

**Appropriate statistics for ordinal level data :  
Should we really be using t-test and Cohen's d  
for evaluating group differences  
on the NSSE and other surveys?**

**Jeanine Romano**

**University of Tampa**

**Jeffrey D. Kromrey**

**University of South Florida**

**Jesse Coraggio**

**University of South Florida**

**Jeff Skowronek**

**University of Tampa**

**Paper presented at the annual meeting of the Florida Association of Institutional Research,  
February 1 -3, 2006, Cocoa Beach, Florida**

## Abstract

Support for data driven decision-making is a large part of institutional research. Institutional comparisons on many surveys, such as the *National Survey of Student Engagement* (NSSE), are analyzed with statistics such as *t*-tests and Cohen's *d* to evaluate differences between group responses. It is important to remember that these methods were developed to evaluate mean differences on variables with interval-level or ratio-level measurement, under assumptions such as population normality and homogeneous variances. Statistical tests such as the *t*-test are used to evaluate the statistical significance of differences between two groups on an item or set of items. Cohen's *d* effect sizes are used to describe the magnitude of the group difference in standard deviation units. It has been argued that Cohen's *d* represents *practical significance* because unlike *t*-tests, it is independent of sample size.

Responses to many survey questions yield discrete ordinal-level data, such that the response values have an inherent order (e.g. strongly agree to strongly disagree) and many responses are identical (e.g., 70% of respondents strongly agree). Even though we assign numerical values to the responses for such items it is important to remember that one should not assume that the difference between strongly agree and agree is the same as the difference between agree and disagree. Because of this it can be argued that there are more appropriate methods for evaluating ordinal data that explore the number of times one group answers higher than the other.

This presentation will explore the issues surrounding the use of *t*-tests and Cohen's *d* and demonstrate alternative methods for evaluating ordinal data such as simple raw differences, odds ratios, and Cliff's Delta (an effect size for ordinal data) by using data from a comparison of a small private master's level university to other Master's level universities, based on the 2005 NSSE results.

## *Introduction*

The external demand to demonstrate evidence of learning and development is greater than ever in the world of institutional research in higher education. One way that this evidence is collected is through the use of student surveys. Specifically the *National Survey of Student Engagement*, (NSSE), is administered to first year students and seniors at many colleges and universities every spring in an attempt to measure the extent to which students are engaged in activities at their institution related to their educational and personal development.

When surveys such as the NSSE are administered to individuals at an institution, the results of these measures have a strong impact on decisions that are made in regards to improving programs at the university. Usually the institution's scores are compared to a peer group of schools such that the "Selected Peers" share similar characteristics to the institution. The NSSE also provides comparisons for the entire population for each class level that took the survey and comparisons to other types of schools in the same classification as determined by Carnegie Commission on Higher Education. We used a small private institution's data, the data from the selected peers group and the data from the other institutions with the same Carnegie classification (Masters) to explore and demonstrate the different statistical indices used in this study.

There are many statistical methods that can be used to describe or make inferences about these data. Statistical methods such as  $t$ -tests and Cohen's  $d$  are usually used to evaluate the differences between group responses. In general, when the  $t$ -test is conducted, if the  $p$  value for the  $t$ -test is less than .05 it is acceptable to state that there is a statically significant difference between the two groups in question. In addition, some surveys, such as the NSSE, will also report Cohen's  $d$  effect size. The NSSE manual indicates that this index is a measure of practical significance, because unlike  $t$ -tests, it is independent of sample size (NSSE Institutional Report, 2005). It is important to remember that these methods were developed to evaluate mean differences on variables with interval-level or ratio-level measurement, under assumptions such as population normality and homogeneous variances.

Most survey questions are discrete ordinal level data such that the items have an inherent order, e.g. strongly agree to strongly disagree. Even though we assign numerical values to the responses for such items, one should not assume that the difference between strongly agree and agree is the same as the difference between agree and disagree. Because of this it can be argued that there are more appropriate methods for evaluating discrete ordinal data that explore the number of times one group answers higher than the other.

This presentation will explore the issues surrounding the use of the  $t$ -test and Cohen's  $d$  and demonstrate alternative methods for evaluating ordinal data such simple raw differences,

odds ratios, and Cliff's Delta (an effect size for ordinal data) by using data from the 2005 NSSE results where a small private master's level university is compared to other Master's level universities that administered the NSSE.

### *Levels of Measurement*

Stevens (1946) proposed that there are four levels that classify the nature of information contained within numbers assigned to the variables of interest in research. According to Stevens, different statistical analyses on variables are appropriate based on the level at which a variable is measured. The four levels of measurement that were proposed by Stevens are nominal, ordinal, interval, and ratio. At the nominal level the variables are categorical and there is no order. For example, religious affiliation would be considered a nominal level variable. The only type of central tendency that can be calculated is the mode. It is possible to evaluate the variation qualitatively, but a standard deviation can not be calculated. At the ordinal level of measurement, the data are ranked such that there is an order to the data but there is no definite interval. The Motion Picture Association of America's movie rating system is an example of ordinal data. Like nominal level data, central tendency can be evaluated using a mode. In addition, since there is an order, the median or 50<sup>th</sup> percentile can be calculated. The third level of Steven's classification is the interval level. Interval data, like ordinal, have a definite rank order. In addition, there is a definite interval between the variable's values and the values can be added and subtracted in a meaningful way. However, there is no inherent zero in an interval scale. For example, Celsius temperature readings provide an interval level variable because zero degrees does not mean an absence of heat but the difference between twenty degrees and thirty degrees is the same value as the difference between thirty degrees and forty degrees, i.e. the values have a definite interval between them. At this level a mode, median, and mean can be calculated to describe central tendency. Finally, ratio is the forth level. Ratio level data are similar to interval level data in the sense that there is a definite interval between the observations but there are meaningful ratios between pairs of numbers. Unlike interval data, there is an inherit zero. For example age is a ratio level variable. Like interval data, the mode, median, and mean can be calculated to describe central tendency.

In the case of survey data, while many of the items are discrete ordinal data by nature it is very common to treat them as interval or ratio and calculate mean differences. While there has been some criticism that the levels are too limited and not necessarily the best way to evaluate data (Velleman & Wilkinson, 1993), there are other assumptions that need be considered before when applying statistics such as *t*-tests and Cohen's *d*.

### *Use of t-tests with NSSE data*

Most NSSE analyses compare the difference between two group means on either individual items or “benchmark” scores of the scalelets. The typical statistic used for these comparisons is the independent samples (or means) *t*-test. The independent samples *t*-test is a parametric test used to determine whether or not two independent sample means are significantly different from one another (where the null hypothesis is  $H_0: \mu_1 = \mu_2$ ). The statistic is calculated as follows:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

where  $\bar{X}_1$  and  $\bar{X}_2$  represent the sample means and  $s_{\bar{x}_1 - \bar{x}_2}$  represents the standard error of the difference between the sample means. This error term is calculated in part by using a weighted average of the variance for each sample, the pooled variance, which is the quotient of an aggregate of each sample’s sum of squares divided by an aggregate of each sample’s degrees of freedom. The *t*-value obtained from this test is compared to the critical values of the sampling *t* distribution, where the degrees of freedom equal  $n_1 + n_2 - 2$ . If the *t*-value obtained, when  $\alpha = .05$  (or sometimes .01), is larger than a specific *t*-value in the distribution for a certain number of degrees of freedom then the test is said to be “significant.” This significance is interpreted to mean there is a difference between the two sample means, which is larger than a difference reasonably attributable to sampling error and more likely was the result of some independent variable or manipulation.

There are a number of assumptions underlying the use of the *t* distribution as the foundation for the comparison of two sample means. Although the *t*-test is fairly robust, these assumptions need to be met for a statistically valid interpretation of a *t*-test result. First, because the independent samples *t*-test is a parametric test, the data for a participant is assumed to be numerical scores on an interval or ratio scale that can be manipulated with basic arithmetic, such as addition or subtraction (Gravetter & Wallnau, 2004). Among other characteristics, the data for both interval and ratio measurements should have an equal distance and equal difference in magnitude between each number. This is not the case for the NSSE data. Even though the responses given by a participant are converted to an interval value for analysis purposes, this is actually just assigning a numerical value meant to represent interval scale data. Answers provided for many of the NSSE questions are ordinal scale data, not interval or ratio scale data. The response choices, such as “Never,” “Sometimes,” “Often,” and “Very Often,” do not share an equal distance or equal difference in magnitude between each other, but rather differ only by a

subjective rank. For example, the difference between “Never” and “Sometimes” represents a change from complete absence to minimal presence. However, the difference between “Sometimes” and “Often” cannot really be qualified in similar terms, this is also true when the responses are assigned numerical values for interval scale data (as scores of 1, 2, 3, and 4). Additionally, because many of the response sets for NSSE are ordinal, they are discrete, not continuous, data. However, when the scores are converted to interval scale data there are often mean values that fall between the assigned intervals, such as 2.5 or 1.7. When the response set is similar to the one described above, these means actually have minimal value in regards to the interpretation of the data. The use of means for these response sets is similar to calculating a mean for a variable that represents gender, often coded as “1-female” and “2-male” for comparison purposes. A value of 1.7 for gender is meaningless because there is no .7 of a gender, just as there is no .5, .7, etc. for the response set of most NSSE items.

