement decision making. In the AES, this choice can have an impact on which areas are focused on in action planning. Missing an important concept due to having a weaker test could lead to weaker action plans. Finding a difference where there is not one could waste the time of decision makers and employees. This is time that could be spent working with veterans or focusing on the areas of organizational health that need help. The importance of choosing the most accurate pairwise comparison cannot be understated.

To summarize, when samples sizes are $n = 20$ and the samples sizes are equal, Cliff's delta is generally more powerful than the other statistics. In equal sample sizes greater than or equal to 50, the results are less consistent. With a few exceptions, the *t*-tests are more powerful except when one of the groups has an SD of 1. The non-parametric tests tended to be more robust especially when the sample sizes were 50 or less and when the mean differences are smaller than .3. We would give a slight edge to Cliff's Delta over the Mann-Whitney *U* in this regard. This is not surprising as the Mann-Whitney *U* does not detect differences in spread or shape of the distributions, whereas Cliff's Delta does. Also, *t* – tests generally had lower error than the non-parametric statistics. The error rates of the non-parametric statistics were almost always larger, and in some cases, were disturbingly high. However, this could be due to the what the statistics test. Two groups could have almost equal means and due to differences in the SDs, the distributions could be radically different. However, it should be noted that these "errors" only occurred in sample sizes of 10,000 which may not be concerning for many applications.

In the unequal sample size conditions, *t*-tests tended to be substantially more powerful in many of the conditions, but the non-parametric statistics tended to be able to detect distributional differences in more situations. The conditions in which the standard *t* – test was generally more powerful were the conditions in which the larger group had a small SD. The opposite is true for Welch's *t*. The most interesting pattern was the similar pattern between Welch's *t* and Cliff's Delta and between Student's *t* and Mann-Whitney *U*.

Recommendations
Because of the low error rates of each statistic for most comparisons, there is not an inherently bad choice of statistic unless there are equal sample sizes of 10,000. Therefore, the concerns about making a Type I error can be mitigated regardless of the test used.

Due to the purpose of the AES and with the low error, we argue that using the test that is the most powerful is the best choice. To maximize the chance of finding significance so decision makers can action plan around those areas is preferable. In this context, there is not a clear, single "winner". Therefore, we suggest using multiple tests depending on the properties of the data. If the larger group has a SD of .75 or less, then Student's *t* would be preferred. If the large group has a SD of 1, then Cliff's Delta or Mann-Whitney *U* would probably be preferred. In the unlikely event of equal-sample sizes in organizational data, it might be better to use Cliff's Delta unless the sample sizes are 1,000 or larger. In that case, one should consider using a *t* – test.

Limitations
This study has limitations on generalizability like many of the previous studies mentioned. The tested parameters are applicable for most AES data and most comparisons, but these results can only generalize to data with similar characteristics. In addition, these results may not generalize to other AES data points such as composites or those rated on a 7-point scale. Kromrey & Hogarty (1998) found some differences in their results based on the number of points on the rating scale.

Another limitation is how error was conceptualized. In large data sets, it is challenging to generate data with variability and mean differences of exactly 0. In addition, if the mean difference of the groups were exactly zero, the *t* – tests would yield a statistic of 0 (and therefore have perfect error rates). This would not be useful. It might be better to conceptualize error in this study as detecting something that is clearly a negligible difference. In addition, with the non-parametric statistics, they do not explicitly test differences in means except under a few exceptional circumstances. Therefore, the high error rates for the non-parametric statistics might not be error at all, but a reflection in this difference between what is tested. This difference may or may not be meaningful. Future research, theory, and situation should guide if this difference is 1) error and 2) if not error, then if the difference in distributions is meaningful.

Conclusions
This study was a formal test of one data analysis aspect of the AES. We agree in part with Kromrey & Hogarty (1998) in that the choice of pairwise comparison statistic should be driven by the research question. However, depending on the situation, researchers should also choose based on the properties of the data beyond that of ordinal vs. interval/ratio data. We recommend that researchers be aware of these considerations, as well as of properties of the data, and how those aspects of the data relate to their questions of interest.

