# NEGLIGIBLE EFFECT (EQUIVALENCE) TESTING BASED PROCEDURES FOR ASSESSING DISTRIBUTIONAL NORMALITY

LINDA SAWA DOROTA FARMUS

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PSYCHOLOGY
YORK UNIVERSITY
TORONTO, ON

MAY 2025

ABSTRACT

Researchers in psychology are often interested in evaluating whether a sample distribution is consistent with a normal (i.e., Gaussian) population distribution, most commonly to evaluate it as an assumption of a statistical model being adopted. In Study 1, a novel negligible effect (equivalence) test (NET) for normality is proposed that evaluates whether a sample distribution is similar enough to a normal distribution to be deemed equivalent (i.e., the difference between the sample and normal distributions is negligible). This NET establishes a negligible effect interval that quantifies coefficients for distribution shape that can be considered approximately normal. Any test statistic which has a 100(1-2α)% confidence interval (CI) that falls between the upper and lower limits of the negligible effect interval leads to the conclusion that the distribution is approximately normal. A series of simulations comparing Type I error and power rates of common traditional (difference-based) approaches (Kolmogorov-Smirnoff and Shapiro-Wilk tests) with the proposed NET-based approach was conducted. Results indicate that in small sample sizes the NET method has low power to detect normality, whereas the traditional methods have low power to detect nonnormality. However, the NET method almost never falsely concludes normality with nonnormal distributions and small samples. With large samples, traditional methods also often indicate that distributions are nonnormal even if the level of nonnormality is very minor (such that it would be unlikely to affect the validity of statistical tests or precision of parameter estimates). This limitation of traditional methods is not a concern for NET tests in large samples, since they rarely falsely conclude that a distribution is nonnormal when it shows minor deviations from normality.

One limitation of the NET-based approach is reduced power to detect normality when distributions are close to normal. Study 2 aimed to improve the calculation of CIs for the NET-based approach and examined alternative methods for computing bootstrap-based confidence

intervals for the NET-SW, including the stochastic bootstrap, parametric bootstrap, and Fisher's *r* to *z* transformation. The stochastic bootstrap approach had the best balance of Type I error to power rates and is the recommended approach to accompany the NET-based normality test.

# ACKNOWLEDGMENTS

Without the generous and enduring mentorship of Rob Cribbie, none of this would be possible. Rob, I am so very grateful for your unwavering support, which has shaped not only this thesis, but my growth as a researcher and graduate student. I always left our meetings with confidence and clarity. Thank you for being such a boundless source of support and expertise over the years—you have left a lasting impact that I will carry forward well into the future.

I am also most grateful to Dave Flora for helping to propel this work to completion – your thoughtful feedback encouraged me to see many facets of the topic in a new light. Over the course of my studies, your steady patience and sense of humour greatly enriched my experience of graduate school, and your advanced statistics classes were truly among my favourites –your lessons deepened my curiosity and appreciation for the subject and I always looked forward to the next.

My earnest thanks to Monique Herbert for providing helpful feedback on this work and challenging me to appreciate the greater implications of this research for teaching and how best to implement knowledge translation so that statistics students can benefit from it. Your pedagogical prowess has greatly inspired me as I've embarked on my own teaching adventures.

To Jodi Martin, my sincerest thanks for chairing the examining committee and raising the important issue of how the methods introduced in this work can be adopted by applied researchers beyond quantitative methods. I have also been so lucky to learn from your teaching expertise and ability to inspire enthusiasm in your students.

Next, I wish to express my gratitude to Johnson Li for serving as the external examiner and sharing reflective questions and insightful feedback, which I greatly valued given your expertise on topics related to this thesis. I am also indebted to Michael Rotondi for many helpful

suggestions and encouraging me to think about the application of this work beyond the social sciences, particularly in relation to emerging best practices in quantitative methodology. Your graduate course on meta-analysis was especially valuable and directly informed parts of the literature review in this work. Thanks are also extended to Phil Chalmers and Ji Yeh Choi for providing valuable suggestions on early drafts of this manuscript.

I am deeply appreciative to Christine Till – your mentorship on earlier research projects immensely impacted my learning and confidence as a graduate student. You continually exemplify rigorous research practices that are always tethered to a genuine curiosity. I also wish to thank Angela Eke at the Ontario Provincial Police for her mentorship and expertise during my internship—I was so privileged to be able to gain new skills required of methodologists working beyond academia. I am also so grateful to Chantal Arpin-Cribbie for her guidance and encouragement during my undergraduate years, which helped to nurture my academic interests and set me on the path to this research.

Of course, I cannot forget my Cribbie Lab mates, who are terrific models of collegiality and have helped to make graduate school such a unforgettable experience. Nataly, I am so thankful for experiencing the first years of graduate life together. Udi, Naomi, and Victoria, I really admire all your incredible research, and you have always been so great to work and hang out with.

I cannot have undertaken this journey without the steadfast support of my family. To my Mamo and Tato, you always believed in me and helped whenever and however you could, supplying countless pots of homemade soup and chocolate to the boys and I, or helping when I needed to meet important deadlines—your unconditional support has meant the world to me. I also wish to thank my brothers Mils and Dorian, my sister Gaja, and my sister-in-law Rumi, who

have always rooted for me and set great examples for me to strive towards—I greatly treasure the time that we spend together and with our kids. Watching them grow in tandem brings me immense joy and pride.

To my dearest buddy Tiff—since we met as undergraduate students, you have been with me every step of the way, never letting me give up, and always seeing the bright side of every situation. Having you in my life has been a tremendous gift and source of strength, much laughter, and endless adventure. To Mark, thank you for always being there for the boys and I, and a huge thanks to Noah—you always make me laugh and I treasure the many fun times we all have together and getting to watch you play hockey and baseball. To Liz, you have been so caring and an incredible friend to me, and my days are always better when I have the chance to talk to you.

I also wish to give thanks to my church community and the St. Patrick's Gregorian Choir. Your beautiful voices made me feel closer to heaven and were a constant source of peace throughout this journey.

To my late brother Jason and Babcia (Grandma) Basia— you often encouraged me to do well in school and when I was with you, I felt as though I was standing with an army. You carry me through every trial and triumph, and I miss you both dearly.

Lastly, I dedicate this work to my boys, Joey and Mitch. Your strength, courage, and patience make being your mom the greatest gift of all. I am so proud of you both and I believe in you. And just remember—never give up.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# CHAPTER ONE

## GENERAL INTRODUCTION

Variables within the social sciences often deviate from normality (Bauer & Sterba, 2011; Blanca et al., 2013; Micceri, 1989), with potentially substantial impacts on the validity of statistical test results. Usual methods to evaluate distributional normality include inspecting descriptive statistics (e.g., skewness and kurtosis) and graphs (e.g., histograms), or using a traditional null hypothesis test (e.g., Shapiro-Wilk test) to check for statistically significant deviations between the sample distribution and a theoretical normal distribution. When researchers are attempting to demonstrate that a distribution is normal, the objective of such tests is to *fail* to identify any significant differences between the sample distribution and a theoretical normal distribution. However, in such a scenario, the goal in checking normality should be to decide whether any deviation from normality is so small that it can be ignored.

Negligible effect testing (NET) offers a framework that aligns with this objective, necessitating that observed effects fall within a predetermined interval, called a negligible effect interval (NEI), in order to conclude that the effect can be discounted. For example, if one were trying to conclude that the difference in two sample means is negligible, a NET could be used to determine whether the difference fell within NEI = {-1, 1}; in this example, any mean difference less than one point in magnitude would be considered negligible.

This research adapted the NET framework to the problem of distributional normality, offering a contrasting approach to traditional normality tests. Specifically, the proposed NETs are used to seek evidence that a sample distribution is similar enough to a theoretical normal distribution, such that any differences can be deemed substantively trivial.

Study 1 presents two NET methods for testing normality, the first based on the Kolmogorov-Smirnov test (NET-KS) and the second based on the Shapiro-Wilk (NET-SW). Both adaptations are effective for minimizing false conclusions of normality, but the NET-SW was superior for detecting true instances of normality, especially in small to moderate sample sizes. However, the NET-SW is limited by conservative confidence intervals (CIs) when data are approximately normal (i.e., their estimates are systematically lower than expected, though the width is unaffected). These CIs are generated using percentile bootstrap methods and tend to be biased downward (often not including the observed sample statistic). This downward bias reduced the ability of the NET-SW to detect negligible differences between the sample distribution and a theoretical normal population distribution, particularly when sample distributions were in fact close to normal. In other words, when the sample distribution shape is near a theoretical normal distribution, we expect the NET-SW to conclude that differences are negligible. However, the downward bias in the CIs results in under-sensitivity, impairing the ability of the NET-SW to confirm negligible differences when they truly exist. The focus of Study 2 was refining the CI used in the NET-SW, exploring alternative CIs that better resolve biases in the original sampling distribution. The performance of alternative CIs was compared to the original percentile method, with an interest in how approaches such as stochastic bootstrapping for calculating CIs may improve the detection of negligible differences from normality.

A practical example is included to demonstrate how the improved NET-SW method can be used to confirm negligible differences in shape between a target distribution and a theoretical normal distribution, even when sample distributions are very close to normal.

# CHAPTER TWO

# NEGLIGIBLE EFFECT (EQUIVALENCE) TESTING BASED PROCEDURES FOR ASSESSING DISTRIBUTIONAL NORMALITY (STUDY 1)

Variables in the social sciences, education, and health are often nonnormally distributed (Bauer & Sterba, 2011; Blanca et al., 2013; Micceri, 1989). Many statistical models (Cribbie et al., 2012; Darlington & Hayes, 2017; Hau & Marsh, 2010; Mills et al., 2009; Stevens, 2009; Westfall & Henning, 2013; Wilcox, 1997; 2005; 2012a, b; Wilcox & Keselman, 2003; Yuan & Bentler, 1998) and effect size indices (e.g., Cohen's *d*; Li, 2016) in the social sciences are influenced by deviations from normality. Researchers often check the assumption of normality using graphs or by conducting null hypothesis significance tests (NHST) of observed data or model residuals. For these NHST approaches, the effect of interest is the *difference* between a normal (i.e., Gaussian) population distribution and the sample distribution of interest.

However, in most situations, the appropriate goal when investigating normality is to obtain evidence that a sample distribution is *negligibly different* from a normal distribution, such that any detected differences are trivial, a goal that can be met within the framework of negligible effect (equivalence) testing (NET). Study 1 is organized in the following manner. First, the impact of nonnormality on statistical inferences, effect size estimates, and so on is outlined. Second, existing tests for normality that are popular or have been found to have good statistical properties are reviewed. Next, important limitations with conventional approaches to evaluating normality are discussed. Fourth, novel NET-based procedures for assessing normality are proposed. Fifth, a simulation study to compare the performance of the proposed NET-based normality tests to traditional tests for normality is conducted. Last, an example demonstrating the application of the novel NET tests is presented.

A normal, or Gaussian distribution, is a probability distribution characterized by symmetry around the mean, forming a bell-shaped curve. Most observations cluster near the mean, and probabilities decrease as the distance from the mean increases in either direction. The mean, median, and mode are equal in a normal distribution, and two parameters, the mean (μ) and standard deviation (σ) describe the central tendency and variability around the mean, respectively. In a perfectly normal distribution, skewness is absent. Kurtosis describes the shape of a distribution's peak and tails; a normal distribution has a moderate peak and tails (i.e., mesokurtic). When a distribution deviates from normality, higher kurtosis (leptokurtic) is associated with a sharper peak and thicker tails, while lower kurtosis (platykurtic) corresponds to a flatter peak and thinner tails. The formula for the probability density function (PDF) for a normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

where $x$ represents the random variable, μ is the mean, and σ is the standard deviation.
The normal distribution has several favourable properties: 1) symmetry around the mean results in equal probabilities of observations below and above the mean; 2) the mean and standard deviation (SD) have straightforward interpretations. As the normal distribution is symmetric around the mean, about 68% of data falls within one SD of the mean, about 95% within two SDs from the mean, and about 99% within three SDs. For other continuous distributions, the mean still represents the arithmetic average or center of the distribution, but the interpretation can be less intuitive if skewness or outliers are present. Similarly, SD may not convey spread as effectively in skewed distributions if the data is concentrated more heavily in a single tail; and 3) many phenomena tend to follow a normal form (e.g., heights among men or women within a specific region or ethnicity, IQ scores, standardized test scores), making it a valuable distribution

for many fields, including psychology. The central limit theorem posits that the sampling distribution of the sample mean approaches normality as the sample size increases, regardless of the shape of the population distribution, and this property facilitates statistical inference about population parameters based on sample statistics using the normal distribution. However, as described in the next section, most phenomena in psychology do not follow a normal form, and research has quantified this prevalence.

**Prevalence of Nonnormality**

Micceri (1989) examined 440 samples of psychometric and achievement measures and found that most were moderately to severely nonnormal; this review led him to memorably compare a normal distribution to a unicorn. Similarly, Blanca et al. (2013) concluded that variables in only 5% of 693 psychology studies were approximately normal and most (74%) mildly to moderately deviated from a normal distribution. Another systematic review of the social sciences, health, and educational fields between 2010 and 2015 found that, across 262 empirical articles, distributions commonly resembled gamma ($n$ = 57), negative binomial ($n$ = 51), multinomial ($n$ = 36), binomial ($n$ = 33), lognormal ($n$ = 29), and exponential ($n$ = 20) population distributions (Bono et al., 2017). These results are not surprising, since it is well known that many variables in psychology are inherently nonnormal (e.g., level of stress, depression, or other psychopathologies in the general population). Due to the prevalence of nonnormality, model or test assumptions may not be satisfied (e.g., errors from statistical models may be nonnormally distributed if the outcome variable or predictor variables are nonnormal). While minor deviations from normality may not have serious implications for inferences, more extreme violations could be problematic. Violations of normality may also impact effect sizes and other statistics that are influenced by the shape of variable distributions (Keselman et al.,

2008), such as pooled estimates from meta-analyses (Sun & Cheung, 2020). Thus, researchers should often consider the extent of deviation from normality and the effect it has on relevant statistical tests. Specifically, the standard error of a statistic is intrinsically linked to the probability model that describes the sampling distribution of the statistic. Standard error measures the variability of a statistic across hypothetical samples. For example, the standard error of the sample mean is $\sigma/\sqrt{n}$, where $\sigma$ is the population standard deviation and $n$ is the sample size when the hypothetical sampling distribution is a normal probability distribution. But the standard error for a proportion assumes a binomial distribution and is calculated with $\sqrt{p(1-p)/n}$, where $p$ is the sample proportion.

Nonnormality can distort descriptive statistics (e.g., mean, standard deviation), complicating meaningful interpretation. The mean is particularly sensitive to skewness and outliers, making it less reflective of central tendency. Similarly, SD, which measures the average distance of each data point from the mean, works well for symmetric distributions where data are evenly spread. However, in skewed distributions or distributions with outliers, the SD may be magnified, overrepresenting extreme cases. As Curran-Everett et al. (1998) explain, "When the tails of a distribution are elongated…. the sample standard deviation will be an inflated measure of variability in the population." (p. 778, citing Mosteller & Tukey, 1997; Snedecor & Cochran, 1980). In such cases, data might cluster on one side of the distributions, while distant outliers stretch out the distribution, leading to an SD that may not accurately represent the spread of the bulk of the data. Consequently, the SD may not be the best measure of variability in a sample distribution that is asymmetric or has outliers.

**Robustness of General Linear Models to Nonnormality**

When model errors in general linear models (GLMs; for instance, models underlying the *t* test for independent means, analysis of variance [ANOVA], and linear regression) are nonnormal, standard errors of the parameter estimates are inaccurate. Thus, any procedure that is a function of the standard errors will also be inaccurate, namely, CIs and significance tests. When nonnormality is combined with heterogenous variance or unbalanced sample sizes, this inaccuracy of standard errors is exacerbated. Although normality is not a necessary requisite if the only research goal is to estimate GLM parameters, it is important for inferences to the population of interest (using traditional parametric inference) via CI estimation or significance testing. Nonnormality can create inflated Type I and Type II error rates (Cribbie et al., 2012; Wilcox, 2012a). Inference with general linear models has several assumptions, including normality (e.g., conditional/marginal normality of population-level errors in linear models), which means that observed sample-level distributions of residuals should be consistent with a *t*-distribution (Fox, 2015). In practical terms, this means that the sample data within each group in a *t* test or cell in an ANOVA should be consistent with a normal population distribution. Formally, this reflects the assumption that the conditional distribution of the outcome variable *Y* given predictors *X* is normal. If the data within any group or cell is nonnormal, then model errors are nonnormal, thereby violating the assumption of normality.

While some simulations have shown that these tests are robust to mild to moderate deviations from normality (Glass et al., 1972; Harwell et al., 1992; Lix et al., 1996), nonnormality can lead to incorrect inference (Cribbie et al., 2012; Wilcox, 1997; 2005; 2012a, b; Wilcox & Keselman, 2003). When nonnormality is paired with variance heterogeneity, the problems are exacerbated. For example, Mills et al. (2009) showed that when distributions are

moderately skewed and variances are unequal, the ANOVA *F* statistic has inflated Type I error rates when within-cell sample sizes and variances are negatively paired (i.e., larger samples are paired with smaller variances and smaller samples are paired with larger variances), and deflated Type I error rates when sample sizes and variances are positively paired (i.e., larger sample sizes are paired with larger variances and smaller samples are paired with smaller variances), for both main effects and interactions. Empirical Type I error rates may be as low as 1.8% or as high as 14% when variances are unequal and are coupled with non-normal distributions (assuming a nominal Type I error rate of $\alpha = .05$). Similarly, Cribbie et al. (2007) found that when distributions are skewed and variances are negatively paired in a one-way ANOVA setting, even the generally robust Welch test could not always maintain acceptable Type I error rates. The Welch test had Type I error rates that approached 17% with nonnormal distributions and unequal variances.

Violation of normality has the most adverse impact on inferences drawn from traditional statistical tests (e.g., *t* test) when samples are relatively small (Lumley et al., 2002). For instance, Boneau (1960) found that when sampling from identically shaped exponential distributions with equal variances, small samples led to a Type I error rate below the nominal rate. When sampling from two different distributions (one normal and one exponential) of $n = 15$ and with unequal variances ($\sigma_1^2 = 1$, $\sigma_2^2 = 4$), Type I error rates were around 12%. Algina et al. (1994) found Type I error rates upwards of 20% in lognormal distributions in small samples. Delaney (2000) found that when two distributions are oppositely skewed, the Type I error rate of the *t* test with homogeneous variances can be higher than 8% with a two-tailed test, and higher than 11% with a one-tailed test. Furthermore, with two distributions of opposite skewness, the Q test of study heterogeneity in meta-analysis has inflated Type I error rates, upwards of 50% (Sun & Cheung,

2020). Wilcox (1997) found that with equal variances, means, and group sample sizes of $n = 21$, simulations indicate Type I error rates of around 13% with different distribution shapes (e.g., two lognormal distributions and two normal distributions in a one-way ANOVA setting, or stated differently, the conditional distributions of the outcome for the groups on the factor variable).

Violation of the normality of model errors assumption impacts standard errors, producing CIs of coefficients that do not properly cover the population effect of interest (e.g., are either too wide or too narrow; Adkins, 2017). CIs for Pearson correlations using the Fisher *z* method may also be inaccurate with nonnormal data (Bishara & Hittner, 2017), leading to inflated or deflated coverage when kurtosis is large, even if skewness is zero (Puth et al., 2014). Although many methodologists have deemed the assumption of normality of model errors irrelevant in large enough samples because of the central limit theorem (CLT; Fox, 1991; Knief & Forstmeier, 2021), including many statistics textbook authors (e.g., Field et al., 2016), what constitutes a 'large enough' sample depends on many factors (Pek et al., 2018). For example, Pek et al. (2018) indicate that nonnormality is impactful in larger samples when the degree of skewness in errors is extreme. For example, when skewness is about 7, even a sample size of 100 is insufficient (Pek et al., 2017). The required sample size increases more when extreme skewness is combined with multiple predictors.

More complex models (e.g., multivariate models, structural equation modeling [SEM], and multilevel modeling [MLM]) are also affected by nonnormality (Hau & Marsh, 2010; Stevens, 2009; Westfall & Henning, 2013; Yuan & Bentler, 1998). For example, multilevel models are sensitive to compounded nonnormality in model residuals, that is, violations of normality that occur at both individual and cluster levels. Man et al. (2022) conducted a simulation study that manipulated normality of data in a multilevel setting, finding that

parameter estimates can be biased, with random effect estimates being more distorted as a result of nonnormal residuals at both levels, compared to fixed effect estimates. This distortion is exacerbated if cluster size is small or when the cluster sizes are large, but the total number of clusters is small. In this case, parameter estimates become more variable, with standard errors increasing, affecting accuracy of inferences. In short, nonnormality increases the likelihood that inferences pertaining to population effects are inaccurate (Darlington & Hayes, 2017).

**Standardized Effect Sizes**

Nonnormality, especially in heavily skewed distributions, can impact the interpretation of Cohen's *d*. When sample data for both groups are approximately normal, *d* provides a meaningful comparison of the distance between means relative to group variability, since the SD captures approximate symmetry in variability. However, if one distribution is nonnormal, even if the pooled SD is the same, *d* becomes less meaningful. The mean is often influenced by extreme cases, which can cause *d* to overestimate the practical difference between groups. Similarly, the SD may be inflated by outliers, further distorting the measure. While Cohen's *d* does not require normality, it assumes that means and SDs are accurate representations of most observations in the sample. Thus, the interpretability and consistency of Cohen's *d* as a meaningful measure of effect size can be impacted by changes in the underlying distributions. Other standardized effect size measures that rely on means and SDs and which may have interpretations distorted by nonnormality include Hedges' *g* and Glass' Δ.

