Is Cliff's δ More Robust to Kurtosis than Robust Cohen's D?

by

Kit Duguay

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

MASTER OF ARTS

Department of Psychology

University of Manitoba

Winnipeg

Abstract

As null hypothesis significance testing (NHST) has been criticized as the sole arbiter of a study's worth, measuring effect sizes has taken greater prominence in psychological research. There are 3 effect size measures (i.e., Cohen's *d*, robust Cohen's d_r, and Cliff's δ) that could compare the difference between two groups of observations (e.g., boys/girls difference on weight). Cohen's *d* (Cohen, 1997) is arguably the most widely employed one, but its' accuracy or robustness depends upon some data assumptions (e.g., the scores such as weight needs to follow normal distribution). In fact, these violations of normality are to be expected when observing data in many research practices (Lyon, 2004). Robust Cohen's d_r and Cliff's (1993) δ, are proposed as alternatives that are robust to violations of those data assumptions. Cliff's (1993) δ describes ordinal data and as such is robust to violations of normality, and robust Cohen's d_r is a proposed modification to Cohen's *d* that trims and winsorizes the data to normalize it (Algina et al., 2005). However, there is no single study that compares and evaluates the robustness of the 3 effect size measures in one single Monte Carlo experiment. Through analyzing data sets created via Monte Carlo simulation—a computer-based experiment that addressed a design with a total of 4 levels (sample sizes) x 4 levels (effect sizes) x 2 levels (normal and mixed-normal) = 32 manipulated levels with 1,000 replications (i.e., generating a total of 32 000 simulated datasets for evaluation)—it has been shown that Cliff's δ is more robust to violations of normality than Cohen's *d*, and Cohen's d_r is more robust to violations of normality than Cliff's δ. However, despite the higher level of robustness shown by Cohen's d_r, Cliff's δ uses cases independent of Cohen's *d* and the high level of data trimming suggested by Algina et al. (2005) may not be epistemologically sound when used as the default option for measuring effect size.

Keywords: Effect sizes, Cohen's *d*, Cliff's δ, robust statistics

ACKNOWLEDGEMENTS

I want to thank Dr. Johnson Li for both his intense support and ability to identify problems with my thesis. I have only come this far with his support. I would also like to thank Drs. Janelle Mann and Ryan Giuliano for acting as the other two members of my thesis committee. While I have not engaged with either group on campus for years, I would like to thank St. John's College (University of Manitoba) as well as the Desautels Faculty of Music for providing me a sense of community during my undergraduate years.

I want to dedicate this to two people – my mother, Roxanna M L Duguay, and my partner, Alyssa Lynch. Both have provided me with the strength and impetus to keep going through my education, and both have believed in me during phases that I did not. I am deeply grateful to them for both seeing the person I can be and inspiring me to be that person.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES

CHAPTER 1 – INTRODUCTION

1.1 Background

The primacy of the normal distribution is accepted in statistics because of the central limit theorem (CLT) stating that if you have many independently scored random variables (even ones that are not normally distributed), summing them together will create a probability distribution that will approach the normal distribution as the sample size approaches infinity (Lyon, 2004). However, this is misinterpreted to argue that probability distributions in nature are themselves normal, which is rarely the case (Lyon, 2004). Lyon approaches this from a philosophical standpoint by demonstrating that it is epistemologically unsound to explain away seemingly normal distributions in nature with the CLT. He quotes Roush and Webb (2000) using the seemingly normal distribution of the tensile strength of steel components to demonstrate the CLT's explanatory power as an example of how the CLT is used to explain phenomena it does not have the power to describe. Most interpretations of the CLT (in fact, a set of theorems with varying constraints) require a set of random variables that are both independently distributed and that all have the same probability distribution. When summing these random variables, the probability distribution of this sum will approach the normal distribution as the number of sums increases. However, Roush and Webb are using variables that are not easy to numerically describe (Lyon's sticking point is them implying the concept of "machining process" (Lyon, p. 630) is a straightforward thing to numerically describe and to sum up). Lyon states that a less strict version of the normal distribution (rarely actually found in nature) is the log-normal distribution, which removes the need to sum sequences of random variables directly. This is because the logarithm of factors multiplied together is equal to the sum of their logarithms. While the log-normal distribution can appear similar to the normal distribution when the

variance is small compared to the mean, the variance can be large, and this will result in a very different looking distribution. It should be noted that another distribution – the mixed-normal distribution presents a slightly different problem in that it can look similar to the normal distribution despite having a larger variance and thus violating normality (Algina et. al, 2005). Lyon cites Limpert et al. (2001)'s study of the use of normal and non-normal distribution (e.g., log-normal distributions) in science to show that log-normal distributions always fit unmanipulated data better than normal distributions. Limpert et al. (2001) explain that normal distributions must allow for negative values (all values in a log-normal distribution must be positive). They then argue that the effects of variables on physical processes are generally multiplicative instead of additive, which means that directly summing factors instead of their logarithms (because the sum of factors' logarithms is equal to the logarithm of the product of factors) will not adequately describe the data. While Lyon's paper is about the limits of the CLT's explanatory power (he points out his criticism of tensile strength being explained through the CLT is still valid if it follows a log-normal distribution), the purpose of citing Lyon is to demonstrate that the normal distribution is not as common as many statistical methods require it to be. Thus, it is necessary to examine what statistical methods are robust to violations of normality as can be seen in even superficially-visually-similar mixed-normal distributions, and we must examine how to create more robust methods.