The second assumption is that the samples should be independent; in other words, each measurement is not influenced by any other prior measurement. For the most part, this assumption is usually met through the use of a random sample (a random sample is also important for the generalizability of the obtained results to the population from which the sample was drawn). However, if the samples are not independent then there may be an inflation of the Type I error rate; suggesting there is a significant difference between samples when in fact there is none (Myers & Well, 1995). Third, it is assumed that the sampling distribution is normal; however, the *t*-test is robust and can withstand this violation fairly well. This is especially true when there is a large sample size or an equal *n* in each group (Myers & Well, 1995; Pagano, 2001).

Lastly, and possibly the most important assumption, is that of homogeneity of variance ( $\sigma_1^2 = \sigma_2^2$ ). Generally, the independent variable should affect the means of the population but not the standard deviations. Because variance can be calculated by taking the square of the standard deviation, homogeneity of variance assumes that the population variances are also equal (Pagano, 2001). Homogeneity is important because the error term of the *t*-statistic is an average of the two group variances. If this assumption is violated, there is little value in pooling the variances and the interpretation of any *t*-test results can be misleading or completely uninformative. The following example illustrates the problem with heterogeneous variances (taken from an actual *t*-test):

<i>t</i> -value	df	<i>p</i> -value
2.034	50	.047

On the surface, the results of this analysis would lead a researcher to conclude there is a significant difference between the means of the two samples under investigation. Without any other analyses to consider, this interpretation would appear to be accurate. However, when a test for equal variances is analyzed, it becomes clear that some adjustments need to be made. There are a number of analyses that can be used to determine if the homogeneity assumption has been violated, including Hartley's  $F$ -max test or Levene's equality test (illustrated below). The results of the test for homogeneity for this particular  $t$  test was:

$F$ value	$p$ -value
4.563	.038

The test for equality measures whether or not the two samples have significantly different variances. If the results of the  $F$  test are significant then the sample variances suggest that the homogeneity of variance assumption has been violated. If the assumption is violated, as it appears to be in this analysis, an adjustment must be made to the degrees of freedom corresponding to the  $t$ -test. With the adjustment made to the degrees of freedom, the results of the  $t$ -test, for the same data as above, were the following:

$t$ -value	Adjusted df	$p$ -value
1.815	27.326	.080

After accounting for heterogeneous variances, a researcher would now conclude there was no significant difference between the means of the two groups under investigation. In this case, prior to the adjustment for heterogeneous variances, the researcher's risk of committing a Type I error was greater than the nominal alpha level of .05. Under conditions in which the sample sizes are relatively close there is rarely a problem with Type I error rate inflation (Myers & Well, 1995). Issues with homogeneity and Type I error control become more of a problem when the ratio between sample variances is large or the  $n$  for each sample is markedly different (Myers & Well, 1995), as is often the case with NSSE data comparisons.

In addition to these assumptions, there are a number of factors that contribute to the results of  $t$ -tests. First, and probably most obvious, is the size of the difference between the sample means ( $\bar{X}_1 - \bar{X}_2$ ). The larger the mean difference between the samples the more likely the  $t$ -test will result in significance. Second, the amount of sample variance influences the results of a  $t$ -test. The sample variances are incorporated in the error term, which is the denominator of the  $t$ -test; therefore, the smaller the sample variances the larger the obtained  $t$ -value will be

(Gravetter & Wallnau, 2004). Lastly, the larger the sample size the easier it will be to obtain a “statistically significant” result. When there is a large sample  $n$  the mean difference between the two samples becomes less of a factor. In fact, with extremely large samples, such as those with the NSSE data, very small differences between sample means may result in a “statistically significant” result. As a result of this last factor, a “statistically significant” result may be just that, statistically significant, but have little practical value or meaning. This has led to the use of effect size statistics, such as Cohen’s  $d$ , which are not affected by sample size, to provide a more meaningful interpretation of the results of a  $t$ -test.

*Use of chi-square with NSSE data*

In addition to  $t$ -test, the chi-square test of significance is another method for comparing the differences in two groups. chi--square is a more appropriate analysis when using categorical or qualitative data as in the case of the NSSE Likert scale items. Unlike the  $t$ -test, the chi-square is a non-parametric test. Rather than comparing the differences in sample means, the chi-square test is used to determine whether the two sample proportions are “significantly” different from one another.

The null hypothesis when using a chi-square test is  $H_0: \pi_1 = \pi_2$ , where  $\pi$  represents a population proportion. The frequency data are generally displayed in a contingency table format as shown in Table 1.

*Table 1*

	Agree	Disagree	Total
Group 1	80	20	100
Group 2	50	50	100
Total	130	70	200

The chi-square statistics is calculated with the following formula:

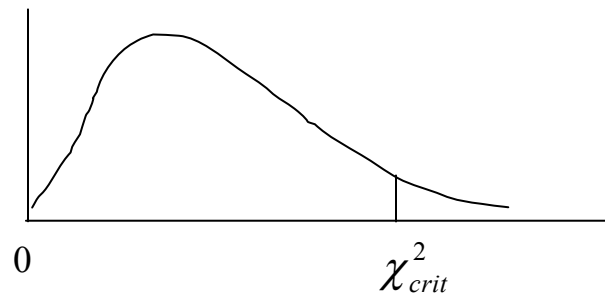
$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

where  $O_i$  represents the observed frequency,  $E_i$  represents the expected frequency, and  $k$  represents the number of cells in the contingency table. In the case of comparing two or more groups  $E$  is represented by the following equation:

$$E = \frac{\text{row total} * \text{column total}}{\text{total}}$$

The chi-square value obtained from the equation is then compared to the critical value of the sampling  $\chi^2$  distribution, where the degrees of freedom ( $df$ ) is equal to  $(r - 1)(c - 1)$ , where  $r$  and  $c$  are the number of rows and number of columns in the contingency table. The chi-square sampling distribution is shown in Figure 1.

Figure 1



As mentioned earlier in the discussion of  $t$ -test, the most common level of Type I error control is an alpha level ( $\alpha$ ) of 0.05. When the obtained chi-square is larger than the critical value ( $\alpha=0.05$ ), the results are “statistically” significant. That is to say that the observed difference in proportions would not be expected in the two samples if the two samples were from the same or identical populations.

As with all test of significance there are some underlying assumptions that must be met in order to ensure accurate results of the analysis. In the case of the chi-square test, most of these assumptions include issues around the assumption of independence. The sample must be drawn randomly. Any systematic sampling may invalidate the results of the analysis. The measured variables must be independent. One person’s response to a survey item must not be influenced by another person’s response to that same item. The categories used the analysis must also be mutually exclusive and when possible exhaustive. The population from which the sample data are drawn should be normally distributed. Finally, the table frequencies should not be too small. One general rule of thumb is that there must be a minimum of five observations expected in each cell of the table.

Response categories are sometimes combined in the analysis of contingency table data. For example, responses of ‘Agree’ and ‘Strongly Agree’ may be combined to allow consideration of the proportion of respondents who agree with an item without differentiation on strength of agreement. For this analysis of the NSSE data, the individual item data were analyzed in both the original response categories and artificially categorized into two groups. This constraint was implemented for demonstrative purposes. In cases where there were four possible options in the

Likert scale (e.g., strongly agree, agree, disagree, and strongly disagree), the data were collapsed into level of agreement or endorsement of the item.

#### *Use of Odds Ratio with NSSE data*

While the chi-square test provides information about the statistical difference in proportions, the odds ratio uses such proportions to provide information about the strength of the relationship between the two variables of interest. As with the chi-square, the odds ratio is designed for use with categorical or qualitative data.