References

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak Measurements vs. Strong Statistics: An Empirical Critique of SS Stevens' Proscriptions in Statistics. *Educational and Psychological Measurement*, *26*(2), 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*(4), 309-335.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological bulletin*, *57*(1), 49.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494.

Kromrey, J. D., & Hogarty, K. Y. (1998). Analysis options for testing group differences on ordered categorical variables: An empirical investigation of type I error control and statistical power. *Multiple linear regression viewpoints*, *25*(1), 70-82.

Lord, F. M. (1953). On the Statistical Treatment of Football Numbers. *American Psychologist, 8*(12), 750-751.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50-60.

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

Poncet, A., Courvoisier, D. S., Combescure, C., & Perneger, T. V. (2016). Normality and Sample Size Do Not Matter for the Selection of an Appropriate Statistical Test for Two-Group Comparisons. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 12*(2), 61-71.

Reed, J. F. (2005). Contributions to two-sample statistics. *Journal of Applied Statistics*, *32*, 37-44.

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

Table 1.

Power rates of pairwise comparison statistics for N=20.

		Normal				Skewed			
Diff	SD (G1, G2)	t	t_w	Δ	U	t	t_w	Δ	U
0.1	(.5, .5)	0	0	0	0	0	0	0	0
	(.5, .75)	0	0	0	0	0	0	0	0
	(.5, 1)	0	0	0	0	0	0	0	0
	(.75, .5)	0	0	0	0	0	0	0	0
	(.75, .75)	0	0	0	0	0	0	0	0
	(.75, 1)	0	0	0	0	0	0	0	0
	(1, .5)	0	0	0	0	0	0	0	0
	(1, .75)	0	0	0	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0	0
0.2	(.5, .5)	0	0	0	0	0	0	0	0
	(.5, .75)	0	0	0	0	0	0	0.01	0.01
	(.5, 1)	0	0	0	0	0	0	0	0
	(.75, .5)	0	0	0	0	0	0	0	0
	(.75, .75)	0	0	0	0	0	0	0	0
	(.75, 1)	0	0	0	0	0	0	0	0
	(1, .5)	0	0	0	0	0	0	0	0
	(1, .75)	0	0	0	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0	0
0.3	(.5, .5)	0.48	0.48	0.48	0	0	0	0.43	0.23
	(.5, .75)	0	0	0.02	0	0	0	0.11	0.09
	(.5, 1)	0	0	0	0	0	0	0.05	0.04
	(.75, .5)	0	0	0.01	0	0	0	0	0.03
	(.75, .75)	0	0	0	0	0	0	0	0
	(.75, 1)	0	0	0	0	0	0	0	0
	(1, .5)	0	0	0	0	0	0	0	0
	(1, .75)	0	0	0	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0	0
0.4	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	0.32	0	0.83	0.38	0.37	0.1	0.89	0.83
	(.5, 1)	0	0	0.02	0.03	0	0	0.4	0.34
	(.75, .5)	0.45	0.11	0.32	0.12	0.32	0	0.84	0.33
	(.75, .75)	0	0	0.01	0.01	0	0	0.27	0.21
	(.75, 1)	0	0	0	0	0	0	0.12	0.04
	(1, .5)	0	0	0.01	0.01	0	0	0	0
	(1, .75)	0	0	0	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0	0

	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	0.29	0.14	0.41	0.46	0.31	0.23	0.94	0.94
	(.75, .5)	1	1	1	1	1	1	1	1
0.5	(.75, .75)	1	1	0.91	0.46	1	1	0.92	0.47
	(.75, 1)	0	0	0.14	0.06	0	0	0.73	0.42
	(1, .5)	0.41	0.18	0.56	0.39	0.29	0	0.06	0.09
	(1, .75)	0	0	0.09	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0.04	0.02

Note. Diff= Mean difference; SD (G1,G2) = Standard deviations of group 1(G1) and group 2 (G2); t = Student's t-test; t_w = Welch's t-test; Δ = Cliff's Delta; U = Mann & Whitney U; Normal = Normal distribution; Skewed = Skewed distribution.