1.2 Brief Outline of my Research

This thesis is divided into five chapters. The first is this introduction, providing a rationale for this study. The second is a review of the literature surrounding effect sizes (seen as deserving increased attention in a null hypothesis significance testing (NHST)-dominated field), focusing on Cohen's *d* (a popular parametric effect size measure) and its robust alternative

(Robust Cohen's d, written as d_r in this study) versus Cliff's δ (a non-parametric effect size measure more robust to violations of normality). The third chapter is a description of the simulation conditions of the study as well as what criteria are being used to determine robustness to violations of normality. The fourth chapter discusses the results of this study, and the fifth discusses pitfalls of Cohen's d (and d_r versus Cliff's δ, possible use cases for each, possible ways to rectify their shortcomings, and possible avenues of further study.

1.3 Values and Impacts of the Study

Using effect size measures to supplement p-values has become a notable trend in psychological research because null hypothesis significance testing (NHST) has structural issues making reliance on it as the only measure of a study's worth inadvisable. The p-value, as it states itself, only shows how likely the results of your specific study are to have occurred by chance. A large effect size could happen at what people call "marginal significance" and be discarded (or kept in clear violation of its own standards of behavior), and an imperceptible effect size could happen at p=0.00001. Thus, determining what measures of effect size are best suited for interpreting data is important for separating the wheat from the chaff in terms of scientific importance. A popular measure of effect size for comparing two groups of scores is the standardized difference between two means, also known as Cohen's *d.* It is susceptible to violations of normality (Algina et al., 2005; Hess & Kromrey, 2004) but because of a longstanding assumption that this is not pervasive in data sets, it sees common use. However, as mentioned, normal distributions are a subset of non-normal distributions, and it is more apt to say that any data distribution specifically follows a non-normal distribution instead of a normal distribution unless the data has been normalized somehow; for example, by the methods described by Algina et. al (2005). Thus, it is important to determine both how robust popular

effect size measures are to non-normally-distributed data, as well as what can be done to make our methods more robust. As such, three effect size measures will be compared. They are Cohen's *d,* a version of Cohen's *d* made more robust to normality violations, and the measure Cliff's δ, which has been shown to have superior confidence band coverage to Cohen's *d* when analyzing data (Hess & Kromrey, 2004). Ultimately, this study may be used to inform the reader of both the need to use robust statistics and give some examples of robust statistics in both robust Cohen's *d* and Cliff's δ.

CHAPTER 2 – Literature Review

2.1 Overview

In psychological research, reporting and interpreting effect size measures in addition to null hypothesis significance testing (NHST) has received increasing attention (Kelley & Preacher, 2012). Jacob Cohen (1977) describes using the standardized mean difference, *d*, as a measure of effect size. It is the difference between two means divided by their pooled standard deviation. While Cohen's *d* is the most popular measure of effect size for tests of two independent samples it also relies on the assumption that the distribution of sample means in the data is normal (Li, 2016). However, it is possible to make a more robust version of Cohen's *d*; this method is described by Algina et al. (2005). On top of this, another measure of effect size, known as Cliff's δ (Hess & Kromrey, 2004), has been found to be robust to violations of normality. These two measures of effect size have key differences and there has not been a comparison of Cliff's δ's performance versus robust Cohen's *d* (d_r).

Cliff (1993) proposed three effect size measures for research scenarios involving two groups of comparisons. That is, Δ_1, which is defined as the percentage of area common to the treatment and control distribution (Tilton, 1937), Δ_2 (stochastic or probabilistic dominance; Grissom & Kim, 2012), which is defined as the proportion of randomly sampled pairs of treatment versus control scores where the treatment score is larger than the control), and Δ_3, which is called the "dominance statistic" (Cliff, 1993) that is defined as the proportion of randomly sampled pairs of treatment versus control scores where the treatment score is larger than the control minus the proportion of pairs where the control score is larger than the treatment score. Cliff (1996) further refined these methods and developed a statistic known as Cliff's δ in the literature. Hess and Kromrey (2004) state that it is for determining how likely it is that

individual observations within the first group created have a higher score than individual observations within the second group. The population parameter for this statistic is the probability that when selecting a random member of one population, its response is higher than a randomly selected member of a second population, after subtracting the probability of the reverse occurring (Hess & Kromrey, 2004). The sample estimate, necessary when measuring the entire population is unfeasible, is calculated simply by dividing Δ_3 by the product of the sample sizes in groups 1 and 2.

In sum, there is a research gap; while we know that Cohen's *d* is not robust to violations of normality and we know that both Cliff's δ and what this paper will refer to as Cohen's d_r (Algina et al., 2005) are robust to violations of normality, we do not know which of the two is more robust. To resolve this, we shall conduct a Monte Carlo experiment to simulate non-normal data and interpret these data with the three previously mentioned statistics. The goal of this proposed study is to provide empirical evidence and guidelines for researchers to select and interpret the most accurate or robust effect size measures, when their research involves comparisons of two groups of observations, but these observations may deviate from normal distribution.