Before discussing the odds ratio, it is important to understand how to calculate odds. Odds are calculated by taking the number or probability of the event ( $p$ ) and dividing by the number or the probability of the non event ( $q$ ). For example, to calculate the odds of a pregnant woman having a boy, one would divide the number of boy births (51 male births in 100) by the number of non-boy births (49 female in 100). So the odds of having a boy would be 51/49 or 1.04. If the odds are greater than 1, than the event is more likely to occur. The odds are not the same as the probability of occurring. If the odds =1, then the probability of the event occurring is the same as the probability of the event not occurring. Odds ratios are essentially the ratio of two independent events.

Odds ratios are calculated dividing the odds in one group (group of interest) by the odds the other group (control group). Using the gender at birth example above, we can compare the odds of a male birth in two different environments. In country 1, the probability of a male birth is .51. So the odds are .51/.49 or 1.04 as stated above. In country 2, we find that the probability of having a male birth is .55. So the odds are .55/.45 or 1.22. So the odds ratio is the odds in country 1 divided by the odds in country 2 or 1.04/1.22 = .852.

The odds ratio must be a non-negative number and can be interpreted by its relationship to a value of one. If the odds ratio is exactly one, than the two groups have the same probability of the event occurring. If the odds ratio is greater than one, than the group of interest has a higher probability of the event occurring. If the odds ratio is less than one, then the comparison group has a higher probability of the event occurring. In our gender at birth example the resulting odds ratio was .852. This value is less than one and the odds of having a boy in country 2 is higher than in country 1.

The odds ratio, while evaluating the relationship in the sample, can also be used as a test of significance to evaluate the differences in the population. This can be calculated using the following formula:

$$\log(\hat{\theta}) \pm z_{1-\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}$$

Where  $\hat{\theta}$  represents the sample log odds ratio,  $\alpha$  = the desired alpha level or Type I control (usually 0.05) and  $n_{xy}$  equals the frequency of observations in each cell as shown in Table 2.

Figure 2

		Response		
		Agree	Disagree	
Group	Group 1	$n_{11}$	$n_{10}$	$n_1$
	Group 2	$n_{01}$	$n_{00}$	$n_0$
		$m_1$	$m_0$	$n$

The created confidence interval can be evaluated to see if it contains a value of one. If it does contain a value of one, than the two proportions while different in the sample (odds ratio) may actually come from the same or equivalent populations indicating no “statistical” differences.

### Effect Sizes

In recent years, a concerted effort has aimed toward encouraging researchers to provide some indication of effect size along with or in place of the results of statistical significance tests. This effort has coincided with renewed debates regarding the over-reliance on hypothesis testing, emphasizing the often misleading nature and inappropriate use of such tests (Nickerson, 2000). Effect sizes provide indices of the extent to which the sample data deviate from the null hypothesis with the impact of sample size removed (Wilkinson & APA Task Force on Statistical Inference, 1999). Effect sizes have been viewed as consistent with null hypothesis significance testing and as an important compliment. Yet, despite urgings for the consistent reporting of effect size, these measures are seldom found in published reports, and are seemingly still far from becoming standard practice (Kirk, 1996, Thompson & Snyder, 1997, 1998). In his critique of statistical hypothesis testing, Carver (1993) pointed out that such tests tell researchers nothing about the size of effects or the size of sampling errors.

One practical impediment to the use and reporting of effect sizes may be the variety of indices that have been proposed. Choosing among the various possible effect-size estimates is not always apparent (Rosenthal, 1991), and opinions vary regarding the merits of the variety of effect sizes available (Crow, 1991; Gorsuch, 1991; McGraw, 1991; Parker, 1995; Rosenthal, 1991; Strahan, 1991). An important consideration is the extent to which indices of effect size calculated from a sample provide information about the magnitude of effect in the population from which the sample was drawn. That is, the potential for statistical bias associated with sample effect size indices should be taken into account in developing accurate interpretations of observed effect sizes. Further, the valid interpretation of sample effect sizes must include a consideration of

the sensitivity of effect size indices to differences in population distribution shape or differences in population variances.

Traditional measures of effect size (Cohen's  $d$  or Hedges'  $g$ ) may be used to describe differences in means relative to an assumed common variance. Cohen's  $d$  is given by

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}}$$

where  $\hat{\sigma}$  is a pooled estimate of the common population standard deviation. Note that this denominator differs from that of the  $t$ -test, in which the difference in sample means is divided by the standard error rather than the standard deviation. This effect size index represents the difference between sample means in standard deviation units. That is, an effect size of 1.0 simply indicates that the sample means are one standard deviation apart.

Hedges and Olkin (1985) suggested that the  $d$  index evidences a small sample bias, and provided an adjusted effect size estimate,  $g$ , designed to reduce such bias:

$$g = d \left( 1 - \frac{3}{4N - 9} \right)$$

where  $N$  is the total sample size (i.e.,  $N = n_1 + n_2$ ).

Problems with both of these effect size indices arise when the samples are drawn from populations that are non-normal or heterogeneous in variances (Fern & Monroe, 1996; Hogarty & Kromrey, 2001; Kraemer & Andrews, 1982; Vargha & Delaney, 2000; Wilcox & Muska, 1999). As alternatives to Cohen's  $d$  or Hedges'  $g$ , robust estimators of location and scale (such as trimmed means and Winsorized variances) have been recommended for computing effect size indices that are analogous to the standardized mean difference (Yuen, 1974; Hedges & Olkin, 1985). In addition, several authors have suggested the use of non-parametric indices of effect size (see, for example, Kraemer & Andrews, 1982; McGraw & Wong, 1992).

A simple non-parametric index, the delta statistic, was proposed by Cliff (1993, 1996) for testing null hypotheses about group differences on ordinal level measurements. The delta statistic is used to test equivalence of probabilities of scores in one group being larger than scores in the other (the property that Cliff (1993) referred to as "dominance"). A sample estimate of the parameter is obtained by enumerating the number of occurrences of an observation from one group having a higher response value than an observation from the second group, and the number of occurrences of the reverse. This gives the sample statistic

$$\hat{\delta} = \frac{\#(x_{i1} > x_{j2}) - \#(x_{i1} < x_{j2})}{n_1 n_2}$$

where  $\#x_{i1} > x_{i2}$  is simply the number of comparisons between observations in the two groups for which the Group 1 observation is larger than the Group 2 observation.

This statistic is most easily conceptualized by considering the data in an arrangement called a dominance matrix. This  $n_1$  by  $n_2$  matrix has elements taking the value of 1 if the row response is larger than the column response, -1 if the row response is less than the column response, and 0 if the two responses are identical. The sample value of Cliff's delta is simply the average value of the elements in the dominance matrix.

Consider a hypothetical example in which the data displayed in Table 2 represent two sets of responses to a single item, with respondent ratings provided on a 5-point scale. For this example, responses were obtained from ten students in program 1 and six students in program 2. The research question seeks to address whether the two populations sampled were different with regard to their response on this item.

Table 2

Sample of Two Groups of Responses to a Single Item.

Group 1	Group 2
1	1
1	2
2	3
2	4
2	4
3	5
3	
3	
4	
5	

Table 3 exhibits these data in a 10 x 6 dominance matrix. The elements of the matrix take the value of 1 if the row (Group 1) value is larger than the column (Group 2) value. The value 0 is assigned if the value for the two groups is the same and the value -1 is given if the row value is less than the column value. These data result in a value for Cliff's delta of -0.25.

Table 3

Dominance Matrix.

	1	2	3	4	4	5	<i>d<sub>i</sub></i>
1	0	-1	-1	-1	-1	-1	-0.833
1	0	-1	-1	-1	-1	-1	-0.833
2	1	0	-1	-1	-1	-1	-0.500
2	1	0	-1	-1	-1	-1	-0.500
2	1	0	-1	-1	-1	-1	-0.500
3	1	1	0	-1	-1	-1	-0.167
3	1	1	0	-1	-1	-1	-0.167
3	1	1	0	-1	-1	-1	-0.167
4	1	1	1	0	0	-1	0.333
5	1	1	1	1	1	0	0.833
<i>d<sub>j</sub></i>	0.8	0.3	-0.3	-0.7	-0.7	-0.9	-0.250

Although not originally intended by Cliff, the delta index provides a useful representation of effect size. When used as an effect size index, Cliff's delta represents the degree of overlap between the two distributions of scores. It ranges from  $-1$  (if all observations in Group 1 are larger than all observations in Group 2) to  $+1$  (if all observations in Group 1 are smaller than all observations in Group 2) and takes the value of zero if the two distributions are identical.