Table 2.

Power rates of pairwise comparison statistics for N=50.

		Normal				Skewed			
Diff	SD (G1, G2)	t	t_w	Δ	U	t	t_w	Δ	U
0.1	(.5, .5)	0	0	0	0	0	0	0	0
	(.5, .75)	0	0	0	0	0	0	0	0
	(.5, 1)	0	0	0	0	0	0	0	0
	(.75, .5)	0	0	0	0	0	0	0	0
	(.75, .75)	0	0	0	0	0	0	0	0
	(.75, 1)	0	0	0	0	0	0	0	0
	(1, .5)	0	0	0	0	0	0	0	0
	(1, .75)	0	0	0	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0	0
0.2	(.5, .5)	0.71	0.71	0.79	0.5	0.52	0.52	0.75	0.65
	(.5, .75)	0	0	0.01	0.02	0	0	0.62	0.56
	(.5, 1)	0	0	0	0	0	0	0.46	0.45
	(.75, .5)	0	0	0	0	0	0	0.02	0.05
	(.75, .75)	0	0	0	0	0	0	0.08	0.09
	(.75, 1)	0	0	0	0	0	0	0.11	0.11
	(1, .5)	0	0	0	0	0	0	0	0
	(1, .75)	0	0	0	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0	0
0.3	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	0.09	0.07	0.31	0.39	0	0	0.97	0.98
	(.75, .5)	1	1	0.95	0.95	1	1	1	1
	(.75, .75)	0.63	0.63	0.49	0.39	0.69	0.69	0.97	0.93
	(.75, 1)	0	0	0.07	0.04	0	0	0.82	0.75
	(1, .5)	0.04	0.02	0.15	0.17	0.07	0.04	0.11	0.19
	(1, .75)	0	0	0.03	0.02	0	0	0.07	0.09
	(1, 1)	0	0	0	0	0	0	0.01	0
0.4	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	1	1	0.97	1	1	1	1	1
	(.75, .5)	1	1	1	1	1	1	1	1
	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	1	1	0.91	0.85	1	1	1	1
	(1, .5)	1	1	0.95	0.96	1	1	1	1
	(1, .75)	1	1	0.93	0.84	1	1	1	1
	(1, 1)	0.6	0.6	0.51	0.38	0.68	0.68	1	0.98

	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	1	1	1	1	1	1	1	1
	(.75, .5)	1	1	1	1	1	1	1	1
0.5	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	1	1	1	1	1	1	1	1
	(1, .5)	1	1	1	1	1	1	1	1
	(1, .75)	1	1	1	1	1	1	1	1
	(1, 1)	1	1	1	1	1	1	1	1

Note. Diff= Mean difference; SD (G1,G2) = Standard deviations of group 1(G1) and group 2 (G2); t = Student's t-test; t_w = Welch's t-test; Δ = Cliff's Delta; U = Mann & Whitney U; Normal = Normal distribution; Skewed = Skewed distribution.

Table 3.

Power rates of pairwise comparison statistics for N=100.