This thesis will thus address two questions. Firstly, it will address the question of Cohen's d_r and Cliff's δ's robustness when applied to mixed-normal distributions showing kurtosis without skew. Secondly, it will address the question of which of these two is more robust to this type of distribution.

2.2 Definition of Terms Used in this Study

In psychological research, effect size is defined "as a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest"

(Kelly & Preacher, 2012, p. 137). Cohen's standardized mean difference *d* is a widely employed effect size measures, and it refers to the quantitative definition of the size of the difference between two groups. It is calculated with the following equation:

$$d = \frac{M_1 - M_2}{s_p}$$

In this equation, $(M_1 - M_2)$ is the difference between two means, and s_p is the pooled standard deviation. It is calculated with the following equation:

$$s_p = \sqrt{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)}$$

In this equation, n_i is the sample size (*i* refers to an arbitrary group level, either 1 or 2) and s_i^2 is the variance for the scores in these groups.

In recent years, measures of effect size have taken precedence over null hypothesis significance testing (NHST), as the weaknesses of this method have become more visible (Cohen, 1994). This is because effect size measures can directly present the strength or magnitude of a study effect, which is often found to be more understandable and interpretable than the conventional, NHST. Moreover, according to the concept of NHST, a small effect size will lead to a significant result (i.e., $p < .05$), when the size of a sample increases. Some authors (e.g., Head, Holman, & Lanfear, 2015) questioned and even criticized the practice of researchers trying whatever strategies (e.g., increasing the sample size, changing the statistical procedures, etc.) in order to obtain a significant *p* value, a situation known as "*p*-hacking".

Robustness refers to the sensitivity (or lack thereof) of a statistic to violations of its underlying assumptions. There are two measures of robustness for a given estimator; these are its influence function (IF) and its breakdown point (BDP). The BDP of an estimator is defined by Hampel (1974) as the smallest proportion of free contaminations which can maintain the value of

an estimator over all bounds, and the influence function describes how much altering a value within the sample will contaminate this estimator.

A normal distribution is a probability distribution with the following probability density function on the domain $x \in (-\infty, \infty)$ (Feller, 1971):

$$n(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

The normal distribution is closely related to the central limit theorem, which states that for a large sample of *n* observations from a population with a finite mean and variance, the sampling distribution of the sum or mean of samples of size *n* is approximately normal (Anderson, 2010, page number for direct quotation). A mixed normal distribution refers to a collection of data obtained from multiple distributions, with one being normal and at least one having a non-normal level of kurtosis.

The term "Monte Carlo simulation" or "Monte Carlo experiment" describes a family of simulations characterized by repeated random sampling from data. This is often used in quantitative-psychology study, in which thousands of samples of data with manipulated levels of conditions (e.g., non-normality) are generated from a computer-simulated platform, and a statistical procedure (e.g., Cohen's *d*) can be applied to these datasets so that the performance or behaviour of this statistic can be examined and evaluated.

A trimmed mean is a mean in which a certain percentage of the smallest and largest data points have been removed (Algina et al., 2005). Winsorizing data refers to a similar practice in which the smallest and largest data points are replaced with the trimmed minimum and maximum data points. This practice was developed by Winsor in 1947 (Hastings, Mosteller, Tukey, & Winsor, 1947). Robust Cohen's *d* (i.e., d_r) was developed based on the idea of trimming and winsorizing, which will be discussed in the following section.

2.3 Background of the Problem

NHST has been severely criticized since the 1950s (Cohen, 1994), but it remains as the primary measure of an experimental result's magnitude. Specifically, it has been criticized for the following reasons. It is misinterpreted so that researchers take non-significant results as explicit support for the null hypothesis, it is often the case that the null hypothesis is much less likely to be true than the alternative hypothesis (making it a less interesting question to ask than how large the effect size is), and it is often used in situations that call for other measures (Lakens, 2019). While Lakens does defend NHST as a valid method of analysis, he describes a burgeoning movement to replace the usage of NHST with other measures, such as effect size. Cohen (1994) was one of the researchers who popularized the usage of effect size, and his effect size measure "*d*" can be read about in many introductory psychological textbooks.

However, Cohen's *d* has a significant problem in that it is not robust to outliers or violations of normality. This is because Cohen's *d* reflects the difference between two means, and a mean can be heavily skewed by outliers. If there is no limit on the range of values in the data set, the breakdown point for the mean – that is, the percentage of outliers/incorrect observations required to render the given estimator (in this case the mean) incorrect – is 0, because a single outlier can be arbitrarily large and thus arbitrarily alter the mean. In order to solve this issue, some researchers proposed that one can use a trimmed mean, which has a breakdown point of X%, where X represents the percentage of data one has trimmed from each end of the data set. Algina et al.'s (2005) d_r is obtained by trimming the mean (the location parameter) by 20% and replacing the standard deviation (the scale parameter) with the square root of the 20% winsorized variance. This measure is recommended by Algina et al. over Cohen's *d* because Cohen's *d* underreports the effect size in distributions with heavy tails