Cohen (1992) provided a useful introduction to the interpretation of effect sizes, especially his early (1969) work on the delineation of small, medium, and large effects. Cohen (1992) noted that a medium effect size represents "an effect likely to be visible to the naked eye of a careful observer," while a small effect is "noticeably smaller than medium but not so small as to be trivial," and a large effect is "the same distance above medium as small was below it." (p. 156). Such an interpretation attempts to anchor the scale of effect size indices in observable phenomena. For Cohen's  $d$ , a value of 0.20 is interpreted as a small effect, 0.50 is a medium effect, and 0.80 is a large effect.

Because Cliff did not suggest the use of his delta statistic beyond hypothesis testing and confidence interval estimation, he did not suggest corresponding values to represent small, medium, and large effects. However, Cohen (1988) presents interpretations of the effect size index  $d$  in terms of the non-overlap between two normal distributions. This provides a direct bridge between  $d$  and delta. That is, with two normal distributions, a difference in means that

represents a  $d$  effect size of 0.20 will have of delta value of 0.147, a  $d$  effect size of 0.50 corresponds to a delta value of 0.33, and a  $d$  effect size of 0.80 corresponds to a delta of 0.474.

#### *Use of Raw Differences with NSSE data*

While all of the above statistical methods may be used to make inferences about the differences between groups, another simple way to evaluate the differences between groups is simply to calculate the “raw” differences in responses. In the case of this study we dichotomized the data such that the ‘top half’ of one group was compared to the ‘top half’ of the comparison group. For example if there was an even number of options, (e.g. disagree, slightly disagree, slightly agree and agree), then the proportion of students from the masters group or peer group that answered slightly agree and agree on an item was subtracted from the proportion of students that answered similarly for the institution. For items that had an odd number of choices, the top half was the smaller of the two “halves” (e.g., if there were seven possible choices then the top three choices were considered the top proportion). An exception was made for the items from number seven of the survey to reflect the methodology applied for the NSSE results. There were seven items from this group that had four possible answers: have not decided, do not plan to do, plan to do and done. According to the NSSE results table all the answers to these items in their analysis were assigned the value zero except for the answer “done”. The value that they calculated was the proportion of respondents that answered “done” among the valid responses (NSSE Institutional Report, 2005).

The raw difference was calculated for each variable. It was decided that a cut off value of a 10% difference between the institution and the peer or master group on an item was an item that warranted concern for the user. It is important that this raw difference calculation is just simply a raw difference. To calculate a statically significant difference between two proportions one would need to conduct a z test that actually considers standard error as well as the proportion difference.

#### *Results of One Evaluation of NSSE*

The variety of issues with the response sets and results suggests that the  $t$ -test may not be the best way to assess items on the NSSE scalelet. Rather non-parametric tests, which do not require means or know population distributions (Pagano, 2001), such has the chi-square or log odds test might be more appropriate. However, because of the robustness properties of the  $t$ -test, this inferential statistic is often used to compare differences between groups. But, if violations of the assumption of variance homogeneity occur then it is especially necessary to consider the appropriate test statistic and effect size calculation. Non-parametric statistics and alternative effect size indices may provide more statistically valid approaches to the analysis and interpretation of such data.

To demonstrate these various methods we used data from a small private university's 2005 NSSE report. Only the responses of first year students were evaluated. The selected peers group had a response rate of 42% with a sample size of 22,276. The master's group, which was this institution's Carnegie classification, had a 33% response rate with a sample size of 59,454 and the private university had a 26% response rate out of a sample size of 1183. Before we proceed with our results it is important to point out that even though we used the NSSE data to discuss the use of these statistical methods, the issues that will be discussed should be considered for all surveys with items of an ordinal nature.

#### *Characteristics of the Data*

As mentioned earlier, one of the assumptions when using *t*-test and Cohen's *d* is that the population distribution is normal in shape. To evaluate the extent to which this assumption was being violated we calculated the skewness and kurtosis of each of the items for the three groups. Skewness and kurtosis were both evaluated using four different intervals: values that fell between -0.5 and 0.5, values that fell between -1.0 and -0.5 or between 0.5 and 1.0, values that fell between -1.0 and -2.0 or 1.0 and 2.0, and values that were less than -2.0 or greater than 2.0. For this evaluation we used the guideline that anything smaller than -1.0 or larger than 1.0, for either skewness or kurtosis, was potentially violating the assumptions of normality (see Tabachnick & Fidell, 2001).

The box plots in Figures 3 and 4 along with Table 5 illustrate the results of this analysis. As one can see from the box plots, that skewness and kurtosis are relatively consistent across all three of the groups evaluated. It should be noted that there were 7 items from the NSSE that, while having four possible options, were dichotomously scored (i.e. a 0 was assigned to three of the four options and 1 was assigned to the option "done"). These variables, not surprisingly, had rather large values for skewness and kurtosis. Overall in each of the groups about two thirds of the variables did have skewness or kurtosis that exceeded the absolute value of one ( $n = 56$  for Private,  $n = 56$  for Peers,  $n = 57$  for master). Thus we can conclude in all three groups that the assumption of normality appeared to be violated for about one third of the items.

In addition to the assumption of normality, the assumption of equal variances was evaluated using Levene's test. The results indicated that there were 17 comparisons for Private vs. Peers and 15 comparisons for Private vs. Master that suggested violation of the assumption of homogeneity of variances.

Along with the *p* values from the original *t*-test and the Cohen's *d* effect sizes, the data were then evaluated using chi square tests (with both dichotomized and the original item response options), odds ratios, raw differences, and Cliff's delta. The extent to which these

indices agreed in terms of significance was evaluated using the criteria in Table 4 below. If the value of the statistic met the criteria for statistical or practical significance it was assigned a yes, if not it was assigned a no.

Table 4  
Criteria for Agreement

Statistic	Criteria for a Yes
T test	$p \leq .05$
Cohen's d	$d \geq .20$
chi- Square	$p \leq .05$
Odds Ratio	Ratio $\leq .50$ or Ratio $\geq 2.0$
Raw Difference	Raw difference $\geq 10\%$
Cliff's Delta	Delta $> .147$

The results of the extent to which these indices agreed are presented in Tables 6 and 7 along with Cohens's Kappa and the percentage of agreement. Cohen's Kappa is a statistic used to assess inter-rater reliability, i.e. agreement for categorical variables. Kappa is considered to be a better means of evaluating agreement than calculating the percentage agreement because it takes into account the extent to which the agreements would have happened by chance.

Table 6 displays the results for the Peer group indices and the extent to which they agree. The percentage of agreement in this table was rather large when the *t*-test results were compared to Cohen's *d* (81.18%), chi- squared (both dichotomized, 83.53% and un dichotomized 82.35%), and for raw differences (74.12%). Agreement was moderately high for Cliff's delta (69.41%). The percentage agreement between the *t*-test and the odds ratios was much smaller (31.76%). The two effect sizes were also compared and the agreement between them was relatively large (88.24%). There were only 10 variables where these two indices did not agree. According to Fleiss (1981) and Brennan and Prediger (1981), if Cohen's Kappa is between 0.40 and 0.60 the agreement is said to be fair, if it is between 0.60 and 0.75 the agreement is good, and if it is 0.75 or greater the agreement is excellent. For the Peers analysis of agreement the Kappa indices ranged from .13 for the odds ratio agreement with the *t*-test to .63 for the dichotomized chi-square agreement with the *t*-test. Only the comparisons between the *t*-test and the two chi-squared tests were above a .60. The Kappa for the comparison Cliff's Delta and Cohen's *d* was .49. The other comparison had smaller values. Perhaps the reason for this is because Kappa adjusts the observed percent agreement by removing this chance agreement. The adjustment is greater for data such as these than it would be if there were more balance between the yeses and the nos. For example in Table 6 the percentage of agreement between Cliff's delta and Cohen's *d* was 88.24% yet the Kappa value was .49. This occurred because the yeses and nos are not balanced, i.e. there are more nos than yeses!

Similar results were seen for the Master group comparisons presented in Table 7. The percentage of agreement between the *t*-test and Cohen's *d* was 88.24%; for the chi-squared it was 80% when the data were not dichotomized data and 92.94% when the data were dichotomized. The agreement between the *t*-test and the raw differences was also large, 88.88%, and the Cliff Delta was 81.18%. The odds ratio was even lower at 21.18% agreement. The two effect sizes, Cliff's Delta and Cohen's *d*, had 92.94 % agreement.. Cohen's Kappa was also similar for the Masters data, however the Kappa for the odds ratio and the *t*-test was -.13. The Kappa value was relatively high for the chi- square test when the data were dichotomized, Kappa = .80, was more modest for the data that were not dichotomized, Kappa = .54.