Nonnormality can also lead to problems when aggregating data for meta-analyses of standardized mean differences with nonnormal data. For example, Sun and Cheung (2020) simulated primary studies and found that when distributions are skewed in opposite directions in independent groups designs, the point estimates of the pooled effect sizes as well as their CIs

were biased relative to when distributions were not oppositely skewed or roughly normal. Moreover, these issues persist when the number of studies meta-analyzed was large and the included studies have a large sample size, meaning that meta-analysis was not robust to the effects of nonnormality. Marfo and Okyere (2019) conducted a simulation study to examine the accuracy of different effect size estimates in meta-analysis, generating random data from normal and contaminated normal distributions to mimic primary studies. They found that under contaminated normal distributions with equal variances, the Probability of Superiority (a non-parametric method) was the most accurate effect-size estimate, with Cohen's *d*, Hedges' *g*, or Glass' $\triangle$ less accurate when normality was violated in primary studies. CIs around *d* may also be biased by nonnormality (Kelley, 2005), especially when combined with unequal variances (Chen & Peng, 2015). Although there are alternatives which are robust to violations of these assumptions (e.g., scaled robust *d* and nonparametric estimators for the common language effect size; Li, 2016; Keselman, 2008), Cohen's *d* is most commonly used (Farmus et al., 2022; Sun et al., 2010).

Generally, meeting normality is important for constructing CIs or conducting significance tests on model parameter estimates, including effect size indices. Thus, appropriate and well-behaving procedures for detecting normality are much needed.

**Traditional Methods to Detect Nonnormality**

Normality is commonly assessed by inspecting graphs of sample data. Examples of graphs include histograms, normal quantile-quantile (QQ) plots, kernel density plots, and violin plots. Researchers also often utilize formal statistical tests, such as the Kolmogorov-Smirnov or Shapiro-Wilk tests (Oztuna et al., 2006). The subjective judgment of graphs by behavioural researchers is often erroneous (Altman & Bland, 1995; Peebles & Ali, 2015; Woller-Carter et al.,

2012; Xiong et al., 2020) and mechanical decisions often outperform human judgment (Meehl, 1954; Swets et al., 2000). Formal tests assess for statistically significant differences between the distribution of interest and a normal distribution.

There are many traditional null-hypothesis based tests for evaluating nonnormality, including the Kolmogorov-Smirnoff (KS; Kolmogorov, 1933; Smirnov, 1948), Anderson-Darling (AD; Anderson & Darling, 1952, Anderson, 1962), Shapiro-Wilk (SW; Shapiro & Wilk, 1965), Jarque-Bera (Jarque & Bera, 1980), D'Agostino (D'Agostino, (1970), Lilliefors corrected KS (Lilliefors, 1967), Anscombe-Glynne (Anscombe & Glynn, 1983), Cramer-von Mises (Cramer, 1928; von Mises, 1931), and the D'Agostino-Pearson (D'Agostino & Pearson, 1973) omnibus tests. Many studies have compared the performance of these tests, with the KS, Anderson-Darling, and SW tests often recommended as powerful and robust (Ghasemi & Zahedias, 2012; Thadewald & Büning, 2007; Yazici & Yolacan, 2006), and widely adopted in research, in part because they are readily available in most statistical software packages (e.g., SW and KS are commonly included in the base R packages and SPSS, SAS, etc.).

Yap and Sim (2011) used simulation to compare power and Type I error rates between the KS (and its modifications, the AD test and Lilliefors test), SW, Cramer-von Mises test, D'Agostino Pearson, chi-squared test, and Jarque-Bera tests. For symmetric, short-tailed distributions, the D'Agostino and SW tests had the best power rates, while the KS and chi-square tests had suboptimal power rates, which supports similar findings of D'Agostino et al. (1990). For symmetric long-tailed distributions, power rates of the Jarque–Bera and D'Agostino tests were comparable to those of the SW test. In asymmetric distributions, the SW test was the most powerful, followed by the AD test, which supports similar findings of Oztuna et al. (2006). Razali and Yap (2011) similarly found that the SW has the best power, followed by the AD,

Lilliefors, and KS test, but all tests perform poorly in small sample sizes. For instance, for an asymmetric gamma distribution with skewness = 1.00 and kurtosis = 4.5, the SW detects nonnormality at a rate of at least 95% at around $N = 100$, the AD at $N = 200$, the Lilliefors at $N = 300$, and the KS at $N = 1000$. However, for more severe deviations from normality (skewness = 2.00, kurtosis = 9.00), the differences in performance narrow, with the SW and AD reaching power = 100% at $N = 50$ while the KS and Lilliefors reaching full power at $N = 100$. However, in samples under $N = 50$, all tests perform poorly. Although the AD and Lilliefors tests are based on the KS test, only the original one-sample KS test is included here given the novelty of the tests to be proposed. To summarize, given their popularity and ease of access (i.e., statistical software), this paper focuses on the KS and SW normality tests.

**One-sample Kolmogorov-Smirnoff**

The KS test is a nonparametric test for continuous data that produces a *D* statistic reflecting the maximum absolute vertical difference in cumulative probability between an empirical cumulative distribution function (CDF) and a specified theoretical CDF. The one-sample KS test compares a sample distribution to a given theoretical distribution and is sensitive to differences in central tendency, variance, and shape. The one-sample KS test may be used to determine whether a sample distribution is drawn from a normal, log-normal, Weibull, exponential, or logistic family; however, it is most often used to evaluate normality (which is the form of the test studied below).

With the single-sample KS test, the null hypothesis ($H_0$) states that the population distribution from which the observations were sampled is normal (more generally, $H_0$ and $H_a$ refer to any reference distribution, but as stated above, the aim here is assessing normality). Observations in the sample distribution are ordered from smallest to largest, $Y_1, Y_2, \ldots Y_N$. The

sample CDF is $Fe(x) = n(i)/N$, where $n(i)$ is the number of points that are less than the ordered $Y_i$ value. This piecewise step function increases by $1/N$ at each value of the ordered observations. In contrast, the reference (i.e., theoretical) CDF does not increase in a stepwise function, but monotonically increases based on the shape of the theoretical distribution.

The *D* statistic reflects the maximum absolute difference in cumulative probability between the CDF of the empirical data [$Fe(x)$] and the CDF of the theoretical distribution [$F(x)$], and thus smaller values suggest greater similarity between the distributions. The KS test statistic can be expressed as:

$$\widehat{D} = max\ \Delta = \ max\ [Fe(x) - F(x)]\ .$$

$\widehat{D}$ is compared to a critical *D* value that is a function of sample size and the nominal α level, and if $\widehat{D}$ is greater than the critical *D* value, the null hypothesis is rejected.

One limitation of the KS test when exploring normality is its greater sensitivity at the center of the distribution compared to its tails (Goldman & Kaplan, 2017) and as explained later, the values of *D* are related to the sample size, such that larger sample sizes result in smaller *D* values.

**Shapiro-Wilk**

The SW test assesses whether a sample distribution is consistent with a normal population distribution using an adjusted correlation between the observed scores from an empirical distribution and ordered theoretical scores from a normal distribution. The null hypothesis for the SW procedure is that the sample is drawn from a normal population distribution, whereas the alternate is that the population distribution is not normal. Proposed by Shapiro and Wilk (1965), the SW test calculates a *W* ratio (a pseudo correlation) that reflects a weighted sum of squares of the differences between the sample datapoints and their expected

values under normality, or the association between the ordered data in the sample and the same length vector of values from a theoretical normal distribution. Lower values suggest a deviation from normality, whereas values near 1 suggest distributions close to normal. In other words, if the distribution is normal, the association would be linear and hence *W* would be close to 1; however, as the distribution becomes more nonnormal, the correlation is reduced and the *W* statistic decreases. The *W* is not a true correlation because it does not fall within the standard range of -1 and 1 and is not symmetric (i.e., a correlation is symmetric in the sense that the correlation between X and Y is the same as the correlation between Y and X).

The SW test determines whether a random sample, $x_i$, $i = 1, 2, \ldots, n$, is sampled from a normal distribution where $X \sim N(\mu, \sigma^2)$. The *W* statistic can be expressed in the following manner for data organized in ascending order (i.e., $x_{(i)}$, $i = 1, \ldots n$, $x_{(i)} \leq x_{(i+1)}$):

$$\widehat{W} = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

$a_i$ are constants derived by:

$$\left(a_1, a_{2,\ \ldots}, a_n\right) = \frac{m^T V^{-1}}{(m^T V^{-1} m)^{1/2}}$$

where $m = \left(m_{1,} m_{2,\ \ldots,}\ m_n\right)^T$ are expected values of the ordered statistics that are independent and identically distributed normal random variables, while *V* is the covariance matrix of the order statistic. The numerator is a sum that is like a covariance and the denominator is a measure of total variability.

Critical values are used to determine whether the sample *W* is statistically significant, which are a function of both the sample size and the nominal significance level (e.g., $\alpha = .05$). Shapiro and Wilk (1965) did not derive critical values from a standard probability distribution

(normal or *t* distribution), but empirically through simulation to understand how *W* behaves under normal data and to develop a reference distribution that could be used to assess normality of sample data in practice. For each *N*, many independent random samples (with replacement) were drawn from a theoretical normal distribution. *W* was calculated on each sample to build an empirical distribution of *W* values, from which critical values were derived. For a given *N* and significance level α, the critical value demarcates the value below which the proportion α of the simulated *W* statistics fall. For instance, if the α = .05, the critical value of *W* corresponds to the value for *W* at the $5^{th}$ percentile of the empirical distribution of *W* for the particular *N*.

Shapiro and Wilk (1965) developed critical values for samples ranging from $N = 3$ to $N = 50$, with subsequent work tabulating critical values for larger sample sizes (Shapiro & Francia, 1972; Royston, 1982; Royston, 1989). However, critical values for every *N* up to 5000 have not been tabulated. Instead, statistical software often interpolates approximations of critical values in larger sample sizes. If the observed *W* is less than the critical value for the given *N* and chosen significance level, then the null hypothesis that the distribution is normal will be rejected.

**Visualizing *D* and *W***

To give readers a sense of the magnitude of *D* and *W* values, continuous distributions with standardized values ($M = 0$, $SD = 1$) can be plotted against a theoretical standard normal distribution. In Figure 1a, the probability density functions (PDFs) of Student's *t*-distributions with different degrees of freedom are compared to the PDF of a standard normal distribution. While all distributions are symmetric, the peak of the distribution is highest for the normal and lowest for the *t* distribution with $df = 1$, while the tails are thinnest for the normal and thickest for *t* distribution with $df = 1$ (i.e., with decreasing *df*, kurtosis of the *t* distribution increases, leading to heavier tails).

For a normal distribution, $W = 1$ and $D = 0$, whereas for a $t$ distribution with $df = 1$, $W = .08$ and $D = .13$. Figure 1b shows these same $t$ distributions but compares the differences between the CDFs of each $t$ distribution to the CDF of a standard normal distribution (i.e., subtracting the quantiles' probabilities from each distribution from the quantiles' probabilities of the standard normal CDF). This difference is shown on the y-axis, while the vertical distance between the quantiles of the sample data and the quantiles of the standard normal distribution are shown at each point along the x-axis. In other words, this difference illustrates how the cumulative probabilities of the sample data deviate from those of the standard normal distribution across different quantiles, with positive values on the *y*-axis indicating that the sample data's CDF is larger than the CDF of the standard normal distribution at that quantile, whereas negative values indicate the opposite. For the roughly normal sample compared with the standard normal distribution, this is a flat line since there are no differences between the CDF of the sample and that of the standard normal distribution. Figure 2a plots the PDFs of $\chi^2$ distributions with different degrees of freedom and compares them to the PDF of a standard normal distribution. The peak of the distributions is lowest for the normal and highest for the $\chi^2$ distribution with $df = 3$, which also has the greatest skewness. For the $\chi^2$ ($df = 3$) distribution, $W = .87$ and $D = .11$. Figure 2b shows these same $\chi^2$ distributions, comparing the differences between the CDF of each $\chi^2$ distribution and the CDF of a standard normal distribution (on the y-axis). Similar to the $t$ distributions, positive values on the *y*-axis indicate that the data from the $\chi^2$ distributions CDF is larger than the CDF of the standard normal distribution at a particular quantile. For more skewed $\chi^2$ distributions (smaller degrees of freedom), larger differences are observed between the CDFs of the $\chi^2$ distribution and the CDF of a standard normal distribution, but these differences become smaller with increasing degrees of freedom.

**Issues with Traditional Tests of Normality**

With traditional tests of normality (e.g., KS, SW), the null hypothesis states that the sample distribution is drawn from a normally distributed population. For example, $H_0$: $\lambda_P \sim N(\mu, \sigma)$, where $\lambda_P$ represents the population distribution from which the sample is drawn. The alternative hypothesis states that the population distribution deviates from a normal distribution; e.g., $H_1$: $\lambda_P \not\sim N(\mu, \sigma)$, where $\not\sim$ indicates "not distributed as". Hence, a significant *p* value is interpreted as evidence that the distribution is not normally distributed (i.e., the data do not follow a theoretical normal distribution). This orientation of null versus alternative hypotheses is problematic for concluding that a distribution is normal because it seeks evidence that a sample distribution is significantly different from normal, when the researcher's goal is (usually) in seeking evidence that a distribution is only negligibly different from normal. Further, it can be challenging to relate quantitative indications of nonnormality (e.g., *p* values) to the implications of those violations on statistical inference, particularly when the statistics measuring normality (e.g., *p* values) are largely a function of sample size.

While these tests seek evidence for significant *differences*, the conceptual goal of meeting normality is usually to find evidence that a sample distribution is *similar* enough to a normal distribution to conclude that the sample plausibly could have been drawn from a normal population. Another way to say this is that the sample distribution might be so similar to a normal distribution that it would likely not have any impact on subsequent statistical inference. Therefore, as outlined above, researchers using traditional tests with small sample sizes may erroneously conclude that a failure to find evidence that the distributions are nonnormal implies that they are normal, whereas, in larger samples, a conclusion of nonnormality would be more likely since the test would be sensitive to even minor deviations from normal.

**Negligible Effect Testing (NET)**

Given these arguments, an appropriate framework to assess the assumption of normality might be to use a NET, which could test whether differences between a theoretical normal distribution and the distribution of interest are so small (or meaningless) that the sample distribution of interest can be considered approximately normal. In other words, NETs are designed to assess whether a difference or association is negligible; the interest in this study is in whether a population distribution from which the observed sample is drawn is negligibly different from a theoretical normal distribution. Within clinical research, NETs are often used to demonstrate the equivalence of different treatments. For instance, NETs have been used to demonstrate that a less costly treatment leads to outcomes that are negligibly different from those of a more costly treatment (Bower & Gilbody, 2005; Hill et al., 2014), that generic formulations of medications for depression or anxiety are equivalent to branded formulations (Ferguson & Clapshaw, 2020; Kharasch et al., 2019; Zhai et al., 2020) or ADHD (Sikes et al., 2017), and to establish that a negligible difference exists between psychological therapy and pharmacotherapy for depression (Dimidjian et al., 2006; Steinert et al., 2017). One source of confusion is the difference between NETs and noninferiority tests, where the latter seeks to establish that, for example, new therapies are equivalent *or superior* to a current therapy (i.e., that they are not inferior; Walker & Nowacki, 2011) and are thus a one-sided variation of the traditional NET.

Briefly, the roles of the null and alternative hypotheses are reversed in NET compared to traditional NHST. Using the basic example of comparing the means of a continuous outcome across two groups, traditional hypothesis testing examines the null hypothesis of no difference between population means (e.g., $H_0$: $\mu_1 = \mu_2$). Failure to reject the null hypothesis implies that there is not enough evidence to conclude that the population means differ. Within an NET

framework, the null hypothesis would state that there is a meaningful difference between the population means, whereas the alternative hypothesis would state that any difference is not meaningful.

The first step with NETs is to choose the smallest degree of association that is of practical significance, herein referred to as the minimally meaningful effect size (MMES; Beribisky et al., 2019; also referred to as the smallest effect size of interest, SESOI). The MMES can be represented by negligible effect bounds (NEBs), denoted using $\{-\delta, \delta\}$; these boundaries represent the largest effects (in the units of the measure), in either direction, from the desired effect (e.g., mean difference = 0) that would still be deemed negligible. The boundaries are usually symmetric but need not be. Any observed effect that falls within the range of the NEBs would be considered negligible.

A common NET strategy is the two-one-sided tests (TOST) approach (Schuirmann, 1987). When adopting the TOST approach for evaluating equivalence, the equivalence null hypothesis is expressed via two inequalities. One states that the observed effect of interest ($\varphi$) is equal to or less than the lower equivalence bound:

$$H_{01}: \varphi \leq -\delta,$$

or that the effect is equal to or larger than the stated upper bound:

$$H_{02}: \varphi \geq \delta.$$

The alternative hypothesis states that the effect is larger in magnitude than the lower bound:

$$H_{11}: \varphi > -\delta$$

and the effect is smaller than the stated upper bound:

$$H_{12}: \varphi < \delta.$$

There are two ways to conclude equivalence or a negligible association. The first method uses hypothesis testing and requires rejecting both null hypotheses, indicating that the effect is significantly greater than -δ (the lower bound) and less than δ (upper bound). Alternatively, a negligible association can be concluded if the $[(1–2\alpha) \times 100]$% CI for the effect falls within the specified bounds $\{-\delta, \delta\}$. In this case, the alternative hypothesis(es) are retained at the given α level. In difference-based NHST, a Type I error relates to incorrectly rejecting the null hypothesis and concluding that a non-zero effect (i.e., difference/association) exists, while statistical power relates to correctly rejecting the null hypothesis. With NETs, a Type I error is incorrectly rejecting the null hypothesis and concluding that the effect is negligible, whereas power is the probability of correctly detecting an effect small enough to be deemed negligible.

Although equivalence tests have been developed for many research contexts, including model fit in regression (Alter & Counsell, 2023), structural equation modeling (Beribisky & Cribbie, 2023; Marcoulides & Yuan, 2017); measurement invariance (Counsell et al., 2020; Yuan & Chan, 2016), comparing correlation/regression coefficients (Counsell & Cribbie, 2015), correlation (Goertzen & Cribbie, 2010), variance homogeneity (Mara & Cribbie, 2018), mediation (Beribisky et al., 2020), negligible interaction (Counsell et al., 2016; Wellek, 2010), and categorical data (Shishkina et al., 2018; Wellek, 2010), I am not aware of any tests for the assumption of normality within an NET framework. Procedures have been developed to evaluate equivalence of two empirical distributions, but these have not focused on testing whether an empirical sample is equivalent to a theoretical normal distribution. Specifically, Ostrovski (2022) developed an equivalence test for the Cramér-von Mises distance to compare two empirical distributions and as well as a modification based on the Anderson-Darling test (Ostrovski, 2023).

## Proposed Negligible Effect Based Tests of Normality

The primary goal of this study was to develop a novel NET for assessing normality, which would evaluate whether the population distribution from which an observed sample is drawn can be considered equivalent to a normal distribution. The basic premise is that if a [(1–2α) × 100]% CI for an effect (i.e., some measure of the extent to which a sample distribution deviates from a normal distribution) falls within the NEBs (where the bounds specify how much a distribution can deviate from normal and still be considered 'practically normal'), then the distribution is concluded to be equivalent to normal. The null hypothesis of the NET states that the distribution deviates meaningfully from a normal distribution, while the alternative hypothesis specifies that the distribution does not deviate meaningfully from a normal distribution. Thus, the NEB is based on an MMES that represents the smallest deviation from normality that would still be meaningful. For example, if a researcher was using this test to determine whether they should move to a robust or non-parametric model in the event that the assumption of normality is untenable, then {-δ, δ} should be set such that any deviation from normality within the bounds would have minimal effect on traditional (non-robust) test statistics (e.g., Type I error, power). The NETs developed here are based on the traditional KS and SW tests because they have been found to have a superior balance of Type I error and power rates within simulation research compared with other tests of normality (Mendes & Pala, 2003; Mohd Razali & Bee Wah, 2011; Öztuna et al., 2006; Thadewald & Büning, 2007; Yap & Sim, 2011). The accuracy (as represented by power and Type I error control) of the proposed NET will be compared to the corresponding conventional difference-based tests for normality, since researchers often use traditional difference-based tests to provide evidence that a distribution is normal (i.e., wanting to not reject the null hypothesis).

**Shapiro-Wilk Negligible Effect Test (NET-SW)**

Recall that the *W* test provides a ratio, or pseudo correlation, to reflect the association between the ordered scores from the sample distribution and the ordered scores from a theoretical normal distribution. Values closer to 1 are indicative of a linear association, and hence a normal distribution. Given the nature of the *W* statistic (always positive), the boundaries of the NET-SW are one-tailed; the research hypothesis (i.e., normality) states that the effect falls above the lower bound of the NEB (ζ). The MMES for the NET-SW is chosen to reflect the largest *W* value for which the observed distribution deviates meaningfully from a normal distribution. Thus, the negligible effect interval for the NET-SW is represented by a pair of bounds {ζ, 1}, in which the upper bound of 1 represents optimal normality (i.e., no deviation from normality) and the lower bound, ζ, represents the largest value of the SW statistic that deviates meaningfully from normality. Any effect that is within the range of the lower and upper bounds is deemed approximately normal. Since only the lower end of the NEB and the lower end of the CI for the point estimate of the effect would be pertinent, the lower end of the NEB retains a Type I error rate of α. Since the critical comparison is whether the lower end of the CI for the *W* is greater than ζ, there is no need to guard against the possibility that the upper bound of the CI is too close to 1 since the deviation can only go in one direction, and therefore it is impossible to commit a Type I error in the upper tail. The Type I error rate is controlled at only the lower bound, meaning that the probability of falsely concluding that the distribution is normal (when it is not) is kept at α and the rejection region is only in the lower tail.

Therefore, the null hypothesis for the NET-SW test states that the population *W* parameter, $W_{pop}$, is equal to or less than the MMES, ζ (i.e., the degree of nonnormality is non-negligible). The alternative hypothesis for the NET-SW test states that the population *W*

parameter is larger than the MMES (i.e, the degree of nonnormality is negligible). To summarize:

$$H_0: W_{pop} \leq \zeta$$

$$H_1: W_{pop} > \zeta$$

If the lower bound of the [(1–2α) × 100]% CI for the observed *W* statistic is greater than the MMES (ζ), then the null hypothesis of meaningful nonnormality is rejected in favour of the alternative hypothesis that there is a negligible difference between the population distribution and a theoretical normal distribution (Figure 3a). The CI for the *W* test statistic is computed using bootstrapping. Specifically, many (e.g., 5000) bootstrap samples are drawn from the sample distribution with replacement and the *W* statistic is computed on each bootstrap sample. The α quantile of the distribution of *W* statistics across all bootstrap samples defines the lower bound of the bootstrapped CI for the test statistic (i.e., the percentile bootstrap method).