(Presumably it also over-reports the effect size in distributions with lighter tails). In addition to these alterations, the sample effect size d_r is multiplied by .642 in order to ensure that the robust population effect size equals Cohen's population effect size when being used on normally distributed data and equal variances. The value of .642 has been found to be the most accurate value in order to adjust for the skewed data based on previous simulation studies, and it is widely employed in the area of robust statistics. In conclusion, the formula for Cohen's d_r is as follows:

$$d_r = .642(\frac{M_{t1} - M_{t2}}{s_w})$$

In this equation, M_{ti} refers to the trimmed mean (once again *i* refers to an arbitrary level of the grouping variable, either 1 or 2), and s_w refers to the square root of the 20% winsorized variance. To obtain this value we use the following equation:

$$s_w = \sqrt{[(n_1 - 1)s_{w1}^2 + (n_2 - 1)s_{w2}^2]/(n_1 + n_2 - 2)}$$

In this equation, s_{wi}^2 refers to the 20% Winsorized variance, and as before, *i* refers to an arbitrary grouping level, in this case either 1 or 2.

It is important to note that Cohen's *d* is only one of a group of effect sizes representing the standardized mean difference (or SMD) (Hogarty & Kromrey, 1999), which in turn describes the difference between location parameters (e.g. the mean) between the treatment and control groups (Hedges & Olkin, 2016). However, the concept of measuring the level of "overlap" between the distribution of observations in treatment and control groups precedes Cohen's *d*, with Tilton (1937) referring to an "ordinary" definition of overlap in use at the time (which in turn refers to the percentage of scores in the treatment group greater than or equal to the median of the control group). His overlap measure, Δ_1, represented by the following equation:

$$\Delta_1 = 100 \int_{-\infty}^{\infty} \min\{f_T(x), f_C(x)\}dx$$

In this equation, $f_T(x)$ and $f_C(x)$ represent the probability density functions of the treatment and control distributions, respectively. Hedges and Olkin (2016) add that $\Delta_1 = 100$ when the two populations are identical and that $\Delta_1 = 0$ when the populations are completely disjoint. Another measure of overlap is "stochastic or probabilistic dominance" (Hedges & Olkin, 2016), which is represented by the following equation:

$$\Delta_2 = P\{y^T > y^C\},$$

in which scores from the treatment and control groups are randomly sampled and paired, and Δ_2 is the proportion of pairs where y^T is greater than y^C. This measure of overlap is closely related to Cliff (1993)'s dominance statistic, which is defined by Hedges and Olkin (2016) through the following equation:

$$\Delta_3 = P\{y^T > y^C\} - P\{y^C > y^T\} = 2\Delta_2 - 1$$

This statistic is referred to by Hess and Kromrey (2004) as "Cliff's Delta", and the sample estimate of this statistic, Cliff's δ, is obtained by dividing Δ_3 by the size of both groups as follows (Hess & Kromrey, 2004):

$$\hat{\delta} = \frac{\#(x_1 > x_2) - \#(x_2 - x_1)}{n_1 n_2}$$

Cliff's δ differs from Cohen's *d* in the following manners: Firstly, it sorts scores in an ordinal instead of interval fashion; secondly, it is bounded from a range of -1.0 to 1.0 (these extremes represent an absence of overlap, with 0.0 representing identical group distributions); and thirdly, it is non-parametric. These qualities mean that Cliff's δ is more robust to outliers and violations of normality (Hess & Kromrey, 2004) without needing to be made robust by trimming the mean or winsorizing the variance.

2.4 Research Gap

Hess and Kromrey (2004) have compared the accuracy and precision of interval estimates of Cliff's δ over Cohen's d when the conditions of normality and homogeneous variances are violated and have found that Cliff's δ has superior confidence band coverage than Cohen's d when skewness and kurtosis are both violated. At their time of publication, Cohen's d_r had not been developed, and it is of interest to determine whether it is more or less robust than Cliff's δ in turn. In addition, kurtosis is harder to infer from visual inspection of the data distribution (Algina et al., 2005) and it may be that Cliff's δ is more robust to one form of normality violation than the other. As Hess and Kromrey (2004) compared Cohen's d to Cliff's δ, while Algina et al. (2005) only compared Cohen's d to Cohen's d_r at an abnormal level of kurtosis of 3.3 with a skew of 0, it is prudent to determine which robust statistic is more robust to this specific violation. In light, there is a research gap; no study has compared Cliff's δ to Cohen's d_r.

CHAPTER 3 – METHOD

3.1 Explanation of Method

The purpose of this study is to compare whether Cliff's δ is more robust to kurtosis than Cohen's d_r and how much more robust they both are to Cohen's d. This has been done by using Algina et al.'s(2005) methodology. To wit, they compared the robustness of Cohen's d and Cohen's d_r on two sets of data distributions. The first set were normal distributions. In this set, the first distribution had a mean of 0 and a standard deviation of 1, and the second had a mean of 1 and a standard deviation of 1. The second set consists of two mixed-normal distributions comprising 90% normally-distributed data with a standard deviation of 1 and 10% normally-distributed data with a standard deviation of 10. This produces a distribution with a kurtosis of 3.3, which is visually similar to a normal distribution but Cohen's d demonstrates downward bias when used to analyze it. This bias is measured in terms of how many percentage points the true effect size is over or underestimated by, to four decimal places. A level of bias of 1.0000 is equivalent to overestimating by 100%, a level of bias of -1.0000 is equivalent to underestimating by 100%.