Figures 4 and 5 present scatter plots of Cliff Delta and Cohen's *d* for both the peers comparison and the masters comparisons. As the figures indicate both indices have relatively similar results with only a few exceptions. There were 6 effect size comparisons that did not agree for the masters comparisons and 10 for the peers comparisons. When examining these items one of their comparisons for the masters group appeared to violate both the homogeneity assumption and the assumption of population normality while three of the items were in violation of normality only. However two of the items that were not in agreement did not show signs of problems for either issue. When examining the items for the Peers comparisons, three of the 10 items that did not agree for the effect size appeared to violate the assumption of homogeneity of variance and only two appeared to be in violation of normality. The other five appeared not to be violating the assumptions.

Finally, correlation analyses were conducted on all of the indices. This information is presented in Tables 8 and 9. In Table 8 the correlations for the peer comparisons are presented. All of the correlations were found to be statistically significant. The Cohen's *d* index was strongly correlated with Cliff's Delta ( $r = .97$ ) and the raw difference scores ( $r = .92$ ). Cohen's *d* was also highly correlated with the *p*-values from the *t*-test ( $r = -.70$ ). and with the odds ratio ( $r = .77$ ). The correlations are a little smaller for both the chi--square test performed when the data were dichotomized ( $r = -.67$ ) and the chi--squared test when they were not dichotomized ( $r = -.56$ ). Cliff's Delta followed a similar pattern where the correlation ranged from  $r = .97$  for Cohen's *d* to  $r = -.51$  for the chi- squared test on the non-dichotomized data. It too was also highly correlated with the raw difference scores. Over all the *t*-test was not correlated that high with any of the other indices including Cohen's *d* where  $r = -.70$ . Its strongest correlation,  $r = .77$ , was with the chi- squared test for the dichotomized data. Similarly the chi –squared test for the non-dichotomized data had somewhat low correlation with the rest of the indices. Its correlation ranged from  $r = -.46$  with the raw differences to as high as  $r = .59$  with the *t*-test.

In Table 9 the results for the master comparisons are presented. In this analysis Cohen's  $d$  was also highly correlated with Cliff's Delta,  $r = .96$  and had its lowest correlation with the chi-squared test for the non-dichotomized data,  $r = -.58$ . The strongest correlation for the  $t$ -test was with Cohen's  $d$ ,  $r = -.80$  and its weakest correlation was with the odds ratio where  $r = -.52$ . Cliff's Delta's weakest correlation was also with the odds ratio where  $r = .50$ . The odds ratio's strongest correlation appeared to be with the chi-squared test for the data when they were dichotomized ( $r = -.68$ ). The Raw differences strongest correlation was also with Cohen's  $d$ ,  $r = -.83$ , and its weakest was also with the chi-squared test using the dichotomized data. Where  $r = -.44$

Figures 5 and 6 display the scatter grams comparing the two different types of effect sizes, Cohen's  $d$  and Cliff Delta for both the analysis for the peers group and the masters group. The vertical line is the location of the cut-off value for Cliff's Delta, .147. The horizontal line is the location for the cut-off value for Cohen's  $d$ , .20. While both of these values are classified as "small" they are not considered trivial. There are four quadrants that the intersect of the two lines creates. The bottom left quadrant contains the effect sizes of the variables that would be considered someone negligible but in agreement. Values in the upper right quadrant contain effect sizes that are also in agreement and are not trivial. The values in the upper left quadrant represent the variables that are important based on Cohen's  $d$  criteria but not for Cliff's delta. In contrast, the values found in the bottom left quadrant represent the variables where Cliff's delta would be considered important but Cohen's  $d$  would be trivial. Notice that in both charts either the indices are in agreement in regards to the quadrant location or Cohen's  $d$  is critical and Cliff's delta is not. This would indicate that Cliff is a more conservative effect size than Cohen's  $d$  and more robust to the violation of assumptions for means tests.

### *Conclusions*

The purpose of this paper was to explore some of the methods that might be used to evaluate discrete ordinal data and the advantages and disadvantages of using these methods. A summary of the various indices used is presented in Table 10. While  $t$ -test and Cohen's  $d$  are parametric types of evaluations, chi-squared, odds ratios, and Cliff's delta are considered non parametric types of evaluation. The main difference between the parametric evaluations and the non parametric is that the non parametric tests or indices only represent the extent to which groups are different, but not how they are different. The parametric tests evaluate the extent to which groups are different for a specific population characteristic, i.e. means as in the case of a  $t$ -test, of the populations. The advantage to using non parametric evaluations is that there are no assumptions, other than independence. Typically, one does not need to be concerned about

distribution shape and variance homogeneity because the non parametric evaluations are not based on such assumptions.

When examining the data used in this study, even when the assumptions are violated the agreement between non parametric indices and parametric is somewhat stable. What was surprising about the results from our data was that while Cohen's  $d$  and the  $t$ -test were strongly correlated; the Cliff's delta was not strongly correlated with any of the non parametric indices but it was strongly correlated with Cohen's  $d$ . In general most of the time when group differences to responses are evaluated in survey research the data are analyzed with statistics such as  $t$ -tests and Cohen's  $d$ . Overall this research suggests that when evaluating group differences for discrete ordinal data, the use of  $t$  test and Cohen's  $d$  is somewhat robust, but Cliff delta is more robust than Cohen's  $d$  when violation of assumptions occur.

In practice, determining which method is "best" depends on what information is needed from the data. The chi-squared test, both dichotomized and un-dichotomized, along with the odds ratio and raw differences are useful in determining if the groups are different but not necessarily how they are different. The chi-squared can be used to determine the extent to which the distributions of answers are the same for both groups. If the responses are dichotomized with the assumption that the "top half" represents agreement with the item, i.e., a "correct" answer, then the chi-square test can be used to determine the extent to which the two groups proportion of agreement, (and disagreement) are the same. Similarly the odds ratio can be used to determine the extent to which one population is likely to agree more or less than the other. The raw difference is an index that represents the raw difference between agreement in two groups without taking into account the variance of the data. While a rather crude measure, this value does represent the actual proportion of difference between groups. However, it was only strongly correlated with the effect sizes and only mildly correlated with any of the other evaluations.

The results of this study suggest that the responses in survey research should be evaluated to determine the extent to which their distributions violate the assumptions and appropriate adjustment should be made. In addition, if the researcher wishes to evaluate the effect size Cliff's delta seems to be the more appropriate index due to its robust nature. Future research should be conducted using the NSSE benchmarks and other influential surveys to determine if similar results would occur.

## References

- Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41, 687–699.
- Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, 61, 230-258.
- Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114, 494-509.
- Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, 31, 331-350.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2<sup>nd</sup> ed.). New York: Academic Press.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- Crow, E. L. (1991). Response to Rosenthal's comment "How are we doing in soft psychology?" *American Psychologist*, 46, 1083.
- Fern, E. F. & Monroe, K. B. (1996). Effect size estimates: Issues and problems in interpretation. *Journal of Consumer Research*, 23, 89-105.
- Fleiss, J. I. (1981). *Statistical methods for rates and proportions*. New York: Wiley.
- Gorsuch, R. L. (1991). Things learned from another perspective (so far). *American Psychologist*, 53, 800-801.
- Gravetter, F.J. & Wallnau, L.B. (2004). *Statistics for the Behavioral Sciences*, 6<sup>th</sup> edition, California: Wadsworth/Thomson.
- Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hogarty, K. Y. & Kromrey, J. D. (2001, April). *We've been reporting some effect sizes: Can you guess what they mean?* Paper presented at the annual meeting of the American Educational Research Association, Seattle.
- Indiana University Center for Postsecondary Research (2005). *National Survey of Student Engagement(NSSE) 2005 Institutional Report*. Bloomington, IN: Author.
- Kirk, R. E. (1996). Practical Significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.
- Kraemer, H. C., & Andrews, G. A. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin*, 91, 404-412.