		Normal				Skewed			
Diff	SD (G1, G2)	t	t_w	Δ	U	t	t_w	Δ	U
0.1	(.5, .5)	0	0	0	0	0	0	0	0
	(.5, .75)	0	0	0	0	0	0	0	0
	(.5, 1)	0	0	0	0	0	0	0	0
	(.75, .5)	0	0	0	0	0	0	0	0
	(.75, .75)	0	0	0	0	0	0	0	0
	(.75, 1)	0	0	0	0	0	0	0	0
	(1, .5)	0	0	0	0	0	0	0	0
	(1, .75)	0	0	0	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0	0
0.2	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	0.95	0.95	1	1	1	1
	(.5, 1)	0.02	0.01	0.17	0.31	0	0	1	1
	(.75, .5)	1	1	0.83	0.83	1	1	0.98	0.98
	(.75, .75)	0.2	0.2	0.3	0.28	0.22	0.22	0.96	0.96
	(.75, 1)	0	0	0.05	0.05	0	0	0.98	0.98
	(1, .5)	0	0	0.13	0.19	0	0	0.03	0.09
	(1, .75)	0	0	0.01	0.01	0	0	0.04	0.07
	(1, 1)	0	0	0	0	0	0	0.11	0.12
0.3	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	1	1	1	1	1	1	1	1
	(.75, .5)	1	1	1	1	1	1	1	1
	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	1	1	1	1	1	1	1	1
	(1, .5)	1	1	0.96	0.98	1	1	1	1
	(1, .75)	1	1	0.95	0.95	1	1	0.99	0.99
	(1, 1)	0.98	0.98	0.82	0.74	0.99	0.99	1	0.99
0.4	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	1	1	1	1	1	1	1	1
	(.75, .5)	1	1	1	1	1	1	1	1
	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	1	1	1	1	1	1	1	1
	(1, .5)	1	1	1	1	1	1	1	1
	(1, .75)	1	1	1	1	1	1	1	1
	(1, 1)	1	1	1	1	1	1	1	1

	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	1	1	1	1	1	1	1	1
	(.75, .5)	1	1	1	1	1	1	1	1
0.5	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	1	1	1	1	1	1	1	1
	(1, .5)	1	1	1	1	1	1	1	1
	(1, .75)	1	1	1	1	1	1	1	1
	(1, 1)	1	1	1	1	1	1	1	1

Note. Diff= Mean difference; SD (G1,G2) = Standard deviations of group 1(G1) and group 2 (G2); t = Student's t-test; t_w = Welch's t-test; Δ = Cliff's Delta; U = Mann & Whitney U; Normal = Normal distribution; Skewed = Skewed distribution.

Table 4.

Power rates of pairwise comparison statistics for N=1,000.

		Normal				Skewed			
Diff	SD (G1, G2)	t	t_w	Δ	U	t	t_w	Δ	U
0.1	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	1	1	1	1	1	1	1	1
	(.75, .5)	1	1	1	1	1	1	1	1
	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	1	1	1	1	1	1	1	1
	(1, .5)	1	1	0.96	0.97	1	1	0.8	0.87
	(1, .75)	1	1	0.91	0.91	1	1	0.6	0.63
	(1, 1)	1	1	0.9	0.9	1	1	0.92	0.92
0.2	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	1	1	1	1	1	1	1	1
	(.75, .5)	1	1	1	1	1	1	1	1
	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	1	1	1	1	1	1	1	1
	(1, .5)	1	1	1	1	1	1	1	1
	(1, .75)	1	1	1	1	1	1	1	1
	(1, 1)	1	1	1	1	1	1	1	1
0.3	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	1	1	1	1	1	1	1	1
	(.75, .5)	1	1	1	1	1	1	1	1
	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	1	1	1	1	1	1	1	1
	(1, .5)	1	1	1	1	1	1	1	1
	(1, .75)	1	1	1	1	1	1	1	1
	(1, 1)	1	1	1	1	1	1	1	1
0.4	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	1	1	1	1	1	1	1	1
	(.75, .5)	1	1	1	1	1	1	1	1
	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	1	1	1	1	1	1	1	1
	(1, .5)	1	1	1	1	1	1	1	1
	(1, .75)	1	1	1	1	1	1	1	1
	(1, 1)	1	1	1	1	1	1	1	1

	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	1	1	1	1	1	1	1	1
	(.75, .5)	1	1	1	1	1	1	1	1
0.5	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	1	1	1	1	1	1	1	1
	(1, .5)	1	1	1	1	1	1	1	1
	(1, .75)	1	1	1	1	1	1	1	1
	(1, 1)	1	1	1	1	1	1	1	1

Note. Diff= Mean difference; SD (G1,G2) = Standard deviations of group 1(G1) and group 2 (G2); t = Student's t-test; t_w = Welch's t-test; Δ = Cliff's Delta; U = Mann & Whitney U; Normal = Normal distribution; Skewed = Skewed distribution.