**Kolmogorov-Smirnov Negligible Effect Test (NET-KS)**

Recall that the KS test produces a *D* statistic to reflect the maximum difference between a theoretical normal distribution and the distribution of interest. A *D* statistic close to 0 indicates very minimal difference in shape between the two distributions and hence leads to the conclusion that the sample distribution is approximately normal. The MMES for the NET-KS test represents the smallest *D* statistic that reflects a meaningful deviation from a normal distribution. The thresholds for the NET-KS test of normality are represented by a pair of bounds $\{0, \theta\}$, in which the lower bound of 0 represents optimal normality (i.e., no deviation from normality) and the upper bound, θ, represents the smallest value of *D* that deviates meaningfully from normality. Any effect that falls within the range of the lower and upper bound is deemed equivalent to normal. Thus, like the NET-SW test interval, the NET-KS normality test is a one-tailed test (i.e.,

since only the upper end of the NEB and the upper end of the CI for the point estimate of the effect would be pertinent, one end of the NEB retains a Type I error rate of α).

The null and alternative hypotheses for the NET-KS are:

$H_0$: The population distribution is non-negligibly different from a theoretical normal distribution.

$H_1$: The population distribution is negligibly different from a theoretical normal distribution.

To test these hypotheses, the sample *D* statistic is compared to θ. If the upper bound of the [(1–2α) × 100]% CI for *D* is less than the MMES (i.e., θ; NEB), then the null hypothesis that the distribution deviates meaningfully from a normal distribution is rejected in favour of the alternative hypothesis that the *D* statistic is small enough for the distribution to be declared negligibly different from a theoretical normal distribution (Figure 3b). To generate the CI via the percentile bootstrap method, samples are drawn with replacement from the sample distribution and the 1-α quantile of the distribution of *D* statistics across all bootstrap samples is used to define the upper bound of CI for the test statistic.

The KS test, as a goodness-of-fit test on continuous data, compares the overall shape of distributions. Since it makes no assumptions about the specific distribution of data, it does not involve estimating population parameters, and hence there is no population *D* parameter estimated. However, the KS test needs to approximate the empirical or sample cumulative distribution function, resulting in the *D* statistic being a function of the size of the sample distribution. Specifically, the *D* statistic decreases with increasing *N* (holding distribution shape constant). Thus, relying on the *D* statistic to set a NEB for the NET-KS test for normality is problematic as the *D* statistic is not agnostic to sample size.

To address this challenge, we adjusted the *D* statistic to minimize the effect of sample size. The adjusted *D* statistic, $D_{adj}$, is calculated as:

$$D_{adj} = \frac{D_1 - D_2}{\sqrt{log\left(\frac{n}{10}\right)}},$$

where the numerator is the difference between the *D* statistic for the distribution of interest, $D_1$, and the population *D* when the distribution is normal ($D_2$). We have found that this adjustment resulted in a trivial variance of the average $D_{adj}$ across sample sizes ranging between 30 and 10,000 (*SD* = 0.002 over 50,000 replications). The adjustment retains the same interpretation and scale of the *D* statistic, such that if the sample is normally distributed, the expected value is $D_{adj} = 0$, and among nonnormal sample distributions, $D_{adj} > 0$.

## Monte Carlo Study (Study 1)

A simulation study was conducted to assess the Type I error and power rates of the proposed NET-SW and NET-KS tests. Further, the proposed tests were compared with their traditional, difference-based counterparts for assessing normality. Because a new method is proposed to assess for distributional normality, the original KS is utilized in the simulation for the sake of evaluating the NET testing approach, although as noted above, there are recent adaptations of the KS (Ostrovski, 2023).

The Tukey *g*-and-*h* (TGH) distributions are a family of parametric distributions that may take on non-normal features. The parameter *g* controls skewness, with positive values indicating positive skewness and negative values indicating negative skewness. The parameter *h* controls the extent of kurtosis of the distribution. Positive *h* values indicate heavier tails and a sharper central peak compared to a normal distribution (i.e., leptokurtic), while negative *h* values indicate lighter tails and a flatter central peak (i.e., platykurtic). When $g = h = 0$, the distribution

is normal (Tukey, 1977). The flexibility of the TGH distributions allows simulation of data with varying levels of skewness and kurtosis.

The simulation study manipulated the population distribution shape (normal distribution and several non-normal distributions) and sample size (*N*) to reflect common distributions found within social and behavioral science research (see Figure 4). The population variance of the distributions was not manipulated as the *W* statistic is not affected by distribution variance and the $D_{adj}$ statistic requires standardization of the sample distribution to compare it to a reference normal distribution.

Five thousand replications were conducted for each condition, with each replication drawing 5000 random samples with replacement from the simulated *g*-and-*h* distribution. During each replication, a unique *g*-and-*h* distribution is created from which 5000 bootstrap samples are drawn with replacement. Sample sizes were $N$ = 30, 50, 75, 100, 150, 250, 500, 1000, and 5000 and a nominal Type I error rate of $\alpha = .05$ was used. To generate the appropriate CI bound for each of the proposed NET-SW and NET-KS tests, each test statistic was computed on each of 5000 bootstrap samples ($D_{adj}$ for KS or *W* for SW).

For the NET-SW and NET-KS tests, normality was concluded if the appropriate bound of the CI for the test statistic (lower bound *W*, upper bound *D*) fell within the NEB (i.e., rejecting the null hypothesis that the distribution deviates meaningfully from normality). For the difference-based tests, normality was concluded if the null hypothesis was not rejected (i.e., the approach often used by researchers is to conclude the distribution is normal if $p > \alpha$). With the NET-based normality tests, a Type I error occurs when the test falsely concludes normality while power is the probability of correctly concluding normality. For the difference-based test, a Type I error occurs when the test falsely concludes nonnormality and power is the probability of

correctly concluding nonnormality. To more easily compare the performance of the NETs for normality to the traditional tests, the rates at which each test concluded that a distribution was normal as defined above were tracked.

**Simulation to Choose an Appropriate MMES**

We used the approach of Chaffin and Rhiel (1993) to select the MMES for the NET-SW and NET-KS. The simulation considers how skewness and kurtosis influence the one-sample *t* test, specifically, how skewness and kurtosis affect the difference between the nominal probability of a Type I error (α) and the observed probability of committing a Type I error across several nonnormal distributions.

The simulations computed the proportion of one-sample *t* statistics falling beyond the critical values for the two-tailed 1% significance level when the parent population mean is set at 0 and the observed sample means and SDs are compared to a reference *t* distribution with μ = 0, σ = 1. A sample size of $N = 25$ was chosen because violations of the assumption of normality are most problematic at smaller sample sizes (i.e., the goal was to identify a condition that would quantify the maximum error that a researcher might experience). Several nonnormal $\chi^2$ and *g*-and-*h* distributions were tested, with 2,000,000 simulations for each distribution. For the simulations using the $\chi^2$ distributions, the distribution ($M = 0$, $SD = 1$) was standardized using the formula:

$$\frac{\chi^2 - df_{\chi^2}}{\sqrt{2df_{\chi^2}}}$$

Since a nominal probability of α = .01 was used across both tails of the reference *t* distribution, the nominal probability that the test statistic would mistakenly fall in either tail rejection region was .005.

The empirical Type I error rate for the lower tail effects ($\alpha_{E,L}$) was computed as the sum of the negative *t* values that were less than or equal to the lower tail critical *t* statistic divided by the number of simulations (*nsim*):

$$\alpha_{E,L} = \frac{\sum t \leq t_{.005,df}}{nsim}$$

Similarly, the empirical Type I error rate for the upper tail effects ($\alpha_{E,U}$) was computed as the sum of the positive *t* values that were greater than or equal to the upper tail critical *t* statistic divided by the number of simulations:

$$\alpha_{E,U} = \frac{\sum t \geq t_{.995,df}}{nsim}$$

The combined empirical Type I error rate (across both tails, $\alpha_{E,C}$) was computed as the sum of all Type I errors across both lower tail and upper tail effects divided by the total number of simulations:

$$\alpha_{E,C} = \frac{\sum t \leq t_{.005,df} + \sum t \geq t_{.995,df}}{nsim}$$

For each of the lower, upper, or combined tails, the percentage error (PE) was computed as:

$$PE = \frac{\alpha - \alpha_E}{\alpha}$$

where α represents the appropriate nominal Type I error rate (e.g., .005 for one-tailed tests, .01 for two-tailed tests) and $\alpha_E$ represents the appropriate empirical Type I error rate (e.g., $\alpha_{E,C}$).

For distributions that produced *W* statistics between .950 and .975, the PE rates across combined tails were generally below 50%. These same distributions produced $D_{adj}$ statistics between .025 and .015. Hence, $W$ = .950 and $D_{adj}$ = .025 were selected as liberal cut-offs, whereas $W$ = .975 and $D_{adj}$ = .015 were selected as conservative cut-offs. Table 1 shows skewed, platykurtic, and leptokurtic distributions that correspond to these criteria (produced using $N$ = 5000 for better depiction of true population shapes). The distributions in the top row of

the table produced a *W* statistic of approximately .975 and a $D_{adj}$ value of approximately .015, which was set as the conservative criteria for the MMES in the NET-SW and NET-KS, respectively (i.e., NET-SW$_{.975}$; NET-KS$_{.015}$). The distributions in the bottom row of the table produced a *W* value of approximately .950 and a $D_{adj}$ value of approximately .025; those values were set as the liberal criterion for the MMES for the NET-SW and NET-KS, respectively (NET-SW$_{.95}$, NET-KS$_{.025}$). The distributions, being at the MMES, reflect minimally meaningful deviations from normality. Anything less extreme would be considered negligibly different from normal and anything more extreme would be considered a meaningful deviation from normality.

Two positively skewed distributions are examples of distributions that are precisely at the chosen liberal and conservative bounds. A distribution with parameters $g$ = .2929 and $h$ = 0 produces *W* and $D_{adj}$ statistics that are within three decimal places of the conservative criterion (SW = .975 & $D_{adj}$= .015). Similarly, a distribution with parameters $g$ = .4410 and $h$ = 0 produce *W* and $D_{adj}$ statistics that are at the liberal criterion with *W* = .950 and $D_{adj}$= .025. These distributions are examples of distributions that have the smallest deviation from a normal distribution that is considered meaningful (Figure 5a and 5b). Any distribution with less extreme deviation would be considered negligibly different from normal. Two conditions were added to the Monte Carlo simulation that reflect these two distributions at the liberal and conservative bounds. The purpose of these conditions was to evaluate the proportion of conclusions of normality of the NETs at the border of the NET interval. Specifically, the proportion of conclusions of normality were evaluated for the NET-SW$_{.950}$; NET-KS$_{.025}$ for a distribution with parameters $g$ = .4410 and $h$ = 0 (since it is at the liberal bound) and the proportion of conclusions of normality of the NET-SW$_{.975}$; NET-KS$_{.015}$ for a distribution with parameters $g$ = .2929 and $h$ =

0 (at the conservative bound). If the tests perform well, the empirical Type I error rates when these distributions are simulated and compared to their matching bound should equal α.

## Results (Study 1)

Figures 6a, 6b, and 6c provide the rates at which each pair of difference-based (KS, SW) and NET-based (NET-KS, NET-SW) tests make the conclusion that the distribution is normal. For the NETs, the rates are displayed for both the liberal and conservative bounds. Figure 6a depicts the rates for distributions with normal or negligibly different from normal shapes, Figure 6b depicts the rates for two distributions with nonnegligible deviations from normality, and Figure 6c depicts the rates for two distributions that deviate non-negligibly from normal, with the deviations being equal to the minimally meaningful deviation from normality (i.e., SW = .975 & $D_{adj}$= .015 and SW = .950 & $D_{adj}$ = .025).

### Normal/Negligibly Different from Normal

In these conditions, the population distribution is normal or negligibly different from normal. For the SW and KS difference-based tests, not rejecting the null hypothesis ($p > \alpha$) is used to imply that the sample distribution does not deviate from a theoretical normal distribution (which is a correct decision when the population from which the sample was drawn is perfectly normal or an incorrect decision if the distribution deviates from normal). Rejecting the null hypothesis ($p \leq \alpha$) suggests that the sample distribution is significantly different from a theoretical normal distribution (which is a false conclusion, or Type I error, when the sample distribution is drawn from a normal distribution or a correct decision when the sample distribution deviates from normal).

For the NET-SW and NET-KS, not rejecting the null hypothesis implies that the sample distribution is not negligibly different from a theoretical normal distribution, which is a false

conclusion (i.e., a Type II error) when the population distribution from which the sample was drawn is perfectly normal or negligibly different from normal. Rejecting the null hypothesis suggests that the distribution is negligibly different from a theoretical normal distribution, which is a correct conclusion in these conditions.

***g* = 0, *h* = 0.** The SW test performs well across all sample sizes, with rates approximately equal to the expected value (.95). On other hand, the KS test has near perfect detection of normality (~ 100%) across most sample sizes, which is unexpectedly high and contrasts with the expected error rate. Given that the data were drawn from a normal distribution, a small proportion of false rejections (about 5%) are expected due to random chance. In other words, within the context of NHST, the KS test should incorrectly reject the null hypothesis of normality and conclude that the distribution is nonnormal about 5% of the time, as seen with the SW.

The best performing NET test is the NET-SW$_{.950}$. It reaches 100% power to detect normality at about $N$ = 150, consistently detecting normality for all samples from that size onward, with no Type II errors (in the NET context, falsely retaining the null hypothesis and concluding nonnormality). The other procedures (NET-SW$_{.975}$, NET-KS$_{.015}$, NET-KS$_{.025}$) do not reach 100% power until at least $N$ = 500. The NET-SW$_{.975}$ performed better than both the liberal and conservative NET-KS tests.

***g* = 0.2, *h* = 0.** This distribution has negligible positive skewness. Since the distribution is not normal, the traditional tests (KS, SW) should not conclude normality. The KS test often mistakenly concludes normality until about an $N$ = 500, when rates of incorrectly concluding normality finally approach 0. For the SW test, rates are near 100% at small $N$, and then steadily decrease as $N$ increases (reaching 0 at around $N$ = 500). Again, among the four NETs, the NET-

$SW_{.950}$ has near-zero power at about $N = 50$ and reaches 100% power between $N = 250$ and $N = 500$. The NET-$SW_{.950}$ performs slightly better than the NET-$SW_{.975}$, which performs poorly with small sample sizes. The NET-$KS_{.015}$ exhibits limited power, reaching its highest rate of normality detection of only .75 even at the largest sample size ($N = 5000$), indicating suboptimal performance for correctly identifying the true distribution as negligibly different from normal. The NET-$KS_{.025}$ performs only slightly better, reaching its maximum power level at $N = 1000$ with rates around .90. Thus, both KS tests perform poorly relative to the NET-SW tests.

***g* = 0, *h* = -0.1.** Here, the distribution is slightly platykurtic. Since the distribution is not normal, the traditional tests (KS, SW) should not conclude normality. The KS test mistakenly concludes normality at rates near 100% until about an $N = 500$, when rates of incorrectly concluding normality approach 0 (i.e., between $N = 30$ and $N = 500$, rates are near 100%, at $N = 1000$, rates are at 87.5%, and at $N = 5000$, the rates plummet to 0).

For the SW test, rates at which the test concludes normality are near 1 at small *N*, and then steadily decrease as *N* increases (reaching 0 at around $N = 500$). Again, among the four NETs, the NET-$SW_{.950}$ has near-zero power at about $N = 50$ and reaches 100% power at around $N = 150$, performing better than NET-$SW_{.975.}$ Both the NET-$KS_{.015}$ and NET-$KS_{.025}$ perform poorly relative to the NET-SW tests.

**Nonnormal/Non-negligibly Different from Normal**

In these conditions, the sample distributions deviate meaningfully from a theoretical normal distribution, and thus the test should conclude that the distribution is not normal. Any conclusion that the distribution is normal or negligibly different from normal is incorrect. Within the hypothesis testing framework of the SW and KS difference-based tests, not rejecting the null hypothesis ($p > \alpha$) implies that there is insufficient evidence to reject the null hypothesis that the

sample distribution was drawn from a normal population distribution, an incorrect conclusion (Type II error). Rejecting the null hypothesis ($p \leq \alpha$) suggests that the sample distribution is significantly different from a normal distribution, which is a correct conclusion. Conversely, for the NET-SW and NET-KS, failing to reject the null hypothesis that the sample is drawn from a (non-negligibly) nonnormal distribution implies that there is insufficient evidence to conclude that the sample was drawn from a distribution that negligibly deviates from normality, which is a correct conclusion. Rejecting the null hypothesis suggests that the sample distribution is negligibly different from a normal distribution, which is an incorrect conclusion (a Type I error).

***g* = 0.5, *h* = 0.1.** This distribution is positively skewed and leptokurtic. The KS test has a high rate of falsely concluding normality until about $N = 100$ (90%), while the SW test performs much better, with rates of falsely concluding normality reduced to 3% around $N = 100$. All four NETs maintain Type I error rates near 0 across all sample sizes, indicating that they are conservative, rejecting the null hypothesis of non-normality virtually always.

***g* = 0.6, *h* = 0.** This is a positively skewed distribution. The KS test has high Type II error rates until about $N = 100$, while the SW test performs much better than the KS test and has much lower Type II rates in smaller samples, reaching a Type II error rate of 0 at $N = 75$. All four NETs tests maintain Type I error rates near 0 across all sample sizes.

**Nonnormal/Non-negligibly Different from Normal but at the Negligible Effect Bound**

As discussed above, distributions where the degree of nonnormality was precisely at the bound for deeming a distribution non-negligibly different from normal were also included. In these conditions, the NETs are expected to conclude normality with an empirical Type I error rate equal to $\alpha$.

***g* = 0.4419, *h* = 0.** This distribution is at the liberal NEB (SW = .950; $D_{adj}$ = .025) and thus Type I error rates for liberal NETs (NET-$SW_{.950}$; NET-$KS_{.025}$) are expected to be α = .05. The NET-$KS_{.025}$ had Type I error rates near 0 across all sample sizes (i.e., is overly conservative), while the NET-$SW_{.950}$ had Type I error rates near 0 at lower *N* but reached the nominal significance level of .05 for *N* > 500.

***g* = 0.2929, *h* = 0.** This distribution is at the conservative NEB (SW = .975; $D_{adj}$ = .015) and so Type I error rates for conservative NETs (NET-$SW_{.975}$; NET-$KS_{.015}$) are expected to be α = .05. Like the liberal-bound conditions, the NET-$KS_{.015}$ has Type I error rates near 0 across all sample sizes (i.e., is overly conservative), while the NET-$SW_{.975}$ reached the nominal significance level of .05 for *N* > 250.

## Discussion (Study 1)

Given the ubiquity of nonnormally distributed variables across the social sciences, and the impact of this nonnormality on parametric model inferences, it is important to have tests of normality that are both valid and accessible. Graphical methods for evaluating normality can be unreliable, whereas traditional test statistics for assessing normality are limited by both unreliability (e.g., sensitivity to outliers and small deviations in large samples, lack of power in small samples, reliance on distributional assumptions that are often not met in practice, and failing to distinguish statistical significance from practical significance) and an inappropriate orientation of the hypotheses (i.e., the null stipulates that the distribution is normal, rather than nonnormal). Specifically, these tests inappropriately evaluate significant differences from a normal distribution rather than a negligible difference from a normal distribution.

We proposed NETs for normality that test for evidence that a sample distribution is negligibly different from a normal distribution and conducted simulations to compare the Type I

error and power rates of the NET-based normality tests to comparable rates for traditional tests of normality. Based on current literature, no tests to date have been developed for assessing normality using a NET framework.

The use of NETs for normality means that the test only concludes normality if the evidence favourably shows that a sample distribution is similar to a normal distribution. The probability that a test concludes that a sample distribution was drawn from a normal distribution was compared across the traditional one-sample KS test, the traditional SW test, and two variations of NET-based tests (i.e., NET-KS and NET-SW). In the simulation study, the first distributions for data generation were either normal or negligibly different from normal. For the normal distribution condition, the SW test performs optimally, correctly concluding that the distribution is normal across all samples sizes at a rate of approximately .95 (i.e., these tests falsely detect significant differences between the sample distribution and a theoretical normal distribution around 5% of the time). However, the KS test has rates of concluding normality near 1.00 across all sample sizes, which is a limitation of the tests' requirement that a researcher specify the sample, rather than the population, means and standard deviations, which biases the KS test towards accepting the null hypothesis. Specifically, because a researcher never knows the population parameters, only the sample statistics are available to them. However, when the distribution was slightly nonnormal, the rate at which the test falsely concludes normality was excessively high. For example, the KS falsely concluded normality close to 100% of the time until around $N = 500$ for both conditions of slightly nonnormal distributions.

When the population distribution was normal, the liberal NET-SW (NEB of .95) reached 80% power to detect a negligible deviation from normality at the smallest sample size of any of the four NET tests ($N = 75$). The remaining NET procedures required much larger sample sizes

to reach power of at least 80% (e.g., $N$ = 250 to 1000). Generally, the NETs performed poorly (i.e., high Type II error rates) in small samples for detecting normal distributions but performed well with moderate to large sample sizes. Similar results were obtained for distributions that were negligibly different from normal; however, as expected, larger sample sizes were required for achieving acceptable power. For example, the liberal NET-SW procedure reached approximately 80% power with $N$ = 250 for the mildly skewed distribution and $N$ = 150 for the mildly platykurtic distribution. These results (NETs are conservative at low $N$) are not unexpected; concluding that a distribution is 'negligibly different from normal' is a very strong statement that naturally should not be concluded with very small sample sizes.

When data was simulated from distributions that were non-negligibly different from normal, the results for the traditional KS and SW tests were like those for the distributions that were negligibly different from normal. With small sample sizes, the Type II error rates were very large (e.g., for $N$ = 50 the Type II error rate was .25 for SW and .99 for KS), but for large sample sizes (at around $N$ = 100 for SW and $N$ = 500 for KS) the Type II error rates were near 0. In contrast, the NETs almost never conclude incorrectly that the distribution is normal, with rates close to 0 across all sample sizes and distributions. Overall, in non-negligibly nonnormal distributions, the NET normality tests perform better than the traditional tests, typically concluding nonnormality across all sample sizes and distributions, whereas the difference-based tests tend to flagrantly conclude normality from smaller samples (especially the KS test).