3.2 Explanation of Simulation

Data has been created via Monte Carlo simulation in the software environment R – it is organized in 32 conditions across three factors. These factors are the type of distribution (normal vs. mixed-normal with a kurtosis of 3.3), the effect size (nonexistent, small, medium, and large), and the sample size (20, 50, 80, and 150). The precise effect size measures are 0, .2, .5, and .8 for Cohen's d and d_r, and 0, .1476, .3297, and .4743 for Cliff's δ. Each condition consists of 1000 replications. This produces a design with a total of 4 levels (sample sizes) x 4 levels (effect size measures) x 2 levels (normal and mixed-normal). Ultimately, this study uses 32 000 simulated

datasets for evaluation. Appendix 1 consists of the R code used for this simulation, compiled in the software package R (R Core Team, 2013).

CHAPTER 4 – RESULTS

4.1 Results

When looking at Table 1 and Figure 1, one can see that Cohen's *d* is, as expected, highly accurate when comparing two sets of scores following normal distribution. In this case, the biases ranged from -.0058 to .0169, with a mean level of bias of .0041 and a SD of .0074. However, Cohen's *d* is less accurate when comparing sets of scores following a mixed-normal distribution. In this case there was a downward bias ranging from -.3854 to .0075, with a mean level of bias of -.1684 and a SD of -.1464. Robust Cohen's *d* was slightly less accurate comparing normally distributed scores but significantly more accurate when comparing mixed-normally distributed scores. When comparing normally distributed scores, the biases of robust Cohen's *d* ranged from -.0052 to .0337 (slightly more upwardly-biased than non-robust Cohen's *d*), with a mean level of bias of .0066 and a SD of .0107 across the 32 simulated conditions. However, when comparing mixed-normally distributed scores, it exhibited bias ranging from -.0220 to .0120 and showed a mean level of bias of .0056, with a SD of .0085. Cliff's δ demonstrated downward bias both for normally distributed and mixed-normally distributed scores. In fact, its downward bias increased when the scores were mixed-normally distributed, and when the population effect size was large (this was considered a score of .8 for Cohen's *d* and a score of .4743 for Cliff's δ). The biases for the normally distributed scores ranged from -.0575 to .0004 and showed a mean level of bias of -.0330, with a SD of .0206. The biases for the mixed-normally distributed scores ranged from -.0701 to .0062, and showed a mean level of bias of -.0437 with an SD of .0291.

CHAPTER 5 – DISCUSSION

5.1 An Explanation of the Simulation Results

Hess and Kromrey (2004) demonstrated that Cliff's δ has superior confidence band coverage to Cohen's *d* and that it is more robust to violations of normality. In addition, Algina et al. (2005) showed that Cohen's d_r is robust to violations of normality that would hamper the usefulness of Cohen's *d*. Cohen's *d* itself has limited usefulness under Lyon's (2004) assertion that assuming most data sets are best described as normally distributed is epistemologically unsound. This renders the advantages that Cohen's *d* has – it is less upwardly-biased when measuring normal distributions compared to Cohen's d_r, and less downwardly-biased when measuring normal distributions compared to Cliff's δ – moot. Thus, the comparison must be drawn between Cohen's d_r and Cliff's δ. In this field, Cohen's d_r performs more admirably. Cliff's δ exhibits downward bias on normally distributed data and this downward bias increases on mixed-normally distributed data, although it should be noted that Cohen's *d* has the potential to be far more downwardly biased than Cliff's δ does, and its mean level of bias is -.1684 with a SD of -.1464, as opposed to Cliff's δ mean level of bias of -.0437 with an SD of .0291. In layman's terms, on average, Cohen's *d* underestimates the true effect size of mixed-normally distributed data by 16.84% and Cliff's δ underestimates the effect size of mixed-normally distributed data by 4.37%. Conversely, Cohen's d_r slightly overestimates the true effect size of normally distributed data and mixed-normally distributed data by roughly the same amount. The mean level of bias for normally distributed data was .0066 and the mean level of bias for mixed-normally distributed data was .0056 (an overestimation of the true effect size by less than one percent). Conversely, the Cliff's δ mean level of bias for normally distributed data is -.0330, compared to the mixed-normally distributed data's level of bias of -.0437. This has demonstrated

that each measure of effect size's measure of downward bias increases under a mixed-normal distribution with a kurtosis of 3.3 when compared to their level of bias under a normal distribution. While Cohen's *d* shows the lowest level of bias under a normal distribution (mean level of .0041, or it overestimates the true effect size by 0.41%), it is by far the least robust to violations of kurtosis. On average, it underestimates the true effect size by 16.84%, and at large effect sizes, it underestimates the true effect size by over 35%. This difference of levels of bias is .1725, which is expectedly large considering Cohen's *d* is not robust to violations of normality. Conversely, Cliff's δ shows a higher level of downward bias under a normal distribution (-.0330, or underestimating the effect size by 3.30%), but is more robust to violations of kurtosis. On average, it underestimates the true effect size by only 4.37%, and at most only underestimated the true effect size by 7.01%. Thus, the difference between the mixed-normal mean level of bias and the normal mean level of bias is .0107; an order of magnitude smaller than the difference for Cohen's *d*. However, Cohen's d_r is more effective still. While its level of bias under a normal distribution is slightly higher than Cohen's *d* (mean level of .0066, or .0025 higher than Cohen's *d*, amounting to an increase in overestimation by 0.25%), it is even less susceptible to kurtosis than Cliff's δ. The mean level of bias for mixed-normal distributions was 0.0056, and the difference between these levels of bias is 0.0010, which is an order of magnitude lower than Cliff's δ. Ultimately, if it is possible to use either Cohen's d_r or Cliff's δ to analyze a dataset, it is advised to use Cohen's d_r.