Table 5.

Error rates of pairwise comparison statistics for N=10,000.

		Normal				Skewed			
Diff	SD (G1, G2)	t	t_w	Δ	U	t	t_w	Δ	U
	(.5, .5)	0.01	0.01	0.02	0.02	0.01	0.01	0	0
	(.5, .75)	0.01	0.01	0.03	0.03	0.03	0.03	0.02	0.02
	(.5, 1)	0	0	0.02	0.04	0	0	0.57	0.58
	(.75, .5)	0.02	0.02	0	0.02	0.05	0.05	0.03	0.03
0 - .02	(.75, .75)	0	0	0	0	0	0	0	0
	(.75, 1)	0	0	0	0	0	0	0.51	0.51
	(1, .5)	0	0	0.02	0.05	0	0	0.56	0.6
	(1, .75)	0	0	0	0	0	0	0.48	0.51
	(1, 1)	0	0	0	0	0	0	0	0

Note. Diff= Mean difference; SD (G1,G2) = Standard deviations of group 1(G1) and group 2 (G2); t = Student's t-test; t_w = Welch's t-test; Δ = Cliff's Delta; U = Mann & Whitney U; Normal = Normal distribution; Skewed = Skewed distribution. All statistics had perfect power ratings for all other conditions.

Table 6.

Power rates of pairwise comparison statistics for different Ns normally distributed data.

		N (G1=17, G2=742)				N (G1=742, G2=17)			
Diff	SD (G1, G2)	t	t_w	Δ	U	t	t_w	Δ	U
	(.5, .5)	0	0	0	0	0	0	0	0
	(.5, .75)	0	0	0	0	0	0	0	0
	(.5, 1)	0	0	0	0	0	0	0	0
	(.75, .5)	0	0	0	0	0	0	0	0
0.1	(.75, .75)	0	0	0	0	0	0	0	0
	(.75, 1)	0	0	0	0	0	0	0	0
	(1, .5)	0	0	0	0	0	0	0	0
	(1, .75)	0	0	0	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0	0
	(.5, .5)	0	0	0	0	0	0	0	0
	(.5, .75)	0	0	0	0	0	0	0	0.05
	(.5, 1)	0	0	0	0	0	0	0	0.03
	(.75, .5)	0	0	0	0	0	0	0	0
0.2	(.75, .75)	0	0	0	0	0	0	0	0
	(.75, 1)	0	0	0	0	0	0	0	0
	(1, .5)	0	0	0	0.01	0	0	0	0
	(1, .75)	0	0	0	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0	0
	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	0	1	1	0	1	0	0.06	0.82
	(.5, 1)	0	1	1	0	1	0	0.01	0.44
	(.75, .5)	1	0	0.05	0.34	0	1	1	0
0.3	(.75, .75)	0	0	0.04	0.01	0	0	0.04	0.02
	(.75, 1)	0	0	0.07	0	0	0	0	0.02
	(1, .5)	1	0	0	0.09	0	1	1	0
	(1, .75)	0	0	0	0	0	0	0.01	0
	(1, 1)	0	0	0	0	0	0	0	0
	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	0.56	1	1
	(.5, 1)	0	1	1	0.47	1	0	0.21	0.94
	(.75, .5)	1	0.64	0.76	1	1	1	1	1
0.4	(.75, .75)	1	0.62	0.74	0.93	1	0.49	0.98	0.85
	(.75, 1)	0	0.72	0.98	0.05	1	0	0.03	0.27
	(1, .5)	1	0	0.03	0.76	0	1	1	0.06
	(1, .75)	1	0	0.04	0.14	0	0.52	0.93	0.02
	(1, 1)	0	0	0	0	0	0	0	0

	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	0.98	1	1	1	1	0	0.16	1
	(.75, .5)	1	1	1	1	1	1	1	1
0.5	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	0.99	1	1	1	1	0	0.22	0.96
	(1, .5)	1	0	0.66	1	0.93	1	1	1
	(1, .75)	1	0.11	0.62	0.85	0.92	1	1	0.97
	(1, 1)	0.95	0.11	0.91	0.67	0.89	0	0.29	0.35

Note. Diff= Mean difference; SD (G1,G2) = Standard deviations of group 1(G1) and group 2 (G2); t = Student's t-test; t_w = Welch's t-test; Δ = Cliff's Delta; U = Mann & Whitney U; Normal = Normal distribution; Skewed = Skewed distribution.