The last set of distributions that data was simulated from had SW and KS values exactly at the bounds of the negligible effect interval. In these non-negligible conditions, the NET-SW and performs very well for moderate to large samples ($N$ > 100 for NET-SW) with Type I error rates near α. However, when sample sizes were very small, the Type I error rates were near 0.

This result naturally follows from the discussion above regarding the difficulty with concluding that a distribution negligibly differs from normal with a small *N*.

To summarize, the proposed NETs performed very well in moderate to large sample sizes; the liberal NET-SW performed the best of all proposed NET procedures. Although it might be easy to recommend the liberal NET-SW without reservation, there might be situations in which only very small deviations from normality can be tolerated; in those situations, the conservative NET-SW would be recommended (or the NEB can be adjusted to impose a stricter threshold for concluding equivalence, requiring narrower and more precise CIs around the estimate of *W*).

Although all NET tests were conservative with small sample sizes (i.e., lower power with small *N*), it could be argued that, at a minimum, a moderate sample size with at least 80% power ($N = 100$ for the NET-SW$_{.95}$ and $N = 250$ for NET-SW$_{.975}$) is required to be able to reject the null that a distribution is non-negligibly nonnormal (i.e., the conservativeness makes sense). Further, the traditional KS and SW tests also had very low power at small sample sizes, but the consequences are reversed; these traditional tests have a high rate of Type II errors (failing to reject the null hypothesis that the data is sampled from a normal distribution) even when the distribution is quite nonnormal.

There are a few limitations to this study. First, regarding the proposed NETs, and like traditional normality tests, they do not highlight the nature or the direction in which the distribution deviates from normality when the null is not rejected (e.g., is the nonnormality based on skewness, kurtosis, presence of outliers?). In other words, these tests act as 'omnibus' tests, providing a general indication that data are nonnormally distributed. Additional analyses are required to understand the characteristics of the deviation (graphs and descriptive statistics), and

these analysis may inform the choice of subsequent analyses (e.g., nonparametric or robust statistical methods) or data transformations.

Second, for distributions that are normal or negligibly different from normal, the bootstrap CI for the NET-SW often produces a conservative lower bound, meaning that the lower bound of the bootstrap CI tends to be less than the NEB (i.e., .95 or .975), leading to a failure to reject the null hypothesis that the sample is drawn from a distribution that is meaningfully different from a normal one. The conservative nature of the bootstrap CIs is obvious with normal (or near normal) distributions because the CI often does not even contain the sample test statistic (e.g., SW = .999; bootstrap CI = {.965, .988}). Due to the conservativeness in these conditions, the NET-SW has reduced power to detect normality. Solutions to this problem were explored in the second study of this dissertation. Finally, the findings in this study cannot be generalized to conditions that go beyond the conditions investigated here. For example, although it is likely that the results will extend to different negligibly or non-negligibly nonnormal distributions, this prediction cannot be verified without further study.

To summarize, the results of this study suggest that, in small samples, the NETs have low power to conclude normality when the degree of nonnormality is negligible (i.e., falls within the negligible effect interval) while the traditional difference-based tests have low power to conclude nonnormality when the distributions are nonnormal. With at least moderate sample sizes ($N$ = 150 for NET-SW$_{.95}$ and $N$ = 250 for NET-SW$_{.975}$), the NETs have upwards of 80% power to detect distributions that are normal or negligibly different from normal. In non-negligibly nonnormal distributions, the NET-SW performs very well even with small sample sizes ($N$ = 30 to $N$ = 75), and in this situation (small $N$, nonnormal distribution) it is better to err towards

concluding nonnormality than normality (when the research hypothesis relates to testing for normality). Given the superior performance of the liberal NET-SW over the conservative NET-SW and both NET-KS tests, we recommend that researchers utilize the liberal NET-SW as a default. However, when only very minor deviations from normality can be tolerated, the conservative NET-SW could be adopted (or the NEB can be made even more stringent). Given the promising results around the liberal NET-based SW statistic over the conservative NET-based SW and both NET-KS tests, we recommend the liberal NET-SW. However, the conservative NET-SW (or an even stricter bound) may be utilized if only slight deviations from normality are tolerable. Future research ought to explore in greater depth the properties of the NET-SW and the potential for other NET-based tests for normality to be developed.

## Applied Example (Study 1)

To demonstrate how the NET-SW procedures can be adopted in a research setting, an example using real data is provided. The performance of the KS test and NET-KS is not reported based on their inferior performance at detecting meaningful deviations from normality. The goal is to encourage researchers to adopt the most rigorous and accurate method, so the focus is on the NET-SW using the liberal NEB of .95.

The dataset comes from an RCT-based study on the effectiveness of an online cognitive-behavioural therapy (CBT) for maladaptive perfectionism among a sample of undergraduate students with high levels of perfectionism (Arpin-Cribbie et al., 2012). The dataset is available within the `negligible` package in R (Cribbie et al., 2023).

For this example, a linear regression model was fit using self-oriented perfectionism scores from the Hewitt and Flett Multidimensional Perfectionism Scale (Hewitt & Flett, 1991) as the outcome, predicted from both Perfectionistic Cognitions Inventory scores (Flett et al., 1998)

and Automatic Thoughts Questionnaire (Hollon & Kendall, 1980) scores, with all variables measured at pretest. The model residuals were extracted ($N$ = 83) and tested for normality using the `neg.normal` function from the **negligible** package (Cribbie et al., 2023) with a liberal NEB of $W$ = .950. The residuals are characterized by negative skewness (-0.44) but little kurtosis (0.01; Figure 7). The 20% trimmed mean of the residuals is 0.41 (i.e., 20% of the smallest values and 20% of the largest values are removed before calculating the mean). If the NET-SW concludes that the sample distribution deviates from a normal distribution meaningfully, these descriptive statistics may help a researcher choose between a transformation, address outliers, or apply a nonparametric or robust approach.

The traditional SW test is nonsignificant, $W$ = .979, $p$ = .19. The traditional null hypothesis that the distribution from which the data were drawn is normal cannot be rejected at the $\alpha$ = .05 level. However, the NET-SW test had a lower bound of the 95% CI for the bootstrap samples from the distribution of residuals of .944, which is below the lower bound of the negligible effect (equivalence) interval of .950 (liberal criterion). Thus, the NET-SW test concludes that the null hypothesis that the degree of nonnormality is extreme cannot be rejected (i.e., the distribution is not concluded to be negligibly different from normal). Therefore, these tests provide conflicting results. On the one hand, the SW test suggests that there is not enough evidence to reject the null hypothesis that the sample comes from a normal distribution, so it cannot confidently be concluded that the data are nonnormal. However, the NET-SW shows that we cannot confidently conclude the data are ‘close enough’ to normal, leaving the possibility of non-negligible deviations (i.e., it may still differ from a normal distribution in a non-negligible way). Given the small skewness and near-zero kurtosis, it is reasonable to expect these ambiguous results. The data exhibit only mild deviations from normality, so while the SW test

may not detect evidence of nonnormality, the NET-SW may still flag small deviations as non-negligible. These differences underscore the sensitivity of the tests to the differing hypotheses.

The NET-SW test is conservative, meaning that it requires stronger evidence to conclude that the sample distribution is non-negligibly different from a normal distribution, providing a strong rationale for choosing a more robust measure in subsequent analyses that mitigate the influence of outliers and nonnormality in parametric methods. So even when the SW test is nonsignificant, the researcher is encouraged to be cautious and assume that normality is violated, ensuring that the subsequent results remain valid as far as violations of normality are concerned.

# CHAPTER THREE

# NEGLIGIBLE EFFECT TESTS FOR DISTRIBUTIONAL NORMALITY: IMPROVING CONFIDENCE INTERVALS FOR THE SHAPIRO-WILK APPROACH (STUDY 2)

Negligible effect tests (NET) are used to determine whether an effect is small enough to be deemed of little consequence for substantive purposes. The effect may be the difference between two means, the correlation between a pair of variables, a regression coefficient, and so on. NETs assess evidence that an effect falls within a predefined range of values that are close to a specific (e.g., nil) effect, known as the negligible effect interval (NEI). NETs for normality were developed in Study 1 as an alternative to traditional statistical tests for normality (e.g., Shapiro-Wilk and Kolmogorov-Smirnov tests). These NETs evaluate whether a sample distribution is negligibly different from a theoretical normal distribution. Specifically, these tests address whether the null hypothesis that a target population distribution is non-negligibly different from a theoretical normal distribution can be rejected.

In Study 1, two NETs for normality were developed: a NET based on the one-sample Kolmogorov-Smirnov test (NET-KS) and a NET based on the Shapiro-Wilk test (NET-SW). Although both tests virtually never falsely concluded normality, the NET-SW was shown to have greater power to conclude normality when distributions were normal or negligibly different from nonnormal, especially in small to moderate sample sizes. Thus, the NET-SW had the best balance of Type I error control and power and was chosen as the focus of this study.

A finding from Study 1 regarding the NET-SW procedure was that power to detect a negligible difference from normality was hindered by the fact that the percentile bootstrap CI was biased downwards (i.e., away from $W = 1$) when population distributions were normal or

close to normal. In other words, the conservative nature of the percentile CI approach when distributions are normal or close to normal reduces the likelihood of detecting that a distribution is negligibly different from normal even when the distribution is normally distributed or almost normally distributed. This issue is the focus of this study.

The layout of Study 2 is as follows. First, the original NET-SW procedure is described. Second, the limitation of the NET-SW (i.e., the reduced power for distributions normally distributed or close to normally distributed) is explained. Third, alternative methods for creating the CIs for the NET-SW are described. Lastly, a Monte Carlo study is used to evaluate the statistical properties of the modified NET-SW procedure. An applied example is also provided, using an R function that applies the recommended alternative for computing CIs for the NET-SW (i.e., the method that optimally corrects the CI for distributions that are normal or close to normal in form).

**Original NET-SW Procedure**

The NET-SW was derived from the traditional SW test, which uses traditional null hypothesis testing to assess whether a population distribution from which sample data are drawn deviates from normal. By incorporating NET-based procedures into the SW test, the NET-SW uses a null hypothesis that the distribution meaningfully deviates from normality, while the alternative hypothesis suggests that any deviation from normality is minimal. The null hypothesis for the NET-SW states that population $W$ is less than or equal to the minimally meaningful effect size (MMES; Beribisky et al., 2019), which, in this context, specifies how much a distribution can deviate from normal without it being meaningful. Generally, a NEI is based on the MMES and is often denoted using $\{-\delta, \delta\}$; these symmetric boundaries represent the smallest effects, in either direction, from the desired effect (e.g., mean difference = 0) that

would be deemed nonnegligible. Each bound of the NEI is referred to as a negligible effect bound (NEB). As the traditional SW is on a positive-only correlation-based metric, *W* ranges from 0 to 1, with values closer to 1 indicative of normality. Therefore, the NEI for the NET-SW is defined as $\{\zeta, 1\}$. Deviations from normality are unidirectional, and so the *W* statistic that demarcates the smallest deviation from normality that is consequential is defined by the lower NEB, ζ. The null hypothesis for NET-SW is stated as $H_0$: $W_{pop} \leq \zeta$, while the alternative hypothesis states that population *W* is larger than the lower NEB: $H_1$: $W_{pop} > \zeta$. Only the lower NEB of the NEI (ζ) is relevant for the NET-SW, since the values of *W* have an upper bound of 1.

In Study 1, the procedure for the NET-SW utilized a basic percentile bootstrap (or empirical percentile bootstrap) method, which constructs CIs using the bootstrap distribution of the statistic of interest. Briefly, bootstrap sampling draws *B* random samples with replacement of size *N* from a distribution of size *N*. This process mimics the process of random sampling from the population, but rather than directly sampling from the population, the sample is drawn with replacement from the observed data, with each observation having an equal chance of being selected at each draw. For each bootstrap sample, the statistic of interest is computed using the resampled data (e.g., *W*), and the distribution of the statistic can then be used to compute CIs. For example, after drawing *X* bootstrap samples, the lower and upper bounds of the CI are determined by selecting the appropriate percentiles from the bootstrap distribution of the statistic calculated in each bootstrap sample (Rousselet et al., 2021). In Study 1, the percentile bootstrap approach was adopted to estimate the sampling distribution of the *W* statistic. If α represents the nominal Type I error rate, then in an NET framework we calculate a 100(1 – 2α)% CI, where the α percentile represents the lower bound of the CI and the (1- α) percentile represents the upper bound of the CI. For instance, for a 90% CI, the 5$^{th}$ percentile of the distribution would represent

the lower bound of the CI and the 95$^{th}$ percentile would represent the upper bound of the CI. Although the percentile method does not require any assumptions about the distribution of the bootstrap distribution, it does not perform optimally with very small samples and when the sample distribution is not an accurate representation of the population distribution (Rousselet et al., 2021).

For the NET-SW, bootstrap samples are drawn from the empirical sample with replacement, and the *W* statistic is estimated for each sample. Since a Type I error is only possible at the lower NEB, the test retains an α level of 5% by placing the entire rejection region in the lower tail. In other words, there is a 5% risk of a Type I error in the lower tail (i.e., a 5% risk of indicating that the distribution is negligibly different from normal when in fact it is non-negligibly different from normal). Specifically, a 100(1 – 2α)% CI was calculated, where the α (5$^{th}$) percentile of the bootstrap distribution represents the lower bound of the CI (and the upper bound is not relevant for the statistic). In Study 1, the lower NEB = .950 was used as the liberal bound and the lower NEB = .975 was used as the conservative bound. The *W* statistics from the *B* bootstrap samples are then combined into a distribution, with the α quantile of the distribution defining the lower bound of the bootstrapped CI for *W*. If the lower bound of the 100(1 – 2α)% percentile CI for the bootstrap distribution is greater than the lower NEB, then the test concludes that the distribution is equivalent to normal. Recall that with the NET-SW, there is only one rejection region (i.e., one-tailed), and hence only one bound of the NEI is meaningful {ζ, 1}; the upper bound of 1 represents optimal normality and the lower bound, ζ, represents the largest value of the *W* statistic that deviates meaningfully from normality.

**Issue with the NET-SW with Distributions Close to Normal in Shape**

In Study 1, the NET-SW procedure provided the best balance of Type I error control and power, compared with the NET-KS test and was recommended. This research also found that bootstrap sampling leads to an approximately normal bootstrap sampling distribution when the sample is not normally distributed; however, if the target distribution is close to normally distributed (i.e., *W* approaches 1), then the bootstrap sampling distribution is negatively skewed and biased downwards. As a result, when the sample is drawn from a distribution that is very close to normal, the bootstrap percentile CIs are not centered over the sample *W* statistic; in fact, the bootstrap percentile CIs often do not contain the sample *W*. Rather, the sample *W* is typically above the upper bound of the percentile CI (Figure 8-A). However, when the sample is not drawn from a normal distribution, the percentile CI is approximately centered over the observed sample *W* (Figure 8-B). Specifically, because the percentile interval is created from quantiles of the bootstrap distribution, rather than being directly formed from the estimate of the statistic in the sample data, there is no guarantee that the CI will contain the sample value of the statistic. Further, as a nonparametric method, the bootstrap percentile CI method does not have a mechanism to adjust for skewness in the distribution of the bootstrap sample. Because of the percentile CI being biased downwards, power to detect negligible differences from normality with the NETs is deflated. Specifically, when the lower bound of the bootstrap CI is deflated (i.e., biased downwards), it is possible that the lower bound of the CI is below the NEB, leading to a conclusion that the distribution is non-negligibly different from normal (a Type II error if the population *W* value falls above the NEB; see Figure 9).

**Solutions**

Below, several alternative solutions for the conservative nature of the CI based on the percentile bootstrap are described, namely the stochastic bootstrap, parametric bootstrap, Fisher's *r* to *z* transformation, and the bias-corrected and accelerated approach.

**Stochastic bootstrap.** Unlike the traditional bootstrap that samples directly from the data, the stochastic bootstrap is a variation of the traditional bootstrap method that adds a random noise component to each resampled observation (Ripley, 1987). With stochastic bootstrap sampling, each of the $i = 1, \ldots, N$ observations from each of the $j = 1, \ldots, B$ bootstrap samples has a random noise component added:

$$BS_{ji'} = BS_{ji} + \varepsilon_{ji},$$

where $BS_{ji'}$ represents the recorded observation for case $i$ in bootstrap sample $j$, $BS_{ji}$ represents the original $i$th observation from bootstrap sample $j$, and $\varepsilon_{ji}$ represents the random noise component [e.g., $\varepsilon_i \sim N(0, \sigma^2)$] for the $i$th observation from bootstrap sample $j$.

For example, if variable $X$ contains the sample observations {2,5,4,7,3}, a bootstrap sample may select as the first observation, 7 (i.e., $BS_{11} = 7$). $BS_{ji}$ is next added to $\varepsilon_{ji}$, where $\varepsilon_{ji}$ is a value selected from a normal distribution with an appropriate population standard deviation [e.g., $\varepsilon_i \sim N(0, 1)$]. Perhaps the $\varepsilon_{ji}$ value selected (rounded to two decimal places) is $\varepsilon_{11} = -0.43$. Thus, the first recorded observation in bootstrap sample 1 $BS_{11'} = 7 - 0.43 = 6.57$). This process continues for all $N$ observations from all $B$ bootstrap samples. The idea is to simulate natural variability in the resampling method, which is lost with direct bootstrapping, to capture potentially important variability.

**Parametric bootstrap.** A parametric bootstrap works under the presumption that the underlying distribution of the data is known. Bootstrap samples are generated by simulating data

from a specified model fitted to the observed data. However, misspecification of the underlying distribution form (e.g., specifying a normal distribution when the distribution is nonnormal) can lead to biased and unreliable results (Cornea-Madeira & Davidson, 2015; Lu & Young, 2012; Milan & Whittaker, 1995; Spokoiny & Zhilova, 2015).

There are three steps to implement a parametric bootstrap that assumes a normal distribution (also called a Gaussian interval, or normal-approximation CI):

1. The bootstrap samples are drawn with a *W* statistic calculated on each bootstrap sample (*B*).
2. The *M* and *SD* are calculated for the bootstrap distribution of the *W* statistic, called $M_B$ and $SD_B$, respectively.
3. $M_B$ and $SD_B$ are used to construct the CI using the quantiles of the standard normal distribution:

$$\text{CI} = M_B \pm z\, SD_B ,$$

where *z* is the critical value that corresponds to the desired confidence level (e.g., for a 95% CI, *z* = 1.96). For example, if $M_B = 0.7$ and $SD_B = 0.1$, assuming $\alpha = .05$, we first find the *z*-value corresponding to the upper tail of the standard normal distribution: $1 - \alpha/2 = .975 = 1.96$. The margin of error is thus $1 - \alpha/2 \times SD_B = 1.96 \times 0.1 = 0.196$. Finally, we create the CI by adding and subtracting the margin of error from the mean of the bootstrap distribution:

$$\text{CI}_{\text{lower}} = M_B - z\, SD_B = 0.7 - 0.196 = 0.504$$

$$\text{CI}_{\text{upper}} = M_B + z\, SD_B = 0.7 + 0.196 = 0.896$$

The 95% CI is (0.504, 0.896). As discussed above, for the NET-SW the interest is in the lower bound of the $100(1 - 2\alpha)\%$ CI.

**Fisher's *r* to *z* transformation.** This is a classic technique used to deal with the fact that the sampling distribution of correlation coefficients is skewed when the population correlation coefficient is near one of the boundaries (e.g., -1 and +1). The transformation converts the correlation (*r*) to a variable (*z*) whose sampling distribution is more normally distributed (especially when *N* is small). The transformation may help when performing hypothesis tests or constructing CIs for the population correlation coefficient, especially when *r* approaches +/- 1 (Fisher, 1915; Fisher, 1921).

Recall that this situation occurs when *W* approaches 1 with the NET-SW. Fisher's *r* to *z* transformation converts *r* to an unbounded scale where the sampling distribution of *z* is approximately normal, particularly as sample size increases. Once transformed, CIs can be computed using the percentile bootstrap method or hypothesis tests can be run within the *z* (standardized normal) scale (Welz et al., 2022). Inverse transformations allow conversion of confidence intervals or results from hypothesis tests from the *z* scale back to the *r* scale for interpretation.

Given a correlation coefficient *r*, Fisher's *r* to *z* transformation is calculated using the following equation:

$$z = .5\ ln\left(\frac{1+r}{1-r}\right),$$

where *z* is the transformed value in the standard normal scale and *r* is the correlation coefficient. For this context, *W* is substituted for *r* in the above equation since, as discussed in Study 1, *W* is a pseudo-correlation statistic.

**Bias Corrected and Accelerated (BCa) Bootstrap.** The bias-corrected and accelerated (BCa) method (Chernick & LaBudde, 2011; Davison & Hinkley, 1997; DiCiccio & Efron, 1996; Efron & Tibshirani, 1994) corrects bias and skewness in a bootstrap distribution and constructs

confidence intervals for the adjusted statistics. Here, the BCa CI is computed on the distribution of bootstrap *W* statistics from each bootstrap sample. To construct a BCa CI, a bias-correction and acceleration factor is calculated from the bootstrap distribution of the *W* statistic. The percentiles of the bootstrap distribution are adjusted, and then a BCa CI is constructed using the adjusted percentiles, providing a more normalized interval than the standard bootstrap percentile method. Specifically:

1. The observed *W* is calculated for the original sample.
2. Bootstrap samples (*B*) are drawn from the data and the *W* statistic is calculated for each bootstrap sample, $W_1, W_2, \ldots W_B$, forming a distribution of the *W* statistic.
3. The acceleration parameter is calculated, $\hat{\alpha}$, which captures any skewness in the bootstrap distribution:

$$\hat{\alpha} = \frac{\sum_{i=1}^{B}(W_i - \bar{W})^3}{6\left(\sum_{i=1}^{B}(W_i - \bar{W})^2\right)^{3/2}},$$

   where $\bar{W}$ is the mean of the bootstrap sample statistics $W_1, W_2, \ldots W_B$.
4. The bias-correction parameter, $\hat{z}_0$, is calculated, which adjusts for the difference between the observed statistic (*W*) and the mean of the bootstrap distribution. $I(W_i \leq W)$ is an indicator function that equals 1 if the bootstrap sample statistic $W_i$ is less than the sample *W* and 0 otherwise. Then

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\sum_{i=1}^{B} I(W_i \leq W)}{B}\right),$$

   where $\Phi^{-1}$ is the inverse of the standard normal cumulative distribution function and the term in parentheses is the proportion of bootstrap samples that are equal to or less than observed sample *W*.