5.2 Implications for Use Cases of Cliff's δ vs Cohen's d_r

Cliff's δ is an ordinal statistic (Cliff, 1993); these does not provide information into the intervals between rankings, but simply that the rankings exist (Stevens, 1946). Cohen's *d* and Cohen's d_r are not used for ordinal data as they measure means, which are a measurement of

quantity and thus one cannot get meaningful information from an ordinal "mean". Ordinal statistics do not allow for some familiar methods of analysis; Cliff (1993) cites *t* tests and ANOVA as examples. However, ordinal data analysis has some advantages over interval or ratio data. These are firstly that they can be more robust to violations of normality and only slightly less powerful than traditional forms of data analysis, secondly that much behavioral data is either much easier to describe or only describable with ordinal scales, and thirdly that mean comparisons such as Cohen's *d* and d_r do not describe the questions researchers ask as well as ordinal data can (Cliff, 1993). The results here, in fact, are in line with Cliff's (1993) assertion that ordinal methods have less power under assumptions of normality but are more robust to these violations than what he calls "normal-based methods" such as Cohen's *d*. While ordinal methods are less commonly studied than normal-based methods and ANOVA – a normal-based method - allows for great variation in design, Cliff (1993) provides extensions for more complicated designs using ordinal analysis and cites Ferguson (1965) with more examples. Thus, the main objection to using ordinal statistics – they have limited research designs – is addressed.

Cohen's d_r, conversely, becomes more robust to violations of normality by transforming the data. Not only is Cohen's d_r more robust than Cliff's δ, it is easier for a less-trained statistician to use because it does not require a background in learning ordinal research design. It is simply making some modifications to the data used to obtain Cohen's *d*. While Cliff (1993) makes a persuasive case for more scholarship and pedagogy regarding ordinal statistics and that they may be more fruitful avenues of research than normal-based methods, Cohen's d_r is easier to use in the current research paradigm and more resistant to violations of normality anyway, at the cost of transforming the data somewhat.

5.3 Final Conclusion and Suggestions

Under the current research paradigm where normal-based methods of study design reign supreme, finding a way to make Cohen's *d* robust to violations of normality (that is also more robust than an ordinal alternative) is an appealing choice. However, Cliff (1993) does make persuasive arguments that using ordinal research designs may be more fruitful for psychological research, and the 20% trimmed mean and 20% winsorized variance that Algina et al. (2005) suggest trimming 40% of the total observations. Further research could be done to determine how heavily one must trim/winsorize data to obtain robust results when using Cohen's *d*. Lyon's (2004) assertion that data only fits normal distributions best once it has been manipulated seems to play out here – discarding 40% of every data set does not seem epistemologically sound. Conversely, studying the ease of learning ordinal statistical methods may be fruitful if one is committed to a holistic approach to data analysis where this practice is used sparingly, if at all. It would also be reasonable to determine whether less-heavily trimmed data compares favourably to using Cliff's δ in the same way that Cohen's d_r did. Ultimately, despite its popularity, Cohen's *d* does not effectively analyze real-world data, and alternatives to it have their use in statistics.

References

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohens standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, *10*(3), 317–328. doi:10.1037/1082-989x.10.3.317

Anderson, C. J. (2010). Central Limit Theorem. In *The Corsini Encyclopedia of Psychology*. Hoboken, NJ: Wiley. doi:10.1002/9780470479216.corpsy0160

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology*, *99*, 332-340. doi:10.1037/a0034745

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. doi:10.1037//0033-2909.114.3.494

Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, *31*, 331-350. doi:10.1207/s15327906mbr3103_4

Cohen, J. W. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. doi: 10.1037/0003-066x.49.12.997

Ferguson, G.A. (1965). *Nonparametric trend analysis.* McGill University Press.

Feller, W. (1971). *An introduction to probability theory and its applications. Vol. 2*. New York: Wiley.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: univariate and multivariate applications*. New York: Routledge.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383. doi:10.2307/2285666

Hastings, C., Mosteller, F., Tukey, J. W., & Winsor, C. P. (1947). Low moments for small samples: A comparative study of rrder statistics. *The Annals of Mathematical Statistics*, *18*(3), 413–426. doi:10.1214/aoms/1177730388

Head, M. L., Holman, L., & Lanfear, R. (2015). The extent and consequences of *p*-hacking in science. *PLOS Biology 13*(*3*). doi:10.1371/journal.pbio.1002106

Hedges, L. V., & Olkin, I. (2016). Overlap between treatment and control distributions as an effect size measure in experiments. *Psychological Methods*, *21*, 61-68. doi: 10.1037/met0000042