Table 7.

Power rates of pairwise comparison statistics for different Ns skewed data.

		N (G1=17, G2=742)				N (G1=742, G2=17)			
Diff	SD (G1, G2)	t	t_w	Δ	U	t	t_w	Δ	U
	(.5, .5)	0	0	0	0	0	0	0	0
	(.5, .75)	0	0	0	0	0	0	0	0
	(.5, 1)	0	0	0	0	0	0	0	0
	(.75, .5)	0	0	0	0	0	0	0	0
0.1	(.75, .75)	0	0	0	0	0	0	0	0
	(.75, 1)	0	0	0	0	0	0	0	0
	(1, .5)	0	0	0	0	0	0	0	0
	(1, .75)	0	0	0	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0	0
	(.5, .5)	0	0	0	0.03	0	0	0	0
	(.5, .75)	0	0	0.24	0	0	0	0.12	0.12
	(.5, 1)	0	0	1	0	0	0	0	0.15
	(.75, .5)	0	0	0	0.09	0	0	0	0
0.2	(.75, .75)	0	0	0	0	0	0	0.12	0
	(.75, 1)	0	0	0	0	0	0	0	0
	(1, .5)	0	0	0	0	0	0	0	0
	(1, .75)	0	0	0	0	0	0	0	0
	(1, 1)	0	0	0	0	0	0	0	0
	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	0	1	1	1	1	0	0.57	0.84
	(.5, 1)	0	1	1	0.29	1	0	0.17	0.61
	(.75, .5)	1	0	0.05	1	0	1	1	0.19
0.3	(.75, .75)	0	0	0.33	0.61	0	0	0.14	0.03
	(.75, 1)	0	0	0.67	0.03	0	0	0.09	0.19
	(1, .5)	1	0	0	0.23	0	1	0.9	0
	(1, .75)	0	0	0	0	0	0	0.02	0
	(1, 1)	0	0	0	0	0	0	0	0
	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	0.72	0.99	1
	(.5, 1)	0	1	1	1	1	0	0.42	0.98
	(.75, .5)	1	0.66	0.98	1	1	1	1	1
0.4	(.75, .75)	1	0.59	1	1	1	0.76	0.97	0.97
	(.75, 1)	0	0.51	1	0.98	1	0	0.31	0.82
	(1, .5)	1	0	0	0.94	0	1	1	0
	(1, .75)	1	0	0	0.83	0	0.66	0.96	0.03
	(1, 1)	0	0	0.09	0.01	0	0	0.19	0.07

	(.5, .5)	1	1	1	1	1	1	1	1
	(.5, .75)	1	1	1	1	1	1	1	1
	(.5, 1)	0.91	1	1	1	1	0.03	0.89	1
	(.75, .5)	1	1	1	1	1	1	1	1
0.5	(.75, .75)	1	1	1	1	1	1	1	1
	(.75, 1)	0.84	1	1	1	1	0.07	0.9	1
	(1, .5)	1	0	0.2	1	0.99	1	1	1
	(1, .75)	1	0	0.28	0.99	0.96	1	1	0.95
	(1, 1)	0.82	0	1	1	0.96	0.14	0.88	0.94

Note. Diff= Mean difference; SD (G1,G2) = Standard deviations of group 1(G1) and group 2 (G2); t = Student's t-test; t_w = Welch's t-test; Δ = Cliff's Delta; U = Mann & Whitney U; Normal = Normal distribution; Skewed = Skewed distribution.