5. Next, the BCa CI is computed for the sample using the acceleration parameter ($\hat{\alpha}$) and the bias-correction parameter ($\hat{z}_0$):

   Lower adjusted percentile: $\Phi\left(\hat{z}_0 + \frac{\hat{z}_0+\Phi^{-1}(\alpha)}{1-\hat{\alpha}(\hat{z}_0+\Phi^{-1}(\alpha)}\right)$,

   Upper adjusted percentile: $\Phi\left(\hat{z}_0 + \frac{\hat{z}_0+\Phi^{-1}(1-\alpha)}{1-\hat{\alpha}(\hat{z}_0+\Phi^{-1}(1-\alpha)}\right)$,

where α is the confidence level of the CI. Because we work with 5% in each tail of the distribution for the NET-SW, α = .10 (a 90% CI is calculated). The resulting BCa interval reflects a 90% CI that adjusts the original percentiles for bias and skewness using the correction parameters, $\hat{z}_0$, and $\hat{\alpha}$.

## Monte Carlo Study (Study 2)

An initial simulation study was conducted to track the rates at which the percentile bootstrap confidence interval contains the observed *W*. Rates (proportion of replications) in which the CI includes the sample *W* value for five sampling methods are compared: percentile bootstrap (usual), stochastic bootstrap, parametric bootstrap (imposing a normal distribution), Fisher's *r* to *z* transformation, and BCa.

The main simulation study assesses the Type I error and power rates of each of the approaches for computing the CI for the NET-SW test. Two NEBs were adopted for the *W* statistic: .975 for a conservative criterion and .95 for a liberal criterion, with the general idea to see how often the lower bound of the CI is greater than the NEB across multiple replications (i.e., how often the procedure declares the population distribution negligibly different from normal). As in Study 1, Tukey *g*-and-*h* distributions were adopted to produce distributions with a range of skewness and kurtosis: normal ($g = 0$, $h = 0$), negligibly different from normal ($g = 0.2$, $h = 0$ and $g = 0$, $h = -0.1$), non-negligibly different from normal ($g = 0.5$, $h = 0.1$ and $g = 0.6$, $h =$

0), at the NEB of .975 ($g$ = 0.2929, $h$ = 0), and at the NEB of .95 ($g$ = 0.4419, $h$ = 0). Sample sizes are the same as in Study 1 ($n$ = 30, 50, 75, 100, 150, 250, 500, 1000, 5000).

In total, there were 630 conditions (5 methods x 2 bounds x 7 distributions x 9 sample sizes). For each condition, 5000 replications were conducted, with 1000 bootstrap samples for each. A nominal confidence level of $\alpha = .05$ was adopted. In each replication, the test statistic ($W$) was computed in each of the 1000 bootstrap samples. For the percentile method, the $\alpha$ percentile ($5^{th}$ percentile) of the bootstrap distribution of $W$ statistics is computed and compared to the lower NEB. The stochastic approach follows the same steps, but with each random bootstrap sample drawn with replacement, normally distributed stochastic noise is added with $M$ = 0 and SD = 5% of the SD of the sample distribution.

In the parametric normal approach, the CI for the mean of the $W$ bootstrap statistics is calculated using the normal approximation method described above (i.e., the CI is computed as the mean +/- the product of the critical value and the $SD$). The lower bound of this CI is then compared to the lower NEB. For the Fisher's $r$ to $z$ approach, each bootstrap sample is drawn and a Fisher's $z$ is computed on each $W$ statistic. Once the $z$ is computed on each bootstrap sample, the CI is calculated and then converted back to the $W$ scale, with the average lower bound compared to the specified NEB to determine the rate at which the test concludes the distribution is negligible to normal. Finally, for the BCa method, the BCa CI is computed from the bootstrap distribution itself using the percentile CI as a starting point, adjusting the distribution for any bias and skewness, and refining the original percentile bootstrap CI.

For the NET-SW, normality was concluded if the lower bound of the CI for the $W$ generated by each CI approach fell within the NEI (i.e., at or above the lower NEB of .975 or .95), leading to a rejection of the null hypothesis that the distribution deviates non-negligibly

from normal. The rates at which the NET-SW test concluded that a distribution was normal was tracked for each of the five methods for computing the CI. In conditions where the population distribution is non-negligibly different from normal, a Type I error occurs when the test falsely concludes the distribution is normal, and in conditions where the population distribution is negligibly different from normal, power is the probability of correctly concluding normality.

## Results (Study 2)

### Computed Percentile CI Containing the Sample *W* Statistic

Table 2 includes the proportion of CIs from the percentile bootstrap method that contain the observed sample *W* statistic. When the population distribution is perfectly normal, the percentage of CIs that contain the sample *W* statistic is 66% with $N = 30$ and 42% when $N = 5000$ and. Conversely, for the perfect normal condition, the sample *W* is not contained in the percentile CI in 34 to 58% of replications. For conditions where the distribution is negligibly nonnormal, 76.8% ($N = 30$) to 100% ($N = 250$) of CIs contain the sample *W* statistic. For conditions that are at the NEB, the sample *W* is contained within the percentile CI between 85% and 100% of replications. For the conditions with nonnegligible nonnormality, the sample *W* is nearly always contained within the percentile CI.

### Rates at which Sample *W* is Contained within Alternative CI Approaches

Figure 10 shows the proportion of replications in which the sample *W* is contained in the CI for each of the alternative procedures discussed above across sample sizes. The usual percentile method is included for reference and starts at around .67 at $N = 30$ and decreases to .42 at $N = 5000$. The parametric normal method shows high rates of coverage in samples up to $N = 150$ and decreases slightly to about 90% in the largest sample size. Fisher's *r* to *z* maintains a perfect coverage rate of 1 across all sample sizes. The stochastic method has slightly improved

coverage compared to the percentile bootstrap approach—an expected pattern given that the stochastic method is a refined variation of the percentile method. Finally, the BCa method starts with very high coverage rates of 84% at $N = 30$ and reaches perfect rates at $N = 75$. Fisher's $r$ to $z$, stochastic, and BCa methods show consistent coverage rates across sample sizes, making them promising methods for the NET-SW test.

**Type I error Rates**

**Non-negligibly nonnormal distributions at the NEB.** When the NEB is .975 and the distribution is $g = 0.2929$, $h = 0$, falling right on the bound for a meaningful deviation from normality when adopting the conservative criterion, Type I error rates are expected to equal α. All methods start with very low rates, never falsely concluding normality from samples $N = 30$ to $N = 100$, and start to incorrectly conclude normality at $N = 150$ (Figure 11a). The percentile method starts with a low rate of .0022 at $N = 150$ and reaches a maximum rate of .0502 at $N = 5000$. The parametric normal interval has slightly higher rates than the percentile method at larger samples, peaking at .0578 for $N = 5000$. The stochastic method starts with a low rate of .0034 at $N = 150$ and peaks at .0594 at $N = 5000$. The BCa is the most conservative method, starting to falsely detect normality at $N = 150$ (.0016) and peaking at $N = 5000$ at a rate of .0344. Fisher's method is very conservative, never concluding normality falsely at any sample size.

When the NEB is .95 and the distribution is $g = 0.4419$, $h = 0$, falling on the bound for a meaningful deviation from normality when adopting the liberal criterion, Type I error rates again should be approximately α. The percentile method has rates that start at 0 for $N = 30$ and $N = 50$, and gradually increase with larger sample sizes, reaching a peak of .0526 for $N = 5000$. This method shows a relatively slow increase in false positive rates. The parametric normal interval increases slightly faster than the percentile method, peaking at .0502 for $N = 5000$, slightly lower

than the percentile method. The stochastic method begins at slightly higher rates compared to the percentile and parametric methods, starting at a rate of .001 at $N = 50$ and increasing more rapidly with larger sample sizes. It reaches its peak of .0644 at $N = 5000$. The BCa is again overly conservative, starting to detect normality at $N = 75$ (.0036) and peaking at a rate of .0404 at $N = 5000$. Fisher's method is most conservative, never falsely detecting normality for any sample size.

**Non-negligibly nonnormal distributions beyond the NEB.** When the distribution is $g = 0.5$, $h = 0.1$ and the conservative NEB of .975 is adopted, for all sample sizes and all methods, the proportion of conclusions of normality is 0 (Figure 11b). Thus, in this condition, the test never falsely concludes that the distribution is normal.

When the distribution is $g = 0.5$, $h = 0.1$ and the liberal NEB of .95 is adopted, all methods maintain very low rates of concluding normality across sample sizes. Specifically, the percentile method maintains very low rates, concluding normality at $N = 75$ (.0088) and peaking at $N = 150$ with a rate of .0096 and then decreasing to 0 at larger sample sizes. The parametric normal interval follows a similar trend as the percentile method but shows slightly higher rates at $N = 100$ (.0136) which then declines as sample size increases. The stochastic method shows slightly higher rates than both the percentile and parametric methods, starting at 0 for $N = 30$, peaking with a false positive rate at $N = 100$ (.0128), declining to 0 thereafter. The BCa is more conservative, initially detecting normality at $N = 75$ (.0046) and peaking at $N = 100$ (.0086) before decreasing back down to a rate of zero. Fisher's method never falsely concludes normality for any sample size.

When the distribution is $g = 0.6$, $h = 0$ and either the conservative NEB of .975 or the liberal NEB of .95 is adopted the rate of false conclusions of normality is 0 for every method and

all sample sizes (Figure 11c). Thus, in this condition, none of the methods ever incorrectly state that the distribution is negligibly different from normal.

**Power**

**Normal distribution.** When $g = 0$, $h = 0$, and the NEB is conservative at .975, none of the methods detect normality when samples are between $N = 30$ and $N = 100$ (Figure 11d). At $N = 150$, all methods have a sharp increase in the detection rate and converge to near perfect detection of normality at $N = 500$. The percentile method reaches high power (.88) at $N = 250$ and near perfect detection (.998) at $N = 500$ and perfect detection from $N = 1000$ onwards. Besides the BCa approach, the performance of the percentile CI is the slowest to ramp up in comparison to other methods.

The parametric normal interval method has similar power rates as the percentile, although at $N = 150$, the detection rate is slightly better (.333) and reaches perfect detection at $N = 500$ and beyond. Fisher's $r$ to $z$ begins to detect normality at a high rate at $N = 150$ (.8354), reaching perfect detection at $N = 250$. This method outperforms all other methods at small and moderate sample sizes. The stochastic method begins to detect normality at $N = 150$ (.3496), increasing its detection sharply at $N = 250$ (.9024). It performs similarly to the percentile and parametric methods in larger sample sizes, reaching perfect detection at $N = 1000$. The BCa method has the poorest detection rates of all methods, only beginning to detect normality at $N = 150$ at a low rate (.1712), but at $N = 250$, it has a slightly better rate than the percentile method at .8538, and improving to near-perfect detection at $N = 500$ (.998) and perfect detection at $N = 1000$.

For $g = 0$, $h = 0$, but using a liberal criterion of the NEB, .95, all methods have improved detection of normality and converge to near perfect detection of normality at $N = 250$. The percentile method weakly detects normality at $N = 50$ (.0056) and improves from $N = 75$ (.447)

onwards, reaching high power at $N = 100$ (.8048). The parametric normal interval begins to detect normality at $N = 50$ (.0166) and consistently improves as sample size increases, reaching a detection rate of .8476 at $N = 100$, near-perfect detection at $N = 150$ (.9778), and perfect detection at $N = 500$. The parametric normal interval improves slightly faster than the percentile method as $N$ increases. The stochastic method begins to detect normality at $N = 50$ (.0236) and reaches high detection at $N = 100$ (.8316), reaching perfect detection at $N = 250$. The BCa has the poorest performance of all methods in small to moderate samples, initially detecting normality at $N = 75$ (.352) and increasing to .7528 at $N = 100$ and converging with other methods to near perfect detection at $N = 1000$. Fisher's $r$ to $z$ begins to detect normality at a rate of .3484 at $N = 75$, reaching perfect detection at $N = 100$. This method outperforms all other methods at moderate sample sizes.

**Negligibly nonnormal distributions.** For $g = 0.2$, $h = 0$, with a conservative criterion of .975, the rates represent how well each method detects negligible differences from normality (Figure 11e). For sample sizes $N = 30$ to $N = 100$, none of the methods detect normality. The percentile method begins to detect normality at $N = 150$ (.0312) and steadily improves with increasing sample size, reaching a high detection rate at $N = 1000$ (.7054) and near-perfect detection at $N = 5000$ (.999). The parametric normal interval slightly outperforms the percentile method, initially detecting normality at $N = 150$ (.0478) and showing higher rates of detection at $N = 150$ (.0478) and showing stronger performance at $N = 1000$ (.7352), with near perfect detection at $N = 5000$ (.9998). The stochastic method performs only slightly better than the parametric method, initially detecting normality at $N = 150$ (.0464). It reaches moderately high detection at $N = 1000$ (.7108) and reaches near-perfect detection at $N = 5000$ (.9998). The BCa performs most poorly of all methods, first detecting normality at $N = 150$ (.0188), reaching only

moderate detection at $N = 1000$ (.6538), and near-perfect detection at $N = 5000$ (.9998). Fisher's method does not detect normality at all from $N = 30$ to $N = 250$ but detects normality perfectly in samples of $N = 500$ and higher.

For $g = 0.2$, $h = 0$, with a liberal criterion of .95, all methods converge to near-perfect detection of normality at $N = 1000$. The BCa method has the lowest power to correctly conclude normality of all methods, initially detecting it at $N = 75$ (.142) and increasing to .8028 at $N = 250$ and peaking at perfect detection at $N = 5000$. The percentile bootstrap method shows the next weakest performance of all methods, with detection starting at $N = 50$ (.0034) and progressively improving from $N = 75$ (.193) and $N = 150$ (.5994) and high detection at $N = 250$ (.8354). The parametric normal interval performs slightly better than the percentile method at smaller sample sizes, with a detection rate of .0096 at $N = 50$ and .2516 at $N = 75$. The parametric method reaches a moderate detection rate of .6566 at $N = 150$ and strong detection of .9812 at $N = 500$. The stochastic method outperforms the percentile and for most smaller sample sizes, the parametric methods, showing a higher detection rate at $N = 50$ of .013, and .2444 at $N = 75$. Detection rates improve rapidly, with .625 at $N = 150$ and high rates by $N = 500$ of .9774. Like other methods, it reaches near-perfect detection at $N = 1000$. Fisher's method does not detect normality at all until $N = 150$, where it detects it perfectly onwards.

For $g = 0$, $h = -0.1$ when the NEB is conservative .975, all methods begin to detect normality at $N = 150$ and converge on near-perfect detection at $N = 1000$ and perfect detection at $N = 5000$ (Figure 11f). The percentile method starts at a rate of .022 (at $N = 150$) and improves to a rate of .2616 at $N = 250$, reaching strong performance at $N = 500$ (.8386). The parametric normal interval performs slightly better than the percentile method, with detection starting at $N = 150$ (.0368), improving at $N = 250$ (.3148), and reaching strong detection at $N = 500$ (.8496). The

stochastic method performs better than the percentile method, detecting normality at $N = 150$ (.0362), improving to a rate of .32 at $N = 250$, and reaching strong detection at $N = 500$ (.871). The BCa method starts to detect normality at $N = 150$ (.014) and increases to .7886 at $N = 500$, thus performing the poorest of all methods. Fisher's method begins to detect normality only at $N = 500$, with perfect detection.

For $g = 0$, $h = -0.1$ when the NEB is liberal at .95, all methods initially detect normality at $N = 50$ and converge on near-perfect detection at $N = 250$. The percentile method starts with very low detection at $N = 30$ (.0024) and improves at $N = 75$ (.303), reaching .9516 at $N = 150$. The parametric normal interval performs slightly better than the percentile method, starting at a rate of .0088 at $N = 30$, improving to .3748 at $N = 75$, and strong detection of .9614 at $N = 150$. The stochastic method performs the best of the three methods, detecting at a higher rate at $N = 50$ (.0134) increasing to .3742 at $N = 75$ and .7194 at $N = 100$. By $N = 150$, the stochastic method achieves .9654 detection. The BCa similarly performs the poorest, starting to detect normality at $N = 75$ (.222) and increasing to .9256 at $N = 150$. Fisher's method detects normality starting at $N = 100$, with perfect rates therein.

Given that the Fisher's *r* to Z results display a rapid increase in power for normal and negligibly different from normal conditions, a supplemental table is included in Appendix A (Tables A1 – A6) with more sample sizes to show granular increases in the proportion of replications in which Fisher's *r* to *z* transformation correctly identifies the distribution as negligibly different from normal. The tables identify the range of sample sizes at which the rates steeply increase in power. Such rapid escalations in power with increased sample sizes relative to the other CI approaches is unexpected and undesirable for negligible effect testing, as more tempered increases are typically expected under normal conditions, suggesting Fisher's *r* to *z* is

over-sensitive and gains power too quickly. For the other methods of calculating the CI, the proportion of correct conclusions rises gradually across increasing larger sample sizes, and this pattern is better aligned with the expectations of negligible effect testing.

These Fisher's *z* approach results are as follows: for a normal distribution ($g = 0$, $h = 0$) and NEB = .975, starting from a sample size of $N = 145$, rates begin to rise (.182) and reach .835 at $N = 150$. For a NEB = .95, rates start to increase steeply around $N = 75$, from .348 to 1 by $N = 80$. In both cases, the steep incline with only 5 to 10 added observations reflects instability and is contrary to expected gradual growth.

When $g = 0.2$, $h = 0$ (NEB = .975), rates start at .007 at $N = 310$ and increase to 1 at $N = 400$. When NEB = .95, this increase begins at $N = 101$ and reaches 1 at $N = 120$. Power increases here are slightly less rapid compared to a normal distribution. When $g = 0$, $h = -0.1$, NEB = .975, a similar pattern is observed as the normal distribution, with a steep incline beginning at $N = 250$ (.001) and reaching 1 by $N = 300$. When NEB = .95, proportions are minimal at $N = 86$, with a rapid climb starting at $N = 88$, reaching 1 at $N = 95$.

## Applied Example (Study 2)

To follow up from the applied example in Study 1, the stochastic bootstrap for the NET-SW using the liberal criterion was used in place of the percentile bootstrap approach to illustrate improved power of the stochastic approach. As in Study 1, data are from an RCT-based study on the effectiveness of an online cognitive-behavioral therapy (CBT) for maladaptive perfectionism among a sample of undergraduate students with high levels of perfectionism (Arpin-Cribbie et al., 2012). The dataset is available within the `negligible` package in R (Cribbie et al., 2023). Although the stochastic bootstrap method was adopted in Study 2 to address limitations of the percentile method in Study 1, reapplying the stochastic method to the NET-SW adopted in the

Applied Example from Study 1 did not yield different conclusions—both the percentile and the stochastic bootstrap methods failed to reject the null hypothesis (of non-normality), concluding insufficient evidence that the sample distribution was negligibly different from normal. As such, a new applied example was used in Study 2 to better demonstrate the improved sensitivity of the stochastic approach.

In this example, a multiple linear regression model was fit using scores on Center for Epidemiological Studies Depression (CESD) scale (Radloff, 1977) as the outcome variable. There were two predictors; 1) socially prescribed perfectionism subscale scores from the Multidimensional Perfectionism Scale (Hewitt & Flett, 1991) and 2) anxiety scores on the Beck Anxiety Inventory (Beck et al., 1998).

All variables were measured at pretest. The model residuals were extracted ($N$ = 83) and tested for normality using the `neg.normal` function from the `negligible` package (Cribbie et al., 2023) with a NEB of $W$ = .950.

The residuals are characterized by minimal skewness (0.29) and kurtosis (-0.23; Figure 12). The 20% trimmed mean is -0.43. If the NET-SW concludes that the distribution is not negligibly different from a normal distribution, these descriptive statistics may help a researcher choose between a transformation, outlier treatment, or a robust approach. The traditional SW test is nonsignificant, $W$ = .984, $p$ = .38. In other words, the traditional null hypothesis that the distribution from which the data were drawn is normal cannot be rejected at the $\alpha$ = .05 level. However, the NET-SW test using the percentile bootstrap method had a lower bound of the 95% CI is .949, which is below the lower bound of the negligible effect (equivalence) interval of .950 (liberal criterion). Thus, the NET-SW test using a percentile bootstrap concludes that the null hypothesis that the degree of nonnormality is extreme cannot be rejected (i.e., the distribution is

not concluded to be negligibly different from normal). However, when the stochastic approach to calculating the CI is adopted, the lower bound of the 95% CI is .953, leading to a rejection of the null hypothesis that the degree of nonnormality is extreme, concluding that the distribution is negligibly different from normal.

Thus, the results of the traditional SW and the NET-SW using a percentile bootstrap CI provided conflicting results. Specifically, the traditional SW failed to find evidence that the distribution deviated significantly from normal, while the NET-SW failed to find evidence of negligible differences from normality. However, the results with the stochastic bootstrap CI for the NET-SW align more closely with that of the traditional SW test because the stochastic approach concludes that the distribution is negligibly different from normal. It is important to reiterate that these tests need not give consistent conclusions, as they assess differing hypotheses; the traditional SW tests the null hypothesis that the distribution is not significantly different from a theoretical normal distribution, whereas the NET-SW assesses a null hypothesis that the distribution *is* meaningfully different from a normal distribution. If the traditional SW fails to reject the null, this result does not suggest that evidence favours a normal distribution. Similarly, if the NET-SW fails to the reject the null, this result does not indicate the evidence favours a distribution that is meaningfully different from normal.