- 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, 70- 82.
- McGraw, K. O. (1991). Problems with BESD: A comment on Rosenthal's "How are we doing in soft psychology?" *American Psychologist*, 46, 1084-1086.
- McGraw, K. O. & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111, 361-365.
- Myers, J.L. & Well, A.D. (1995). *Research Design and Statistical Analysis*. Hew Jersey: Lawrence Erlbaum Associates
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241-301.
- Pagano, R.R. (1995). *Understanding Statistics in the Behavioral Sciences*, 6<sup>th</sup> edition. California: Wadsworth/Thomson.
- Parker, S. (1995). The "difference of means" may not be the "effect size." *American Psychologist*, 50, 1101-1102.
- Rosenthal, R. (1991). Effect sizes: Pearson' correlation, its display via the BESD, and alternative indices. *American Psychologist*, 46, 1086-1087.
- Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680.
- Strahan, R. F. (1991). Remarks on the binomial effect size display. *American Psychologist*, 46, 1083-84.
- Thompson, B., & Snyder, P.A. (1997). Statistical significance testing practices in the *Journal of Experimental Education*. *Journal of Experimental Education*, 66, 75-83.
- Thompson, B., & Snyder, P.A. (1998). Statistical significance and reliability analyses in recent JCD research articles. *Journal of Counseling and Development*, 76, 436-441.
- Vargha, A. & Delaney, H.D. (2000). A critique and improvement of the CL Common Language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25, 101-132.
- Velleman, P. F. & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 47(1), 65-72.
- Wilcox, R. R. & Muska, J. (1999). Measuring effect size: A non-parametric analogue of  $\omega^2$ . *British Journal of Mathematical and Statistical Psychology*, 52, 93-110.
- Wilkinson, L., and American Psychological Association (APA) Task Force on Statistical Inference.(1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604. )

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

Table 5

## Skewness and Kurtosis of the data

<b>Small Private</b>	<b>Kurtosis</b>				
<b>Skewness</b>	$-.5 < x < .5$	$-1 < x < -.5$ or $.5 < x < 1$	$-2 < x < -1$ or $1 < x < 2$	$x < -2$ or $x > 2$	Grand Total
$-.5 < x < .5$	7	29	11		47
$-1 < x < -.5$ or $.5 < x < 1$	14	6			20
$-2 < x < -1$ or $1 < x < 2$	2	2	5		9
$x < -2$ or $x > 2$				9	9
Grand Total	23	37	16	9	85

<b>Selected Peers</b>	<b>Kurtosis</b>				
<b>Skewness</b>	$-.5 < x < .5$	$-1 < x < -.5$ or $.5 < x < 1$	$-2 < x < -1$ or $1 < x < 2$	$x < -2$ or $x > 2$	Grand Total
$-.5 < x < .5$	6	37	9		52
$-1 < x < -.5$ or $.5 < x < 1$	11	2	1		14
$-2 < x < -1$ or $1 < x < 2$	3	3	2	3	11
$x < -2$ or $x > 2$			2	6	8
Grand Total	20	42	14	9	85

<b>Masters</b>	<b>Kurtosis</b>				
<b>Skewness</b>	$-.5 < x < .5$	$-1 < x < -.5$ or $.5 < x < 1$	$-2 < x < -1$ or $1 < x < 2$	$x < -2$ or $x > 2$	Grand Total
$-.5 < x < .5$	7	35	10		52
$-1 < x < -.5$ or $.5 < x < 1$	11	4			15
$-2 < x < -1$ or $1 < x < 2$	3	2	3	1	9
$x < -2$ or $x > 2$				9	9
Grand Total	21	41	13	10	85

Figure 3 Skewness of the Data

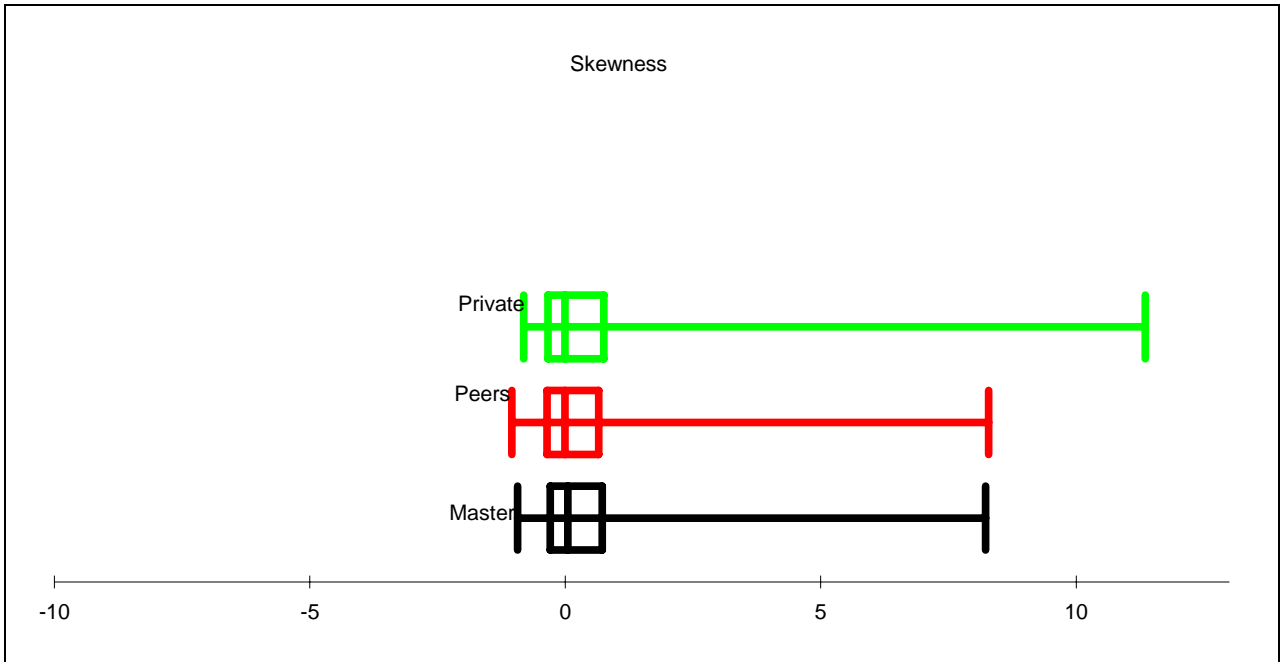
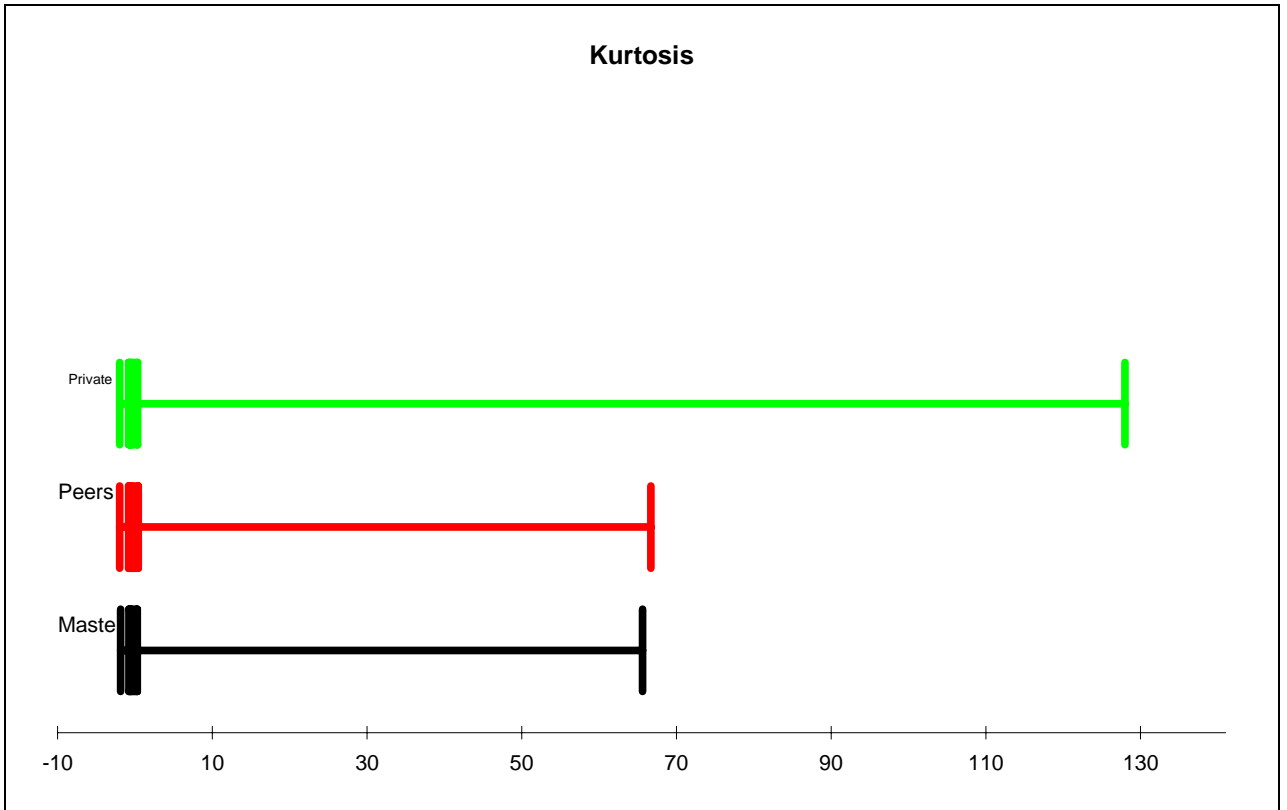