Hess, M. R., & Kromrey, J. D. (2004, April). *Robust confidence intervals for effect sizes: A comparative study of Cohen's d and Cliff's delta Under non-normality and heterogeneous variances.* Paper presented at the annual meeting of the American Educational Research Association, San Diego,

Hogarty K. Y. & Kromrey, J.D. (1999, August). *Traditional and robust effect size estimates: Power and Type I error control in meta-analytic tests of homogeneity*. Paper presented at the Joint Statistical Meetings, Baltimore

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137-152. doi:10.1037/a0028086

Lakens, D. (2019, April 9). *The practical alternative to the p-value is the correctly used p-value*. Retrieved from https://doi.org/10.31234/osf.io/shm8v

Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, *51*(5), 341. https://doi.org/10.1641/0006-3568(2001)051[0341:lndats]2.0.co;2

Lyon, A. (2014). Why are normal distributions normal? *The British Journal for the Philosophy of Science, 65*(3), 621-649. https://doi.org/10.1093/bjps/axs046

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Roush, M. L., & Webb, W. M. (2000). *Applied Reliability Engineering*. Center for Reliability Engineering, University of Maryland.

Tilton, J. W. (1937). The measurement of overlapping. *Journal of Educational Psychology, 28*(*9*), 656–662

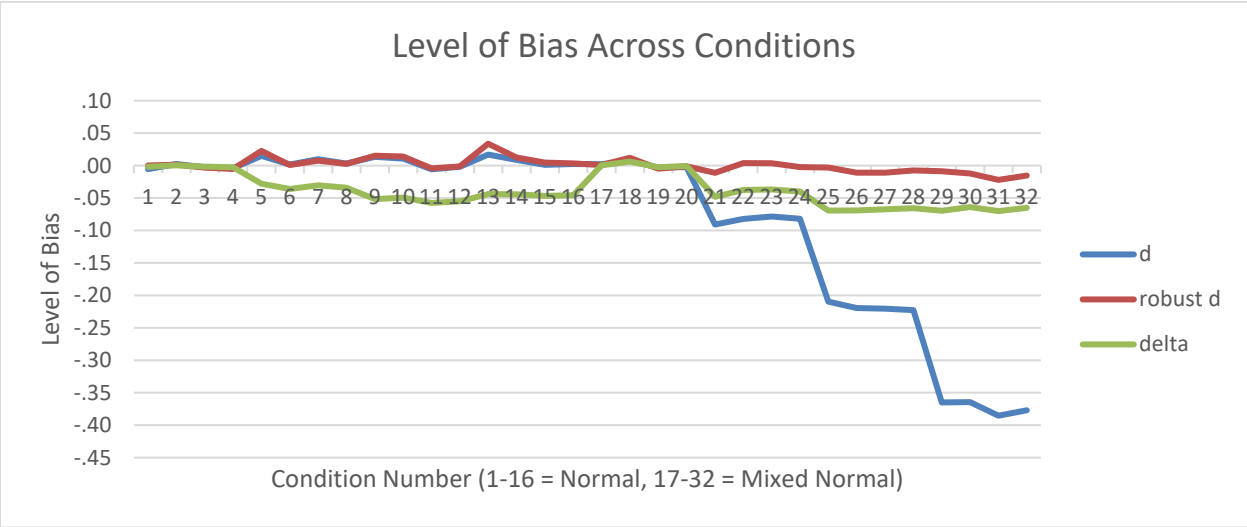
Table 1. Biases of the means of the 1,000 replicated effect size measures (Cohen's *d* (d), Cohen's d_r (robust d), and Cliff's δ (delta)) compared to the associated population values across 32 simulated conditions

Distribution	Population Effect Sizes	n	*d*	robust *d*	delta
Normal	(0, 0)	20	-.0054	.0002	-.0018
		50	.0023	.0013	.0004
		80	-.0027	-.0031	-.0015
		150	-.0039	-.0052	-.0026
	(.2, .1476)	20	.0149	.0227	-.0279
		50	.0017	.0010	-.0361
		80	.0099	.0076	-.0304
		150	.0032	.0027	-.0340
	(.5, .3297)	20	.0135	.0152	-.0517
		50	.0107	.0140	-.0495
		80	-.0058	-.0043	-.0575
		150	-.0019	-.0013	-.0551
	(.8, .4743)	20	.0169	.0337	-.0440
		50	.0088	.0125	-.0443
		80	.0011	.0045	-.0466
		150	.0023	.0033	-.0456
Mixed Normal	(0, 0)	20	.0020	.0010	.0005
		50	.0075	.0120	.0062

	80	-.0031	-.0047	-.0025
	150	-.0024	-.0012	-.0005
(.2, .1476)	20	-.0909	-.0112	-.0481
	50	-.0821	.0038	-.0376
	80	-.0786	.0037	-.0369
	150	-.0819	-.0022	-.0400
(.5, .3297)	20	-.2097	-.0032	-.0693
	50	-.2195	-.0108	-.0691
	80	-.2205	-.0109	-.0673
	150	-.2228	-.0074	-.0657
(.8, .4743)	20	-.3651	-.0087	-.0697
	50	-.3645	-.0121	-.0640
	80	-.3854	-.0220	-.0701
	150	-.3770	-.0153	-.0652

Note. The population effect size values for d and delta are presented in brackets (d, delta). N is the sample size.