## Discussion (Study 2)

This study expanded on the exploration of procedures for detecting normality using NETs, focusing on the NET-SW, a modification of the traditional Shapiro-Wilk test, and methods to calculate CIs for the NET-SW. The NET-SW developed in Study 1 assesses whether the deviation of a sample distribution from a normal distribution is negligible, contrasting it with traditional tests for normality (e.g., SW and KS tests), which assess whether deviations are significantly different from normal. In this context, the NEI defines a range within which deviations from normality are considered negligible. For example, the NEI from the NET-SW may be set between .95 as a lower bound and an upper bound of $W$ =1 representing perfect normality.

Study 1 found that the NET-SW performed well compared to the NET-KS with respect to balancing Type I errors and power. However, a key limitation was uncovered: the bootstrap percentile CI approach used to estimate the CI for the NET-SW was too conservative when the sample distribution was normal or close to normal. This method led to reductions in power for detecting negligible deviations from normality and it was common for the bootstrap CI to not include the sample $W$, especially as sample size increased ($W$ was greater than the upper bound of the CI). Specifically, with normal or near-normal distributions, the distribution of bootstrap $W$ statistics is negatively skewed, shifting the lower bound of the CI below the sample $W$. In other words, with normal distributions, the NET-SW may fail to detect negligible deviations from normality, leading to an underpowered test.

The primary issue explored in Study 2 was alternative methods of CI estimation for the NET-SW. Four alternative bootstrapping methods were examined to improve the CI estimation in the NET-SW procedure (i.e., stochastic bootstrap, parametric bootstrap, Fisher's $r$ to $z$

transformation, and bias-corrected and accelerated bootstrap). Monte Carlo simulations evaluated the performance of these methods (power or true detection of negligible deviations from normality, and Type I errors or incorrect conclusions of normality when deviations were meaningful) under varying sample sizes and distribution shapes (normal, near-normal, and non-normal). As in Study 1, dual NEBs were adopted (conservative = .975 and liberal = .95) to evaluate sensitivity to detections of negligible deviations from normality and Tukey *g*-and-*h* distributions allowed for control over skew and kurtosis. The rates that each method concluded negligible differences from normality were tracked, which meant a Type I error condition if the population distribution had *W* less than or equal to the lower bound of the NEI or a power condition if the distribution had *W* greater than the lower bound of the NEI.

For normal distributions, Fisher's *r* to *z* performed the best in samples greater than 75, while the parametric and stochastic methods had similar performance, having the best rates in samples of $N = 75$ (for an NEB = .95). Both methods performed better than the percentile method. In distributions that were negligibly different from normal, again the stochastic and parametric methods had the best power rates in samples up to $N = 100$ (NEB = .95), while Fisher's *r* to z had the best rates above $N = 100$. For distributions in which deviations were substantive ($W \leq \text{NEI}_{\text{lower}}$), rates of false conclusions of normality were maintained below α across all methods and sample sizes. For conditions where the population *W* parameter was at the lower NEB, all methods maintained error rates very close to α.

The stochastic percentile bootstrap, parametric normal, and Fisher's *r* to *z* were the most reliable across various conditions. Both had strong power rates and conservative Type I error rates and consistent power in mid to large sample sizes. As expected from Study 1, the percentile bootstrap method showed greater variability, especially with smaller sample sizes, and showed a

slow increase in power rates. The parametric method works well up to moderate sample sizes but does not perform optimally in larger sample sizes. The BCa performed worse to the percentile bootstrap method and so both methods should be avoided.

As mentioned above, the BCa method performed worse than the percentile bootstrap method. This finding could be explained by that fact that if the original sampling distribution is negatively skewed (i.e., like for the *W* statistic of bootstrap samples when the population being sampled is approximately normally distributed), the adjusted BCa CI will be shifted to the left (i.e., downwards) (DiCiccio & Efron, 1996). Due to negative skewness, the estimate of average $\bar{W}$ may be closer to the lower tail. The bias correction parameter of the BCa adjusts for asymmetry in the bootstrap distribution while the acceleration constant accounts for the rate of change in the standard error for the estimated *W* parameter. It measures the sensitivity of the standard error to changes in the true parameter value. The standard error may be small near the bulk of the sampling distribution, but larger near the tails based on a few cases, and so the acceleration constant accounts for this fluctuation in the standard error. The width of the BCa CI is adjusted depending on the asymmetry of the distribution. If the distribution is negatively skewed, the acceleration parameter will shift the lower bound of the CI in the negative direction, resulting in a lower bound that is moved downwards to better estimate the true *W*, lest the longer tail underestimates the range of possible lower values. This correction is meant to compensate for lack of coverage of the lower tail when using traditional percentile bootstrap intervals. In the present research context, to improve the rates at which the NET-SW detects negligible deviations from normality, the lower CI bound must shift to the right so that the lower bound is higher relative to what it is in the percentile CI. Thus, the BCa method is ineffective for dealing with the conservative nature of the percentile CI. The bias-correction term that adjusts for asymmetry in

the *W* distribution will likely shift the CI interval leftwards when the distribution is negatively skewed due to a higher density in the upper range of the bootstrap distribution. The BCa method corrects the CI by shifting the lower bound downwards in consideration of the longer left tail.

The purpose of this study was to explore alternative options to calculating confidence intervals for a negligible effect test for normality based on the SW test. Specifically, the NET-SW with CIs calculated using a percentile confidence interval approach has reduced rates to conclude negligible differences from normality when the sample is approximately normally distributed. The stochastic bootstrap method should be adopted when calculating NET-SW CIs given its power and reliable performance across conditions.

# CHAPTER FOUR

# GENERAL DISCUSSION

This research introduced negligible effect testing (NET) for distributional normality. The proposed tests were designed to conclude normality when a target distribution is negligibly different from normal. In contrast to traditional null-hypothesis methods, such as the Shapiro-Wilk or Kolmogorov-Smirnov tests, which assess whether a sample distribution significantly differs from a theoretical normal distribution, NETs for normality evaluate whether differences between a target distribution and theoretical normal distribution are so small that they may be discounted.

Simulations compared the proposed NET-based normality procedures (NET-SW and NET-KS) to the traditional tests, using both liberal (.95) and conservative (.975) NEBs. Results showed that the traditional normality tests lack statistical power to detect meaningful deviations from normality when sample sizes are small and tend to conclude significant deviations from normality in large samples even when the differences are negligible. On the other hand, NET-based normality tests maintain low errors rates for both negligible and meaningful deviations, rarely falsely concluding normality when deviations were substantive. In normal and negligibly different from normal conditions, the NET-based normality tests require moderate to large sample sizes (e.g., $N > 100$) to have sufficient power to detect normality. Specifically, the NET-SW with a liberal criterion outperformed other NETs in terms of the best balance of Type I error and power rates. The primary limitation of this test in Study 1 was the conservative percentile bootstrap CIs when the target distribution is normal or near-normal, leading to a reduction in power to detect negligible differences from normality.

In Study 2, several alternative methods were examined for calculating the CI for the NET-SW procedure. Monte Carlo simulations assessed the performance of each CI method across varying sample sizes and distribution conditions using stochastic, parametric, Fisher's *r* to *z*, and bias-corrected and accelerated (BCa) methods. The BCa and percentile methods performed similarly and were poorest due to their inability to adjust for skewness in the sampling distribution in near-normal conditions. Fisher's *r* to *z* transformation was too sensitive in moderate sample sizes, with large leaps in statistical power within small sample size ranges. The stochastic bootstrap approach performed the most consistently and reliably across conditions, improving detection of negligible deviations from normality while still minimizing false conclusions of normality when deviations from normality were meaningful, especially in moderate to large samples. Together, the two studies underscore the utility of NETs for detecting negligible differences between a target distribution and a theoretical normal distribution. Specifically, the NET-SW based on the stochastic bootstrap method and liberal NEI is a fitting alternative to traditional tests for detecting when a target distribution differs negligibly from normal.

# REFERENCES


Adkins, M. C. (2017). Best practices for constructing confidence intervals for the general linear model under non-normality (Unpublished master's thesis). York University, Toronto.

Algina, J., Oshima, T. C., & Lin, W.-Y. (1994). Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups. *Journal of Educational Statistics, 19*(3), 275–291. https://doi.org/10.3102/10769986019003275

Alter, U., & Counsell, A., (2023). Determining negligible associations in regression. *The Quantitative Methods for Psychology*, *19*(1), 59-83. https://www.tqmp.org/RegularArticles/vol19-1/p059/

Altman, D. G., & Bland, J. M. (1995). Statistics notes: the normal distribution. *BMJ (Clinical research ed.), 310*(6975), 298. https://doi.org/10.1136/bmj.310.6975.298

Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises Criterion. *The Annals of Mathematical Statistics, 33*(3), 1148-1159. https://doi.org/10.1214/aoms/1177704477

Anderson, T. W., & D. A. Darling. (1952). Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. *Annals of Mathematical Statistics, 23*(2), 193-212. https://doi.org/10.1214/aoms/1177729437

Anscombe, F. J., & Glynn, W. J. (1983). Distribution of the kurtosis statistic $b_2$ for normal samples. *Biometrika, 70*(1), 227–234. https://doi.org/10.1093/biomet/70.1.227

Arpin-Cribbie, C., Irvine, J., & Ritvo, P. (2012). Web-based cognitive-behavioral therapy for perfectionism: A randomized controlled trial. *Psychotherapy Research: Journal of the*

*Society for Psychotherapy Research, 22*(2), 194–207. https://doi.org/10.1080/10503307.2011.637242.

Bauer, D. J., and Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: performance of alternatives specifications and methods of estimation. *Psychological Methods, 16*(4), 373–390. https://doi.org/10.1037/a0025813

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology, 56*(6), 893-897. https://doi.org/10.1037/0022-006X.56.6.893

Beribisky, N., Alter, U., & Cribbie, R. A. (2019). *A multi-faceted mess: A systematic review of statistical power analysis in psychology journal articles* [Preprint]. PsyArXiv. https://osf.io/3bdfu

Beribisky, N., & Cribbie, R. A. (2023). Evaluating the performance of existing and novel equivalence tests for fit indices in structural equation modelling. *British Journal of Mathematical and Statistical Psychology, 77*(1), 103–129. https://doi.org/10.1111/bmsp.12317

Beribisky, N., Mara, C. A., & Cribbie, R. A. (2020). An equivalence testing approach for evaluating substantial mediation. *The Quantitative Methods for Psychology, 16*(4), 424–441. https://doi.org/10.20982/tqmp.16.4.p424

Bishara, A. J., & Hittner, J. B. (2017). Confidence intervals for correlations when data are not normal. *Behavioral Research, 49*, 294-309. https://doi.org/10.3758/s13428-016-0702-8

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*: *European Journal of Research Methods for the Behavioral Social Sciences, 9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Boneau, C. A. (1960). The effects of violations of assumption underlying the *t* test. *Psychological Bulletin, 57*(1), 49-64. https://doi.org/10.1037/h0041412

Bono, R., Blanca, M.J., Arnau, J., & Gómez-Benito, J. (2017). Non-normal distributions commonly used in health, education, and social sciences: A systematic review. *Frontiers in Psychology*, 8:1602. https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01602/full

Bower, P., Gilbody, S. (2005). Stepped care in psychological therapies: access, effectiveness and efficiency: Narrative literature review. *British Journal of Psychiatry,186*(1), 11-17. doi:10.1192/bjp.186.1.11

Chaffin, W. W., & Rhiel, S. G. (1993). The effect of skewness and kurtosis on the one-sample *T* test and the impact of knowledge of the population standard deviation. *Journal of Statistical Computation and Simulation, 46*(1–2), 79–90. https://doi.org/10.1080/00949659308811494

Chen, L. T., Peng, C. Y. J. (2015). The sensitivity of three methods to nonnormality and unequal variances in interval estimation of effect sizes. *Behavioral Research, 47*, 107–126. https://doi.org/10.3758/s13428-014-0461-3

Chernick, M. R., & LaBudde, R. A. (2011). *An introduction to bootstrap methods with applications to R* (1st ed.). Wiley.

Cornea-Madeira, A., & Davidson, R. (2015). A parametric bootstrap for heavy-tailed distributions. *Econometric Theory*, *31*(3), 449–470. http://www.jstor.org/stable/24537626

Counsell, A., & Cribbie, R. A. (2015). Equivalence tests for comparing correlation and regression coefficients. *British Journal of Mathematical and Statistical Psychology, 68*(2), 292–309. https://doi.org/10.1111/bmsp.12045

Counsell, A., Cribbie, R. A., & Flora, D. B. (2020). Evaluating equivalence testing methods for measurement invariance. *Multivariate Behavioral Research, 55*(2), 312–328. https://doi.org/10.1080/00273171.2019.1633617

Counsell, A., Ragoonanan, C., & Cribbie, R. A. (2016). Testing for negligible interaction: A coherent and robust approach. *British Journal of Mathematical and Statistical Psychology, 69*(2), 159-174. https://doi.org/10.1111/bmsp.12066

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal, 1928*(1), 13–74. https://doi.org/10.1080/03461238.1928.10416862

Cribbie, R., Alter, U., Beribisky, N., Chalmers, P., Counsell, A., Farmus, L., Martinez Gutierrez, N. (2023). *Negligible: A collection of functions for negligible effect/equivalence testing.* R package version 0.1.6. https://CRAN.R-project.org/package=negligible.

Cribbie, R., Fiksenbaum, L., Keselman, H., & Wilcox, R. (2012). Effect of non-normality on test statistics for one-way independent groups designs: Effects of non-normality on test statistics. *British Journal of Mathematical & Statistical Psychology*, *65*(1), 56–73. https://doi.org/10.1111/j.2044-8317.2011.02014.x

Curran-Everett, D., Taylor, S., & Kafadar, K. (1998). Fundamental concepts in statistics: Elucidation and illustration. *Journal of Applied Physiology, 85*(3), 775-786. https://doi.org/10.1152/jappl.1998.85.3.775open_in_new

D'Agostino, R. B. (1970). Transformation to normality of the null distribution of $g_1$. *Biometrika, 57*(3), 679–681. https://doi.org/10.1093/biomet/57.3.679

D'Agostino, R. B., & Belanger, A. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44(4), 316–321. https://doi.org/10.2307/2684359

D'Agostino, R. B., & Pearson, E. S. (1973). Tests for departure from normality. Empirical results for the distributions of $b_2$ and $\sqrt{b_1}$. *Biometrika, 60*(3), 613–622. https://doi.org/10.1093/biomet/60.3.613

Darlington, R. B., & Hayes, A. F. (2017). *Regression analysis and linear models: Concepts, applications, and implementation*. The Guilford Press.

Davison, A.C., & Hinkley, D.V. (1997). *Bootstrap methods and their application (Cambridge series in statistical and probabilistic mathematics, series number 1)* (1st ed.). Cambridge University Press

Delaney, H. D., & Vargha, A. (2000). The effect of nonnormality on Student's two-sample t test. *ERIC Document Reproduction Service No. ED443850*.

DiCiccio, T. J., Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*(3), 189 – 228. https://doi.org/10.1214/ss/1032280214

Dimidjian, S., Hollon, S. D., Dobson, K. S., Schmaling, K. B., Kohlenberg, R. J., Addis, M. E., Gallop, R., McGlinchey, J. B., Markley, D. K., Gollan, J. K., Atkins, D. C., Dunner, D. L., & Jacobson, N. S. (2006). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology, 74*(4), 658–670. https://doi.org/10.1037/0022-006X.74.4.658

Dixon, P.M. (2006). Bootstrap Resampling. In Encyclopedia of Environmetrics (eds A.H. El-Shaarawi and W.W. Piegorsch). https://doi.org/10.1002/9780470057339.vab028

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

Efron, B., & Tibshirani, R. J., (1994). *An Introduction to the Bootstrap.* Chapman & Hall/CRC.

Farmus, L., Beribisky, N., Martinez Gutierrez, N., Alter, U., Panzarella, E., & Cribbie, R. A. (2022). Effect size reporting and interpretation in social personality research. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues.* Advance online publication. https://doi.org/10.1007/s12144-021-02621-7

Ferguson, W., & Clapshaw, L. (2020). Study of mental health outcomes associated with different brands of venlafaxine at the Kumeu Medical Centre from January 2017 to October 2018. *Therapeutic Advances in Psychopharmacology, 10*, 2045125320927309. https://doi.org/10.1177/2045125320927309

Field, A., Miles, J., & Field, Z. (2016). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE Publications.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, *10*(4), 507–521. https://doi.org/10.2307/2331838

Fisher, R. A. (1921). On the ‘probable error’ of a coefficient of correlation deduced from a small sample, *Metron, 1,* 3-32. http://hdl.handle.net/2440/15169

Fox, J. (1991). *Regression diagnostics: An introduction*. Sage Publications.

Fox, J. (2015). *Applied regression analysis and generalized linear models* (3rd ed.). Sage Publications.

Flett, G. L., Hewitt, P. L., Blankstein, K. R., & Gray, L. (1998). Psychological distress and the frequency of perfectionistic thinking. *Journal of Personality and Social Psychology, 75(5),* 1363-1381. https://doi.org/10.1037/0022-3514.75.5.1363

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, *10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Review of Educational Research, 42(3), 237–288. https://doi.org/10.3102/00346543042003237

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology, 63*(Pt 3), 527–537. https://doi.org/10.1348/000711009X475853

Goldman, M., & Kaplan, D. M. (2017). Comparing distributions by multiple testing across quantiles or CDF values. *arXiv preprint arXiv:1708.04658*. https://arxiv.org/abs/1708.04658

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics, 17*(4), 315–339. https://doi.org/10.3102/10769986017004315

Hau, K. T., & Marsh, H. W. (2010). The use of item parcels in structural equation modelling: Non-normal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology*, *57*(2), 327-351. https://doi.org/10.1111/j.2044-8317.2004.tb00142.x

Hewitt, P. L., & Flett, G. L. (1991). Perfectionism in the self and social contexts: Conceptualization, assessment, and association with psychopathology. *Journal of*

*Personality and Social Psychology, 60*(3), 456–470. https://doi.org/10.1037/0022-3514.60.3.456

Hill, J. J., Kuyken, W., & Richards, D. A. (2014). Developing stepped care treatment for depression (STEPS): study protocol for a pilot randomised controlled trial. *Trials, 15*(1), 452. https://doi.org/10.1186/1745-6215-15-452

Hollon, S. D., & Kendall, P. C. (1980). Cognitive self-statements in depression: Development of an automatic thoughts questionnaire. *Cognitive Therapy and Research, 4(4)*, 383–395. https://doi.org/10.1007/BF01178214

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters, 6*(3), 255–259. https://doi.org/10.1016/0165-1765(80)90024-5

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement, 65*(1), 51–69. https://doi.org/10.1177/0013164404264850

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., Deering, K. N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods 13*(2), 110-129. https://doi.org/10.1037/1082-989X.13.2.110

Kharasch, E. D., Neiner, A., Kraus, K., Blood, J., Stevens, A., Schweiger, J., Miller, J. P., & Lenze, E. J. (2019). Bioequivalence and therapeutic equivalence of generic and brand bupropion in adults with major depression: A randomized clinical trial. *Clinical pharmacology and therapeutics, 105*(5), 1164–1174. https://doi.org/10.1002/cpt.1309

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods, 53*(6), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari,* 4, 83-91.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill Education.

Li, J. C. H. (2016). Effect size measures in a two-independent-samples case with nonnormal and nonhomogeneous data. *Behavioural Research, 48*, 1560–1574. https://link.springer.com/content/pdf/10.3758/s13428-015-0667-z.pdf

Lilliefors, H. W. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association, 62*(318), 399–402. https://doi.org/10.1080/01621459.1967.10482916

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579–619. https://doi.org/10.3102/00346543066004579

Lu, H.Y.K., & Young, G.A. (2012). Parametric bootstrap under model mis-specification, *Computational Statistics & Data Analysis*, *56*(8), 2410-2420. https://doi.org/10.1016/j.csda.2012.01.018.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23*:151–69. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Mara, C. A., & Cribbie, R. A. (2018). Equivalence of population variances: Synchronizing the objective and analysis. *Journal of Experimental Education, 86*(3), 442–457. https://doi.org/10.1080/00220973.2017.1301356

Marcoulides, K. M, & Yuan, K. H. (2017). New ways to evaluate goodness of fit: A note on using equivalence testing to assess structural equation models. *Structural Equation Modeling*: *A Multidisciplinary Journal*, *24*(1), 148-153. https://doi.org/10.1080/10705511.2016.1225260

Man, K., Schumacker, R., Morell, M., & Wang, Y. (2022). Effects of compounded nonnormality of residuals in hierarchical linear modeling. *Educational and Psychological Measurement*, *82*(2), 330–355. https://doi.org/10.1177/00131644211010234

Marfo, P., & Okyere, G. A. (2019). The accuracy of effect-size estimates under normals and contaminated normals in meta-analysis. *Heliyon*, *5*(6), e01838. https://doi.org/10.1016/j.heliyon.2019.e01838

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Lawrence Erlbaum Associates.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis, MN: University of Minnesota Press.

Mendes, M., & Pala, A. (2003). Type I error rate and power of three normality tests. *Information and Technology*, *2*(2), 135–139. https://scialert.net/abstract/?doi=itj.2003.135.139

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. https://doi.org/10.1037/0033-2909.105.1.156

Milan, L., Whittaker, J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Journal of the Royal Statistical Society Series C: Applied Statistics, 44*(1), 31–49. https://doi.org/10.2307/2986193

Mills, L., Cribbie, R. A., & Luh, W.M. (2009). A heteroscedastic, rank-based approach for analyzing 2 x 2 independent groups designs. *Journal of Modern Applied Statistical Methods, 8*(1), 322-336. https://doi.org/10.56801/10.56801/v8.i.426

Mohd Razali, N., & Bee Wah, Y. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21–33. https://www.nrc.gov/docs/ml1714/ml17143a100.pdf

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Addison-Wesley.