Figure 4 Kurtosis of the Data



**Table 6**  
**Statistics for Peers Analysis**

<b>Cohen's d</b>			
t test	Yes	no	Grand Total
Yes	16	16	32
No		53	53
Grand Total	16	69	85
	<b>% Agreement</b>		<b>Kappa</b>
	81.18%		0.55

<b>Odds Ratio</b>			
t test	Yes	No	Grand Total
Yes	26	6	32
No	52	1	53
Grand Total	78	7	85
	<b>% Agreement</b>		<b>Kappa</b>
	31.76%		0.13

<b>Un-dichotomized Chi Square</b>			
t test	Yes	No	Grand Total
Yes	26	6	32
No	9	44	53
Grand Total	35	50	85
	<b>% Agreement</b>		<b>Kappa</b>
	82.35%		0.63

<b>Raw Difference</b>			
t test	Yes	No	Grand Total
Yes	10	22	32
No		53	53
Grand Total	10	75	85
	<b>% Agreement</b>		<b>Kappa</b>
	74.12%		0.36

<b>Dichotomized Chi Square</b>			
t test	Yes	No	Grand Total
Yes	21	11	32
No	3	50	53
Grand Total	24	61	85
	<b>% Agreement</b>		<b>Kappa</b>
	83.53%		0.63

<b>Cliff's Delta</b>			
t test	Yes	No	Grand Total
Yes	6	26	32
No		53	53
Grand Total	6	79	85
	<b>% Agreement</b>		<b>Kappa</b>
	69.41%		0.22

<b>Cliff's Delta</b>			
Cohen's d	Yes	No	Grand Total
yes	6	10	16
no		69	69
Grand Total	6	79	85
	<b>% Agreement</b>		<b>Kappa</b>
	88.24%		0.49

Statistic	Criteria for a Yes
T test	$p \leq .05$
Cohen's d	$d \geq .20$
Chi Square	$p \leq .05$
Odds Ratio	Ratio $\leq .50$ or Ratio $\geq 2.0$
Raw Difference	Raw difference $\geq 10\%$
Cliff's Delta	Delta $> .147$

**Table 7**  
**Statistics for Masters Analysis**

<b>Cohen's d</b>			
t test	Yes	no	Grand Total
Yes	12	10	22
No		63	63
Grand Total	12	73	85
		<b>% Agreement</b>	<b>Kappa</b>
		88.24%	0.64

<b>Odds Ratio</b>			
t test	Yes	No	Grand Total
Yes	16	6	22
no	61	2	63
Grand Total	77	8	85
		<b>% Agreement</b>	<b>Kappa</b>
		21.18%	-0.13

<b>Un-dichotomized Chi Square</b>			
t test	Yes	No	Grand Total
Yes	18	4	22
No	13	50	63
Grand Total	31	54	85
		<b>% Agreement</b>	<b>Kappa</b>
		80.00%	0.54

<b>Raw Difference</b>			
t test	Yes	No	Grand Total
Yes	10	12	22
no		63	63
Grand Total	10	75	85
		<b>% Agreement</b>	<b>Kappa</b>
		85.88%	0.55

<b>Dichotomized Chi Square</b>			
t test	Yes	No	Grand Total
Yes	16	6	22
No		63	63
Grand Total	16	69	85
		<b>% Agreement</b>	<b>Kappa</b>
		92.94%	0.80

<b>Cliff's Delta</b>			
t test	Yes	no	Grand Total
Yes	6	16	22
no		63	63
Grand Total	6	79	85
		<b>% Agreement</b>	<b>Kappa</b>
		81.18%	0.36

<b>Cliff's Delta</b>			
Cohen's d	Yes	no	Grand Total
Yes	6	6	12
no		73	73
Grand Total	6	79	85
		<b>% Agreement</b>	<b>Kappa</b>
		92.94%	0.63

Statistic	Criteria for a Yes
T test	$p \leq .05$
Cohen's d	$d \geq .20$
Chi Square	$p \leq .05$
Odds Ratio	Ratio $\leq .50$ or Ratio $\geq 2.0$
Raw Difference	Raw difference $\geq 10\%$
Cliff's Delta	Delta $> .147$

**Table 8**

<b>Correlations for Peers Statistics</b>							
	Cohen d	p-value from t-test	Cliff delta master	Odds Ratio	p-value from Chi-square Dichotomized data	p-value from Chi-square Un-Dichotomized data	Raw Difference Score
Cohen d	1	-0.70314	0.97321	0.77068	-0.67008	-0.56795	0.91647
p-value from t-test	-0.70314	1	-0.65391	-0.59334	0.77497	0.5983	-0.5712
Cliff delta	0.97321	-0.65391	1	0.70282	-0.59772	-0.51889	0.90136
Odds Ratio	0.77068	-0.59334	0.70282	1	-0.73335	-0.55413	0.75614
p-value from Chi-square Dichotomized data	-0.67008	0.77497	-0.59772	-0.73335	1	0.48556	-0.72714
p-value from Chi-square Un-Dichotomized data	-0.56795	0.5983	-0.51889	-0.55413	0.48556	1	-0.46308
Raw Difference Score	0.91647	-0.5712	0.90136	0.75614	-0.72714	-0.46308	1

**Table 9**

**Correlation for Masters Statistics**

	Cohen d	p-value from t-test	Cliff delta master	Odds Ratio	p-value from Chi-square Dichotomized data	p-value from Chi-square Un-Dichotomized data	Raw Difference Score
Cohen d	1	-0.79849	0.96286	0.62515	-0.63136	-0.58588	0.83579
p-value from t-test	-0.79849	1	-0.75034	-0.52314	0.60174	0.6192	-0.55661
Cliff delta	0.96286	-0.75034	1	0.50218	-0.5568	-0.53798	0.77352
Odds Ratio	0.62515	-0.52314	0.50218	1	-0.68971	-0.48643	0.61612
p-value from Chi-square Dichotomized data	-0.63136	0.60174	-0.5568	-0.68971	1	0.48587	-0.79571
p-value from Chi-square Un-Dichotomized data	-0.58588	0.6192	-0.53798	-0.48643	0.48587	1	-0.44947
Raw Difference Score	0.83579	-0.55661	0.77352	0.61612	-0.79571	-0.44947	1