Figure 1. Biases of the means of each of the 1,000 replicated effect size measures (Cohen's d (d), Cohen's d_r (robust d), and Cliff's δ (delta)) compared to the associated population values across 32 simulated conditions



Appendix 1. R code used for the Monte Carlo simulation

```
library(orddom)
library(psych)
library(effsize)
library(boot)
labn <- '/Users/'
set.seed(38383388)
sd.t <- 1
sd.c <- 1
true.d.v <- c(0,0.2,0.5,0.8)
n.v <- c(20,50,80,150)
nor.v <- c(0,1) # 0 = normal; 1 = mixed normal

replications <- 1000

des.len <- length(true.d.v)*length(n.v)*length(nor.v)
save.results <- array(0, dim = c(des.len, 3))
save.sig.results <- array(0, dim = c(des.len, 8))
save.des <- array(0, dim=c(des.len,4))
results <- array(0, dim=c(replications,3))
sig.results <- array(0, dim=c(replications,8))

pos <- 1
for(nor.pos in 1:length(nor.v)){
nor <- nor.v[nor.pos]
for(pos.d in 1:length(true.d.v)){
true.d <- true.d.v[pos.d]
for(pos.n in 1:length(n.v)){
n1 <- n2 <- n.v[pos.n]
sig.results <- array(0, dim=c(replications,8))
t.mean <- true.d*sqrt(((n1-1)*sd.t^2 + (n2-1)*sd.c^2)/(n1+n2-2))

for(rep in 1:replications){
  treatment <- rnorm(n1,t.mean,sd.t)
  control <- rnorm(n2,0,sd.c)

  #mixed normal
  if(nor==1){
  draw <- sample(n1)[1:(n1*.1)]
  treatment[draw] <-  treatment[draw]*10
  control[draw] <-  control[draw]*10}
  dat <- array(0, dim=c((n1+n2),2))
  dat[(1:n1),1] <- 1
  dat[((n1+1):(n1+n2)),1] <- 0
  dat[(1:n1),2] <- treatment
```

```
  dat[((n1+1):(n1+n2)),2] <- control
  colnames(dat) <- c("G","X")

  #Cohen's d
  d.results <- cohen.d(treatment,control)[[3]]

  #robust d
  #sdw <- sqrt(((n1-1)*winsor.var(treatment, trim = 0.2) + (n2-1)*winsor.var(control, trim =
0.2))/(n1+n2-2))
  #dr.results <- .642*((mean(treatment, trim = 0.2)-mean(control, trim = 0.2))/sdw)
  boot.d.r <- function(dat){
  return(d.robust(as.data.frame(dat),"G",trim=.2)[[4]][[2]])}

  fc <- function(d, i){
    d2 <- d[i,]
    return(boot.d.r(d2))}

  bootdr <- boot(dat, fc, R=400)
  reci <- boot.ci(boot.out = bootdr,type=c("norm", "basic", "perc", "bca"))

  if(reci[[4]][2] > 0 | reci[[4]][3] < 0){sig.results[rep,3] <- 1}
  if(reci[[5]][4] > 0 | reci[[5]][5] < 0){sig.results[rep,4] <- 1}
  if(reci[[6]][4] > 0 | reci[[6]][5] < 0){sig.results[rep,5] <- 1}
  if(reci[[7]][4] > 0 | reci[[7]][5] < 0){sig.results[rep,6] <- 1}

  if(cohen.d(treatment,control)[[5]][[1]] > 0 | cohen.d(treatment,control)[[5]][[2]] < 0)
    {sig.results[rep,8] <- 1}

  #cliff's delts
  delta.results <- cliff.delta(treatment,control,return.dm=TRUE)

  #Welch's t test
  if(t.test(treatment,control)[[3]]<0.05){sig.results[rep,1] <- 1}

  #t test
  if(t.test(treatment,control,var.equal = TRUE)[[3]]<0.05){sig.results[rep,2] <- 1}

  #Wilcoxon test

  if(wilcox.test(treatment, control)[[3]]<0.05){sig.results[rep,7] <- 1}

  results[rep,1] <-   d.results
  results[rep,2] <- bootdr[[1]]
  results[rep,3] <-  delta.results[[1]]
}
```

```
for(i in 1:3){
save.results[pos,i] <- mean(results[,i]) }

for(j in 1:8){
  save.sig.results[pos,j] <- mean(sig.results[,j])}

save.des[pos,] <-c(nor, true.d, cohd2delta(true.d),n1)

pos <- pos+1
}}}# end of loop

colnames(save.results) <- c("d","robust_d","delta")
colnames(save.sig.results) <- c("Welch","t.test","normal","basic","perc","bca","Wilcoxon","d")
colnames(save.des) <- c("distribution","true_d","true_delta","n")

write.table(save.results,paste(labn,'EffResults.csv',sep=''),row.names=FALSE,sep=",")
write.table(save.sig.results,paste(labn,'EffSignResults.csv',sep=''),row.names=FALSE,sep=",")
write.table(save.des,paste(labn,'Design.csv',sep=''),row.names=FALSE,sep=",")
```