Ostrovski, V. (2022). On application of the Cramér-von Mises Distance for equivalence testing. *Journal of Statistics and Computer Science, 1*(1), 1-9. https://DOI: 10.47509 /JSCS.2022.v01i01.01

Ostrovski, V. (2023). Equivalence tests based on weighted $L_2$-distance between cumulative distribution functions. *Journal of Statistics and Computer Science, 2*(1), 147-159. https://DOI: 10.47509 /JSCS.2023.v02i02.06

Oztuna, D., Elhan, A. H., & Tuccar, E. (2006). Investigation of four different normality tests in terms of Type I error rate and power under different distributions. *Turkish Journal of Medical Sciences, 36*(3), pp. 171–176. https://dergipark.org.tr/tr/download/article-file/129239

Peebles, D., & Ali, N. (2015). Expert interpretation of bar and line graphs: The role of graphicacy in reducing the effect of graph format. *Frontiers in Psychology, 6*, 1673. https://doi.org/10.3389/fpsyg.2015.01673

Pek, J., Wong, O., & Wong, C. M. (2017). Data transformations for inference with linear regression: Clarifications and recommendations. *Practical Assessment, Evaluation, Research & Assessment, 22*(9), 1-11. https://openpublishing.library.umass.edu/pare/article/id/1619/

Pek, J., Wong, O., & Wong, A. C. M. (2018) How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Front. Psychol*. 9:2104. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6232275/

Puth, M., Neuhäuser, M., & Ruston, G. D. (2014). Effective use of Pearson's product-moment correlation coefficient. *Animal Behavior, 93*, 183-189. https://doi.org/10.1016/j.anbehav.2014.05.003

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21–33. https://www.nrc.gov/docs/ML1714/ML17143A100.pdf

Ripley, B. D. (1987). *Stochastic simulation*. Wiley Series in Probability and Mathematical Statistics.

Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2021). The percentile bootstrap: A primer with step-by-step instructions in R. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920911881

Royston, J. P. (1982). An extension of the Shapiro and Wilk's *W* test for normality to large samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 31*(2), 115–124. https://doi.org/10.2307/2347973

Royston, J. P. (1989) Correcting the Shapiro-Wilk *W* for ties. *Journal of Statistical Computation and Simulation, 31*(4), 237-249. https://doi.org/10.1080/00949658908811146

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*(6), 657–680. https://doi.org/10.1007/BF01068419

Shapiro, S. S., & Francia, R. E. (1972). An approximate analysis of variance test for normality. *Journal of the Americal Statistical Association, 67*(337), 215-216. https://doi.org/10.1080/01621459.1972.10481232

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*(3-4), 591–611. https://doi.org/10.2307/2333709

Shishkina, T., Farmus, L., & Cribbie, R. A. (2018). Testing for a lack of relationship among categorical variables. *The Quantitative Methods for Psychology, 14*(3), 167–179. https://doi.org/10.20982/tqmp.14.3.p167

Sikes, C., Stark, J. G., McMahen, R., & Engelking, D. (2017). A Single-Dose, Two-Way Crossover, Open-Label Bioequivalence Study of an Amphetamine Extended-Release

Oral Suspension in Healthy Adults. *Journal of Attention Disorders*, *24*(3), 414-419. https://doi.org/10.1177/1087054717743329

Smirnov, N. V. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics, 19*(2), 279–281. https://doi.org/10.1214/aoms/1177730256

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Iowa State University Press.

Spokoiny, V., & Zhilova, M. (2015). Bootstrap confidence sets under model misspecification. *The Annals of Statistics*, *43*(6), 2653–2675. http://www.jstor.org/stable/43818864

Steinert, C., Munder, T., Rabung, S., Hoyer, J., & Leichsenring, F. (2017). Psychodynamic therapy: As efficacious as other empirically supported treatments? A meta-analysis testing equivalence of outcomes. *American Journal of Psychiatry, 174*(10), 943–953. https://doi.org/10.1176/appi.ajp.2017.17010057

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons, *Journal of the American Statistical Association*, *69*(347), 730-737. https://doi.org/10.2307/2286009

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Mahwah, NJ: Routledge Academic.

Sun, W. R., & Cheung, S. F. (2020). The influence of nonnormality from primary studies on the standardized mean difference in meta-analysis. *Behavior Research Methods*, 52, 1552–1567. https://doi.org/10.3758/s13428-019-01334-x

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102*(4), 989–1004. https://doi.org/10.1037/a0019507

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1–26. https://doi.org/10.1111/1529-1006.001

Thadewald, T., & Büning, H. (2007). Jarque–Bera test and its competitors for testing normality – A power comparison. *Journal of Applied Statistics, 34*:1, 87-105. https://doi.org/10.1080/02664760600994539

Tukey, J. W. (1977) Modern techniques in data analysis. NSF-sponsored regional research conference at Southeastern Massachusetts University, North Dartmouth, MA.

von Mises, R. (1931). *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik.* Leipzig: F. Deuticke.

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of general internal medicine, 26*(2), 192–196. https://doi.org/10.1007/s11606-010-1513-8

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/EBK1439808184

Welz, T., Doebler, P., & Pauly, M. (2022). Fisher transformation based confidence intervals of correlations in fixed- and random-effects meta-analysis. *The British journal of mathematical and statistical psychology, 75*(1), 1–22. https://doi.org/10.1111/bmsp.12242

Westfall, P. H., & Henning, K. S. S. (2013). Texts in statistical science: Understanding advanced statistical methods. Boca Raton, FL: Taylor & Francis.

Wilcox, R. R. (1997). A bootstrap modification of the Alexander-Govern ANOVA method, plus comments on comparing trimmed means. *Educational and Psychological Measurement, 57*(4), 655–665. https://doi.org/10.1177/0013164497057004010

Wilcox, R. R. (2005). Introduction to robust estimation and hypothesis testing (2nd ed.). San Diego, CA: Academic Press.

Wilcox, R. R. (2012a). Introduction to robust estimation and hypothesis testing, (3rd ed.) San Diego, CA: Academic Press.

Wilcox, R. R. (2012b). Modern statistics for the social and behavioral sciences: A practical introduction. New York: Chapman & Hall/CRC Press.

Wilcox, R. R. & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8*(3), 254-274. https://doi.org/10.1037/1082-989X.8.3.254

Woller-Carter, M. M., Okan, Y., Cokely, E. T., & Garcia-Retamero, R. (2012). Communicating and distorting risks with graphs: An eye-tracking study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56*(1), 1723–1727. https://doi.org/10.1177/1071181312561345

Xiong, C., Ceja, C. R., Ludwig, C. J. H., & Franconeri, S. (2020). Biased average position estimates in line and bar graphs: Underestimation, overestimation, and perceptual pull. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 301–310. https://doi.org/10.1109/TVCG.2019.2934400

Yap, B. W. & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation, 81*(12), 2141-2155. https://doi.org/10.1080/00949655.2010.520163

Yazici, B., & Yolacan, S. (2006). A comparison of various tests of normality. *Journal of Statistical Computation and Simulation, 75*(2), 175–183. https://doi.org/10.1080/00949650410001669914

Yuan, K. H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 51*, 289–309. https://doi.org/10.1111/j.2044-8317.1998.tb00682.x

Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods, 21*(3), 405-426. http://dx.doi.org/10.1037/met0000080

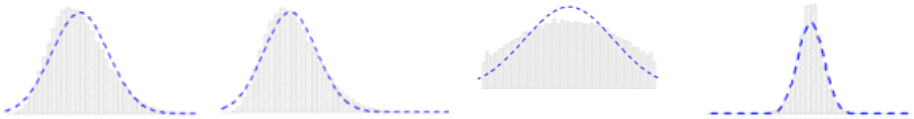
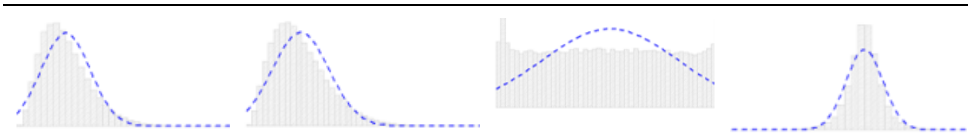
Zhai, Y., Wu, L., Zheng, Y., Wu, M., Huang, Y., Huang, Q., Shentu, J., Zhao, Q., & Liu, J. (2020). Bioequivalence study of amitriptyline hydrochloride tablets in healthy Chinese volunteers under fasting and fed conditions. *Drug Design, Development and Therapy, 14*, 3131–3142. https://doi.org/10.2147/DDDT.S258173

Zimmerman D. W., & Zumbo, B. D. (1992). Parametric alternatives to the student *t* test under violation of normality and homogeneity of variance. *Perceptual and Motor Skills, 74*(3), 835–44. https://doi.org/10.2466/pms.1992.74.3.835

TABLES

## Table 1

*Percentage Error Rates for Skewed, Platykurtic, and Leptokurtic Distributions using Conservative and Liberal Criterions for the Minimally Meaningful Effect Size when Testing a One-Sample* t *Test.*

| | | $\chi^2$ | ***g* & *h* (Skewed)** | ***g* & *h* (Platykurtic)** | ***g* & *h* (Leptokurtic)** |
|---|---|---|---|---|---|
| | | $\chi^2_{18}$ | $g$ = .3, $h$ = 0 | $g$ = 0, $h$ = -.15 | $g$ = 0, $h$ = .12 |
| **Conservative criterion for MMES** | $PE_{lower}$ | -113 | 42 | -13 | 16 |
| **SW ~ .975** | $PE_{upper}$ | 58 | -77 | -12 | 16 |
| **KS ~ 0.015** | $PE_{combined}$ | -27 | -17 | -13 | 16 |
| | |  | | | |
| | | $\chi^2_8$ | $g$ = .45, $h$ = 0 | $g$ = 0, $h$ = -.25 | $g$ = 0, $h$ = .18 |
| **Liberal criterion for MMES** | $PE_{lower}$ | -202 | 55 | -18 | 26 |
| **SW ~ .950** | $PE_{upper}$ | 73 | -135 | -18 | 27 |
| **KS ~ 0.025** | $PE_{combined}$ | -65 | -40 | -18 | 26 |
| | |  | | | |

*Note.* MMES = Minimally meaningful effect size; $PE_{lower}$ = percentage error in the lower tail. $PE_{upper}$ = percentage error in the upper tail; $PE_{combined}$ = percentage error across both tails combined.

**Table 2**

*Rates of Percentile Bootstrap Confidence Intervals Containing Observed* W *Using 5000 Simulations, 5000 Bootstrap Samples Per Simulation*

| **Conditions** | **30** | **50** | **75** | **100** | **150** | **250** | **500** | **1000** | **5000** |
|---|---|---|---|---|---|---|---|---|---|
| ***Negligible*** | | | | | | | | | |
| $g = 0$, $h = 0$ | .659 | .640 | .620 | .592 | .579 | .542 | .504 | .486 | .418 |
| $g = 0.2$, $h = 0$ | .768 | .809 | .855 | .898 | .943 | .986 | 1 | 1 | 1 |
| $g = 0$, $h = -0.1$ | .780 | .860 | .921 | .956 | .986 | 1 | 1 | 1 | 1 |
| ***Nonnegligible*** | | | | | | | | | |
| $g = 0.5$, $h = 0.1$ | .931 | .975 | .992 | .999 | 1 | 1 | 1 | 1 | 1 |
| $g = 0.6$, $h = 0$ | .991 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| ***At bounds*** | | | | | | | | | |
| $g = 0.4419$, $h = 0$ | .950 | .990 | .999 | 1 | 1 | 1 | 1 | 1 | 1 |
| $g = 0.2929$, $h = 0$ | .850 | .910 | .953 | .979 | .995 | 1 | 1 | 1 | 1 |

FIGURES

**Figure 1a**

*Probability Density Functions for Student's* t*-Distributions vs. Standard Normal Distribution*

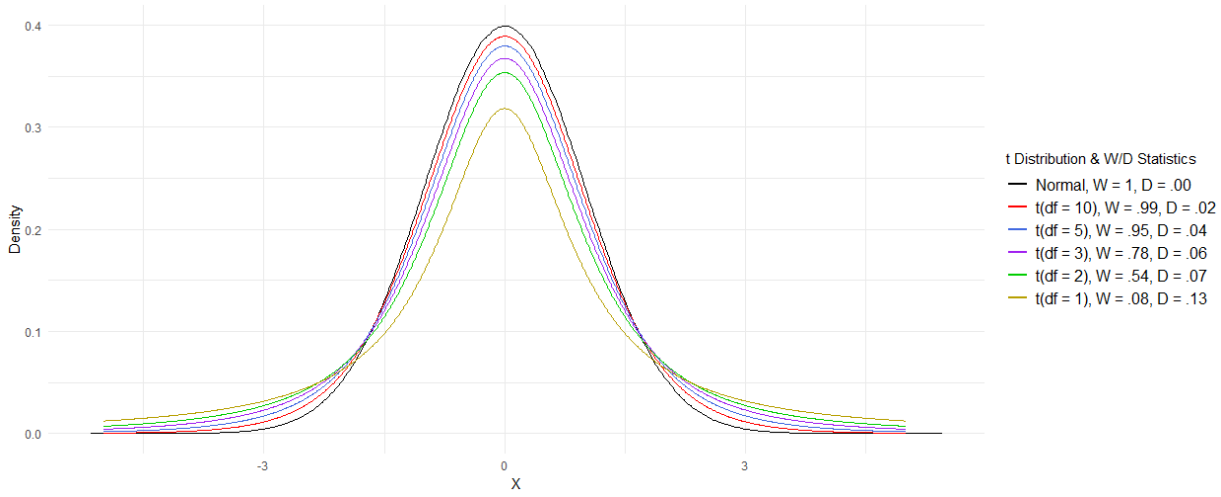


**Figure 1b**

*Differences in Cumulative Distribution Functions: Student's* t*-Distributions vs Standard Normal Distribution*

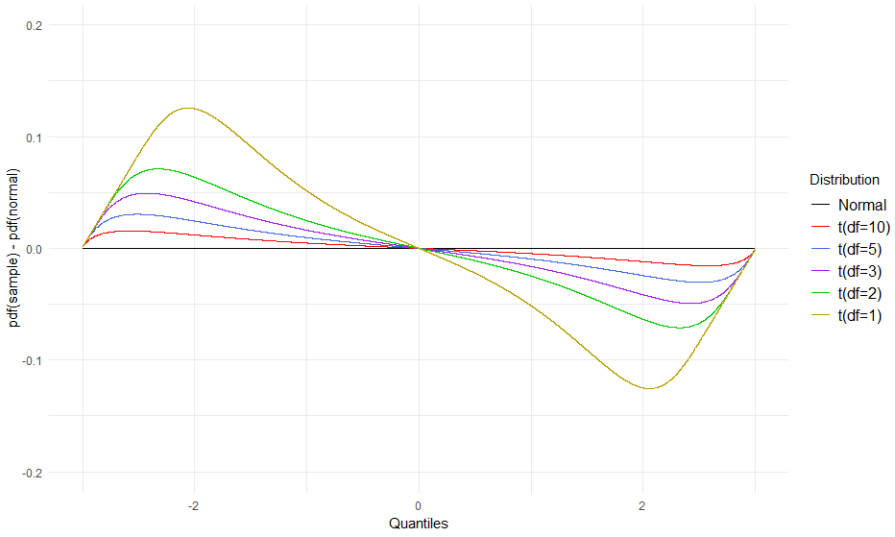


**Figure 2a**

*Probability Density Functions for $\chi^2$ Distributions vs. Standard Normal Distribution*

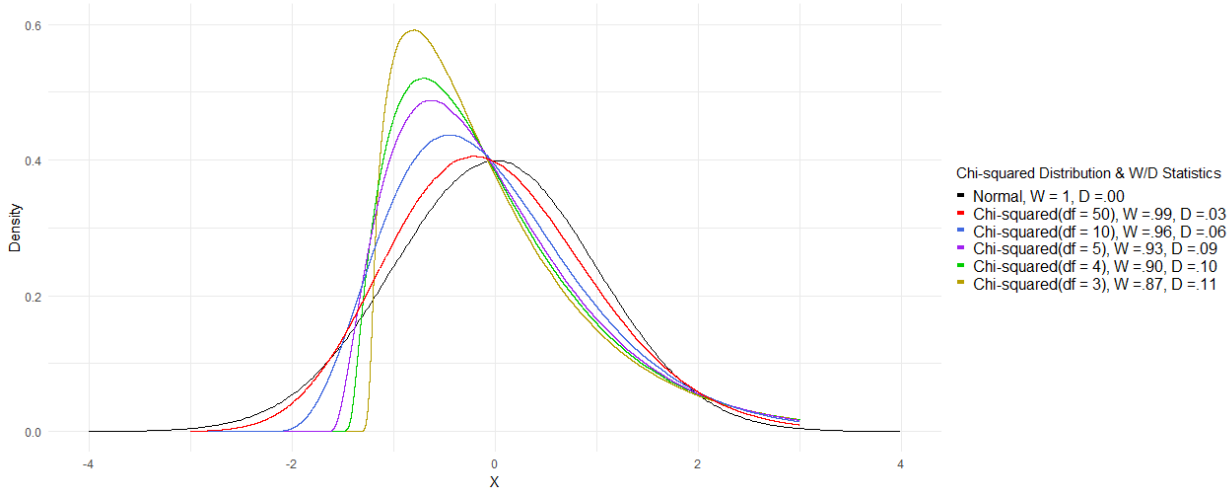


**Figure 2b**

*Differences in Cumulative Distribution Functions: $\chi^2$ Distributions vs Standard Normal Distribution*

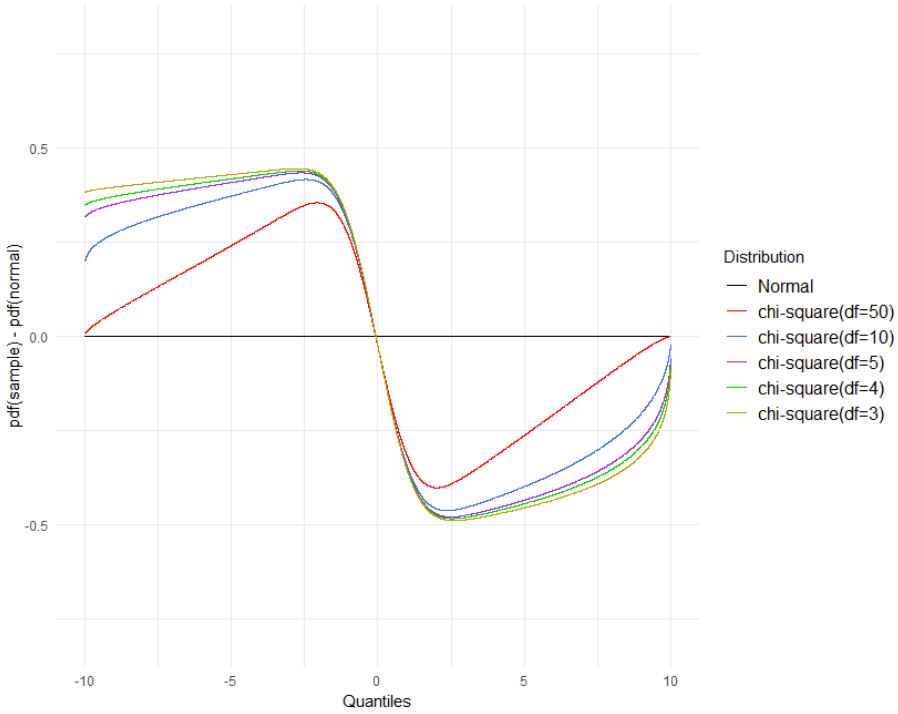
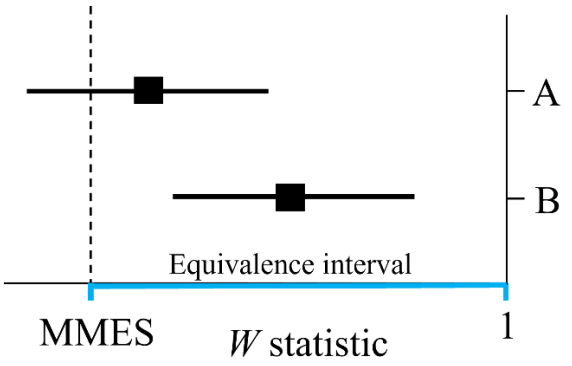


**Figure 3a**

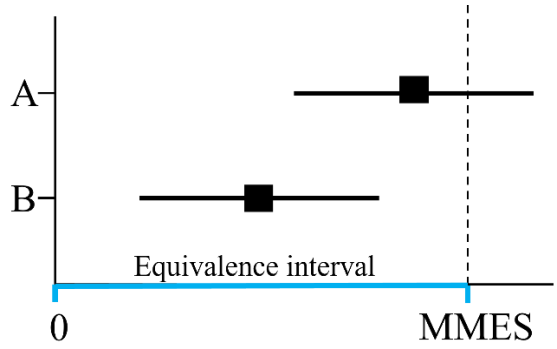
*Hypothesis Decisions for a Negligible Effect Test for Shapiro Wilk (NET-SW)*




*Note.* MMES = minimally meaningful effect size. A = example of observed *W* and CI when the lower bound of the 100(1-2α)% CI for NET-SW is not greater than the MMES, leading to a failure to reject the null hypothesis for the NET-SW. B = observed *W* and CI when the lower bound of the 100(1-2α)% CI for NET-SW is greater than the MMES, leading to rejection of the null hypothesis of the NET-SW.

**Figure 3b**

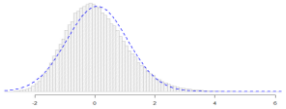
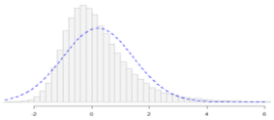
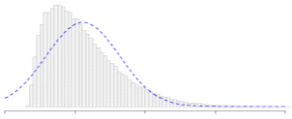
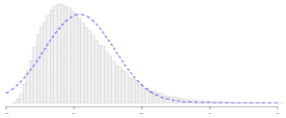
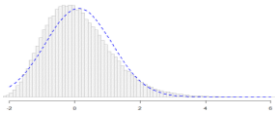
*Hypothesis Decisions for a Negligible Effect Test for Kolmogorov-Smirnov (NET-KS)*




*Note.* MMES = minimally meaningful effect size. A = example of observed *W* and CI when the upper bound of the 100(1-2α)% CI for NET-KS is greater than the MMES, leading to a failure to reject the null hypothesis for the NET-KS. B = observed *W* and CI when the lower bound of the 100(1-2α)% CI for NET-KS is less than the MMES, leading to rejection of the null hypothesis of the NET-KS.