Figure 5

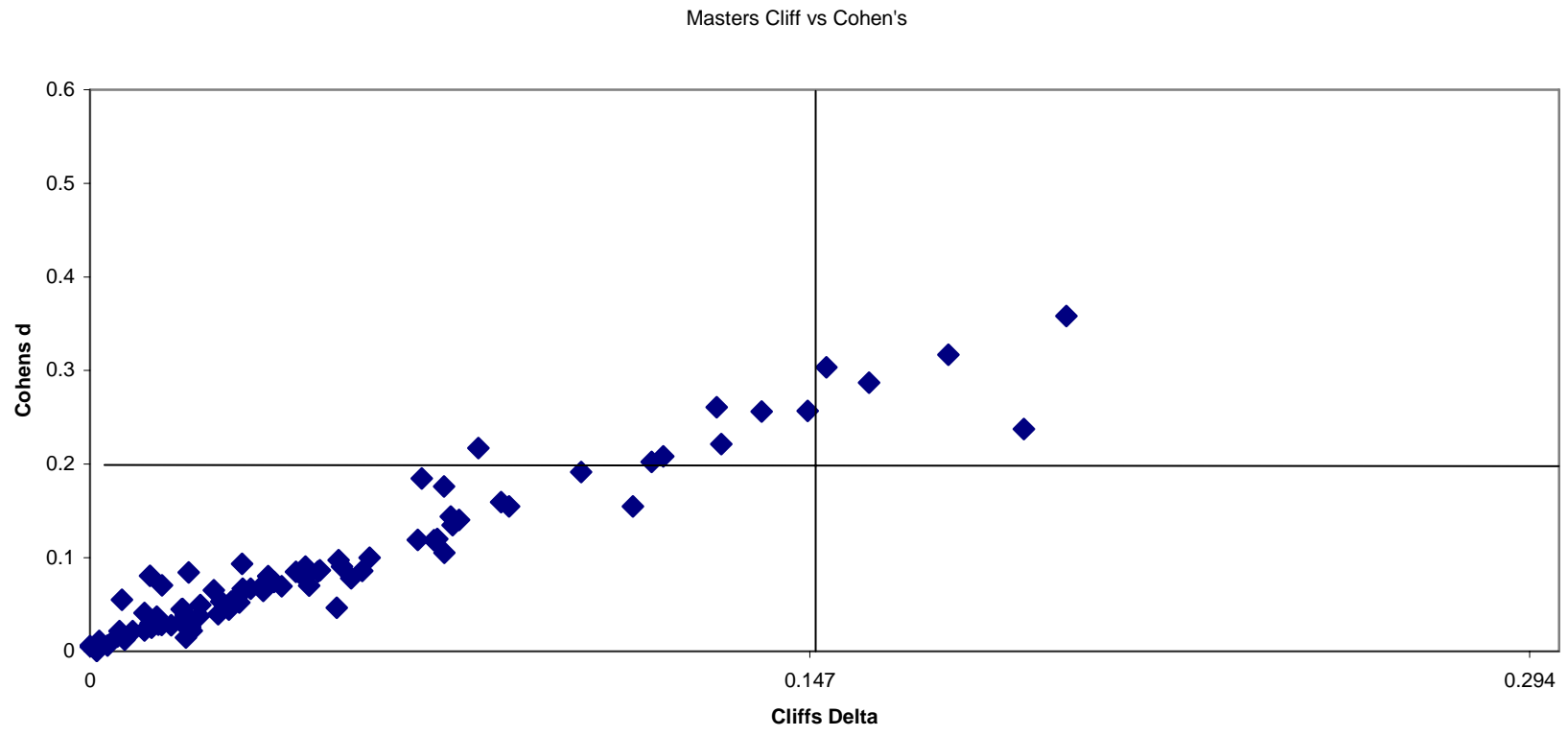
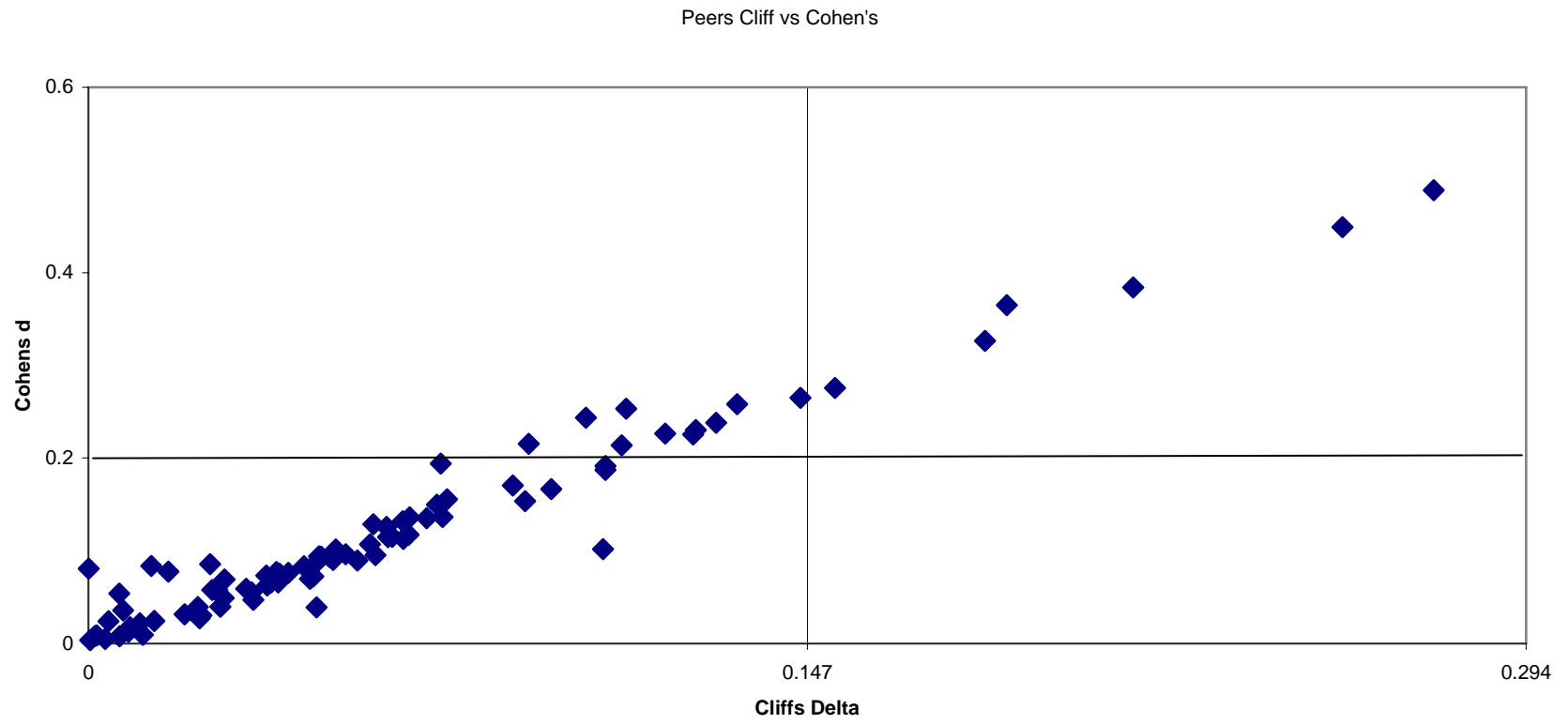


Figure 6



**Table 10 Overview of the Various Methods**

<i>Statistical Method</i>	<i>What is it?</i>	<i>How do interpret the values</i>	<i>Advantage and Disadvantages</i>
<i>T tests</i>	Parametric test used to determine whether or not two unrelated sample means are significantly different from one another	If the <i>p</i> value for the test is less than .05 then the difference between the two group means are said to be statistically significant	<i>Advantage</i> - evaluates the difference between the based on the means of each group. <i>Disadvantage</i> - the Type I error rates can be in be inflated if the assumption of normality or equal variances are violated. Also, larger samples can result in statistical significance that is not necessarily “practical” significance
<i>Cohens d</i>	Effect size represents the difference between sample means in standard deviation units.	The <i>d</i> represents the difference between the two groups in standard deviation units. A <i>d</i> =1 would indicate that the two groups are one standard deviation apart	<i>Advantage</i> - evaluates the difference between the based on the means of each group. Is not a function of sample size. <i>Disadvantage</i> - can be inflated if the data violates the assumptions of normality or equal variances
<i>Chi Squared Test using dichotomized data</i>	Non parametric test used to determine whether the two sample proportions are “significantly” different from one another. Specifically if the data is dichotomized, it is used to determine the differences between two samples with only two levels	If the <i>p</i> value for the test is less than .05 then the “top half” of one group is said to be statically significantly different than the “top half” of the second group.	<i>Advantage</i> – this is a non parametric test therefore the only assumption is independence. <i>Disadvantage</i> - Only indicates if the groups are different but not what is different about the two groups.
<i>Chi Squared Test using Un - dichotomized data</i>	Non parametric test used to determine whether the two sample proportions are “significantly” different from one another.	If the <i>p</i> value for the test is less than .05 then the distribution of the answers for one group is said to be statistically significantly different than the distribution of answers for the second group	<i>Advantage</i> – this is a non parametric test therefore the only assumption is independence. <i>Disadvantage</i> - Only indicates if the groups are different but not what is different about the two groups.
<i>Cliffs Delta</i>	Effect size used to test equivalence of probabilities of scores in one group being larger than scores in the other	The delta value represents the degree of overlap between the two distributions of scores. It ranges from -1 (if all observations in Group 1 are larger than all observations in Group 2) to +1 (if all observations in Group 1 are smaller than all observations in Group 2) and takes the value of zero if the two distributions are identical.	<i>Advantage</i> – this is a non parametric test therefore the only assumption is independence. Also, it is fairly simple value to interpret. <i>Disadvantage</i> - Only indicates if the groups are different but not what is different about the two groups.
<i>Odds Ratios</i>	Statistic that provides information about the strength of the relationship between the two variables of interest.	If the odds ratio is exactly one, than the two groups have the same probability of the event occurring. If the odds ratio is greater than one, than the group of interest has a higher probability of the event occurring. If the odds ratio is less than one, than the control group has a higher probability of occurring.	<i>Advantage</i> – this is a non parametric test therefore the only assumption is independence. <i>Disadvantage</i> - Is a measure of likelihood and sometimes difficult to interpret.
<i>Raw Differences</i>	An index that represents the raw difference in agreement between two groups	The value would indicated the extent to which one group is agreeing more than the other group.	<i>Advantage</i> – is a simple value to interpret and calculate <i>Disadvantage</i> – Does not take into consideration the “bottom half” of the data or the variance in the data.