**Figure 4**

*Monte Carlo Simulation Distribution Conditions*

| *g*-and-*h* distributions | Histogram |
|---|---|
| $g = 0, h = 0^1$ |  |
| $g = 0.2, h = 0^1$ |  |
| $g = 0, h = -0.1^1$ |  |
| $g = 0.5, h = 0.1^2$ |  |
| $g = 0.6, h = 0^2$ |  |
| $g = 0.4419, h = 0^3$ |  |
| $g = .29286, h = 0^4$ |  |

*Note*. The blue overlay represents a standard normal curve. 1 = negligibly different from normal; 2 = non-negligibly different from normal; 3 = at the liberal negligible effect bound (SW = .950, KS = .025); 4 = at the conservative negligible effect bound (SW = .975, KS = .015).

**Figure 5a**

*Histogram of a* g*-and-*h *Distribution with Skewness Parameter* g *= 0.2929 and Kurtosis Parameter* h *= 0 Overlaid with a Normal Distribution Curve, Representing the Conservative Criterion*

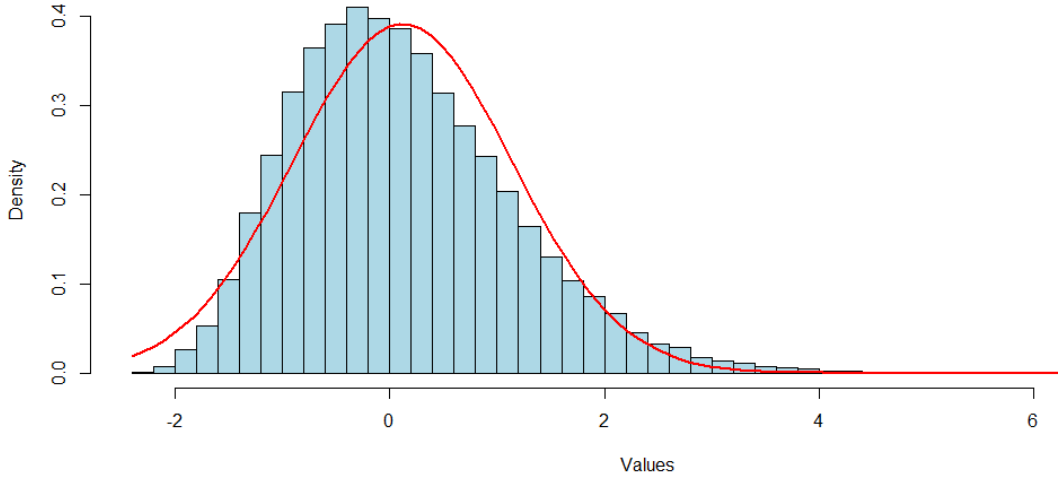



**Figure 5b**

*Histogram of a* g*-and-*h *Distribution with Skewness Parameter* g *= 0.4419 and Kurtosis Parameter* h *= 0 Overlaid with a Normal Distribution Curve, Representing the Liberal Criterion*

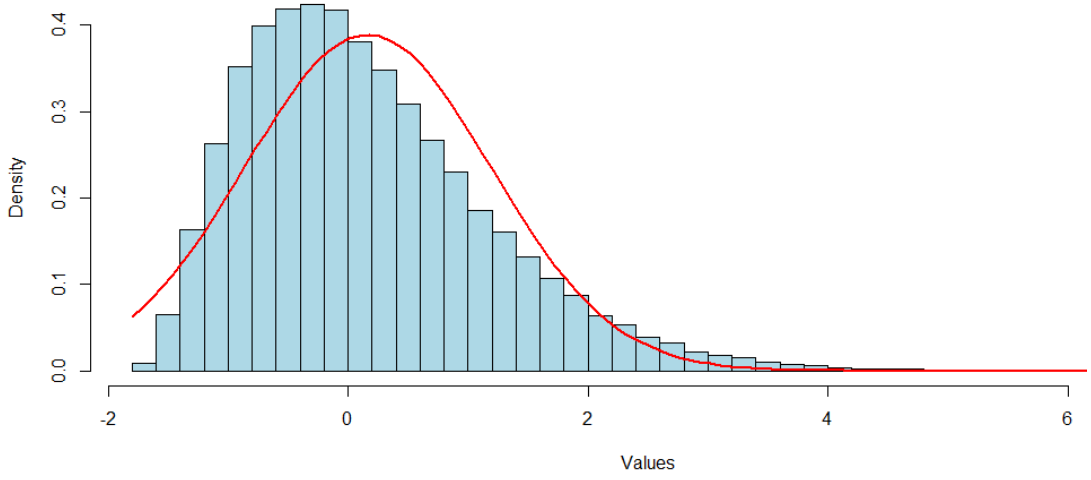
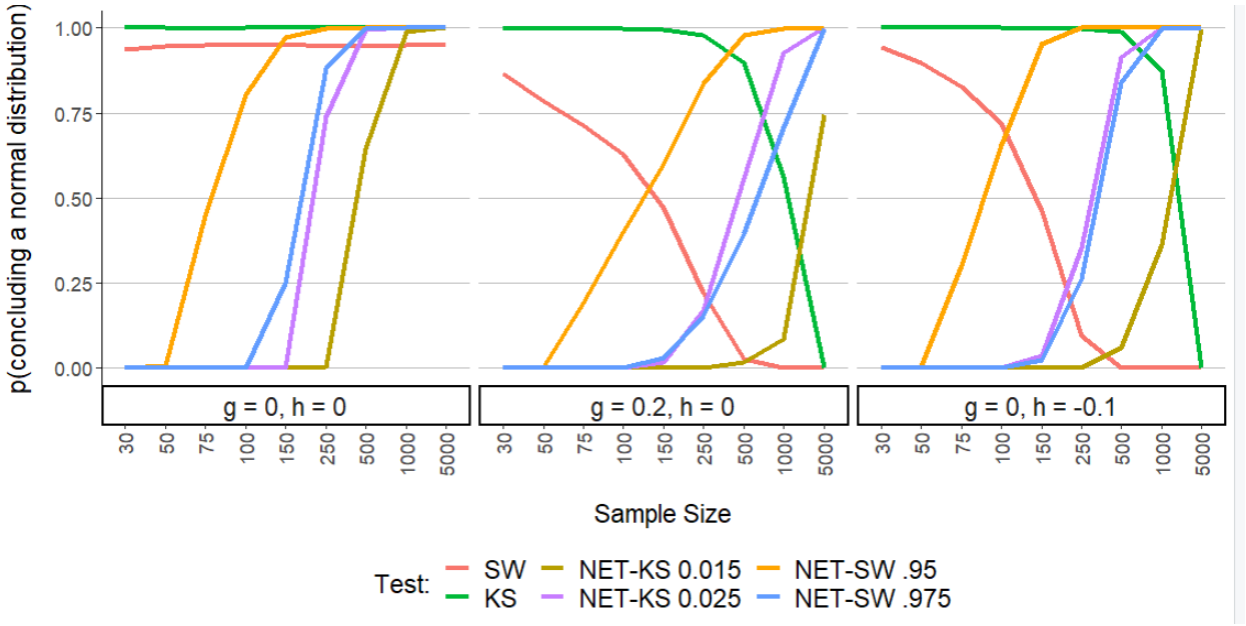


**Figure 6a**

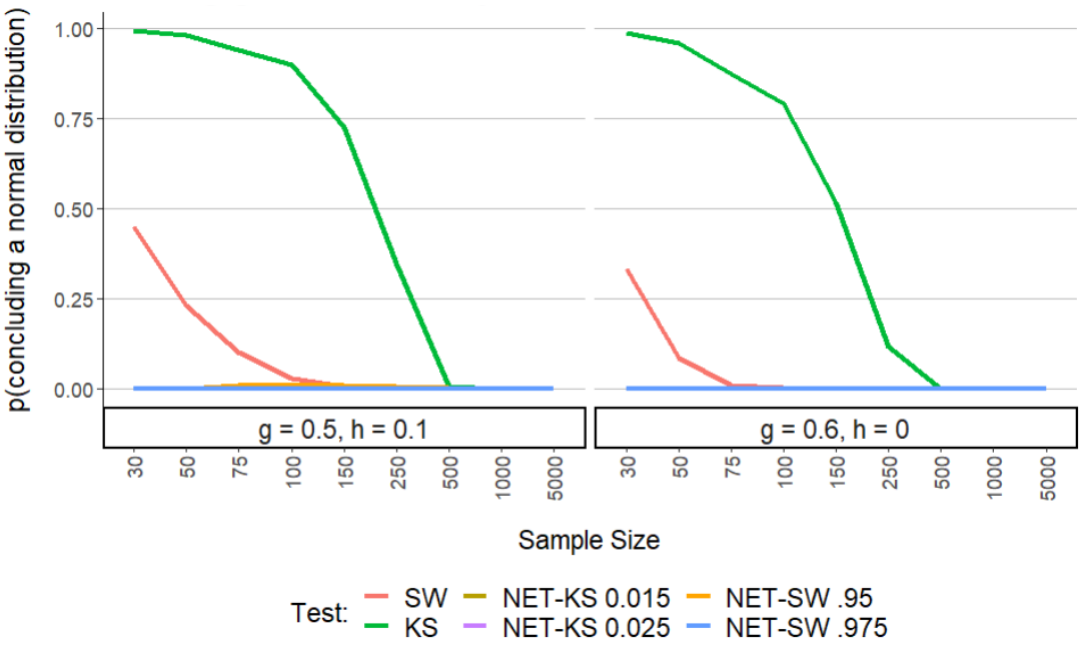
*Proportion of Conclusions of Normality Among Normal and Negligibly Different from Normal Conditions*




*Note.* SW = Shapiro-Wilk; KS = Kolmogorov-Smirnov; NET-KS = negligible effect test – Kolmogorov-Smirnov; NET-SW = negligible effect test – Shapiro-Wilk.

**Figure 6b**

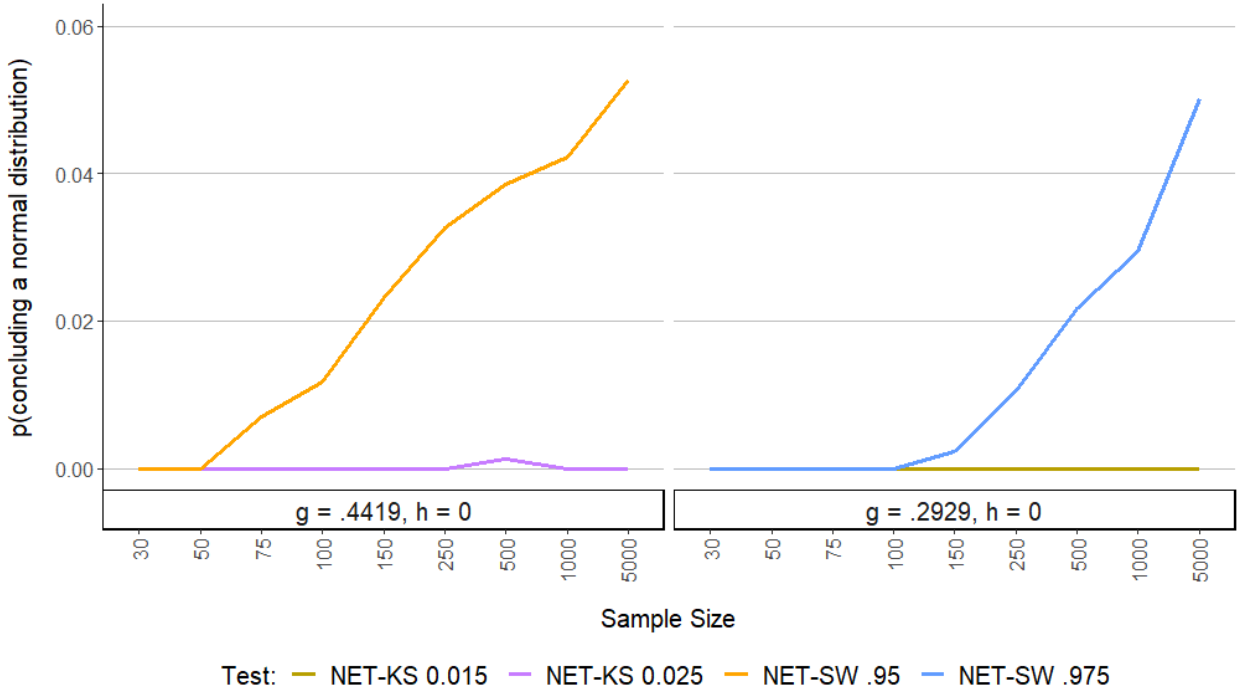
*Proportion of Conclusions of Normality Among Non-negligibly Different from Normal Conditions*




*Note.* SW = Shapiro-Wilk; KS = Kolmogorov-Smirnov; NET-KS = negligible effect test – Kolmogorov-Smirnov; NET-SW = negligible effect test – Shapiro-Wilk.

**Figure 6c**

*Proportion of Conclusions of Normality Among Nonnegligible Conditions at the Negligible Effect Boundary*




*Note.* NET-KS = negligible effect test – Kolmogorov-Smirnov; NET-SW = negligible effect test – Shapiro-Wilk.

**Figure 7**

*Model Residuals from Multiple Regression of Self-Oriented Perfectionism Predicted from Perfection Cognition and Automatic Thoughts*

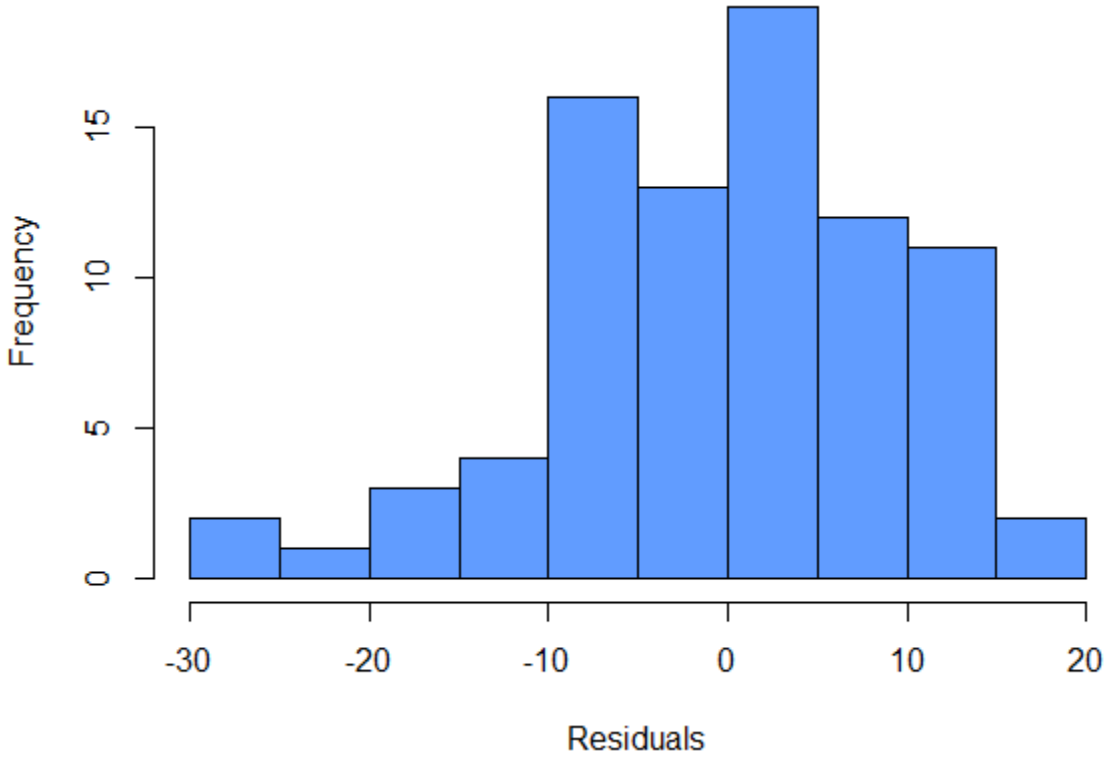
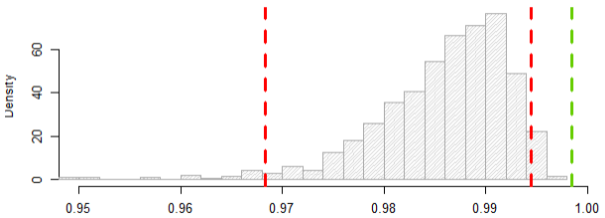
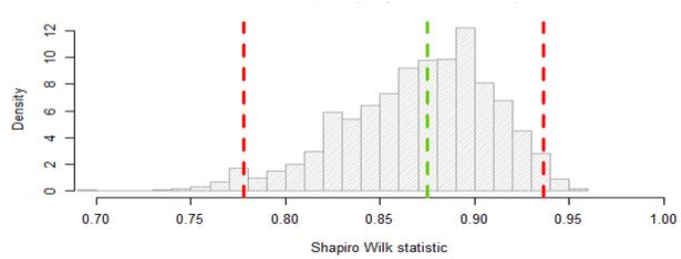


**Figure 8**

*Sampling Distributions of* W *Statistics when Parent Samples are Normal (A) and Nonnormal (B)*



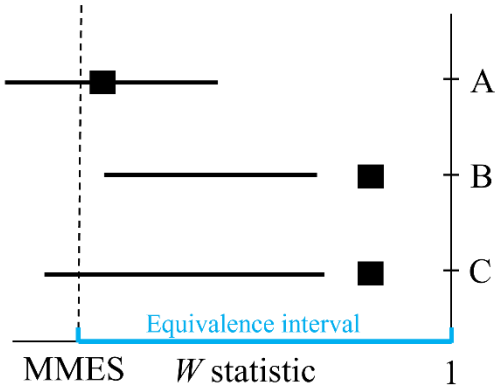

A. Sampling Distribution of *W* Statistic when *W* = .999, 95% Bootstrap CI: .969, .994.
*Note.* Green dashed line is observed *W*, and red dashed lines are percentile CI bounds.




B. Sampling Distribution of *W* Statistic when *W* = .875, 95% Bootstrap CI: .777, .936.
*Note.* Green dashed line is observed *W*, and red dashed lines are percentile CI bounds.

**Figure 9**

*Correct and Incorrect Decisions for a Negligible Effect Test for Normality*




*Note.* MMES = minimally meaningful effect size. A = observed *W* and CI when distribution is not normal (correct decision). B = observed *W* and CI when distribution is normal (correct decision, but CI does not contain observed *W*). C = observed *W* and CI when distribution is normal (incorrect decision and CI does not contain observed *W*).

**Figure 10**

*Comparison of Methods for Coverage of Sample Shapiro-Wilk* W *Statistics within Confidence Intervals Across Sample Sizes for a Normal Distribution (*g *= 0,* h *= 0)*

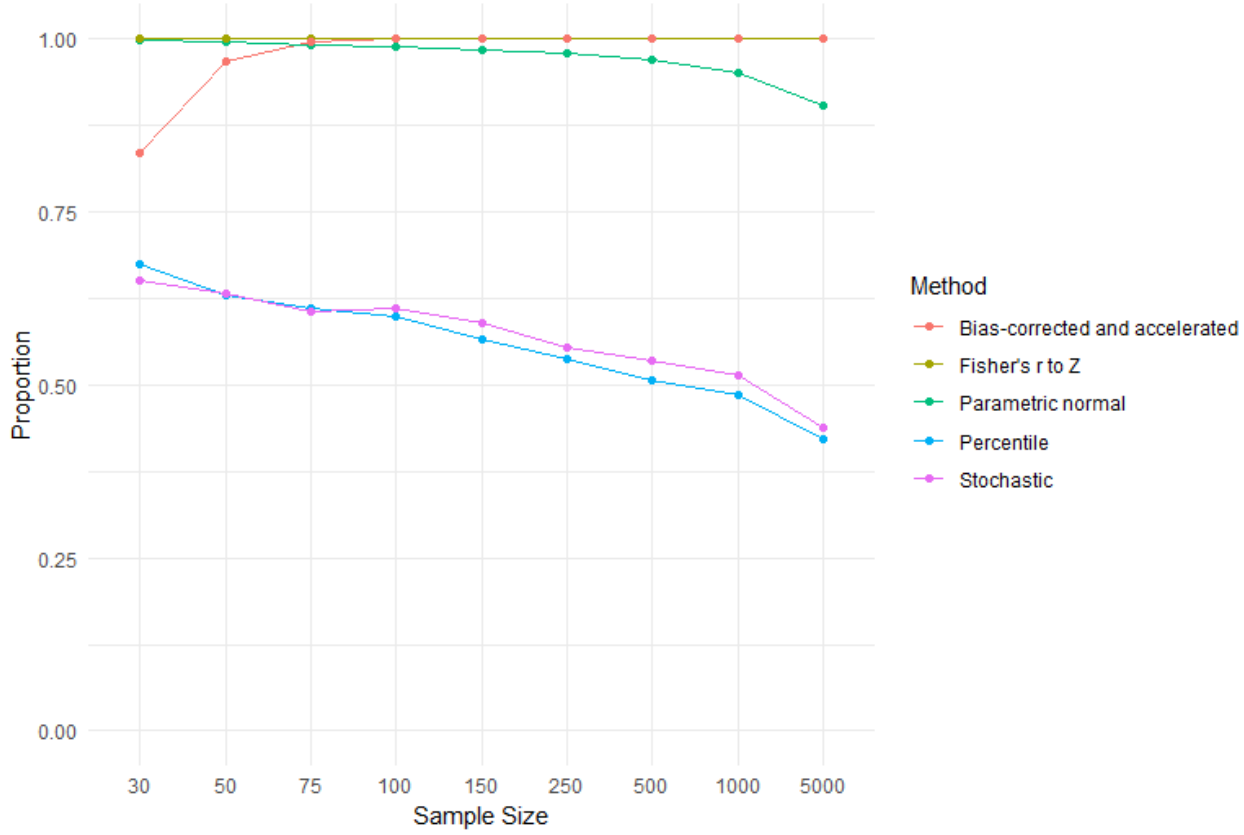


**Figure 11a**

*Proportion of Conclusions of Normality Across Confidence Interval Estimation Methods for On-Bound Conditions with Conservative Negligible Effect Bound (NEB) of .975 (*g *= 0.2929,* h *= 0) and Liberal NEB of .95 (*g *= 0.4419,* h *= 0)*

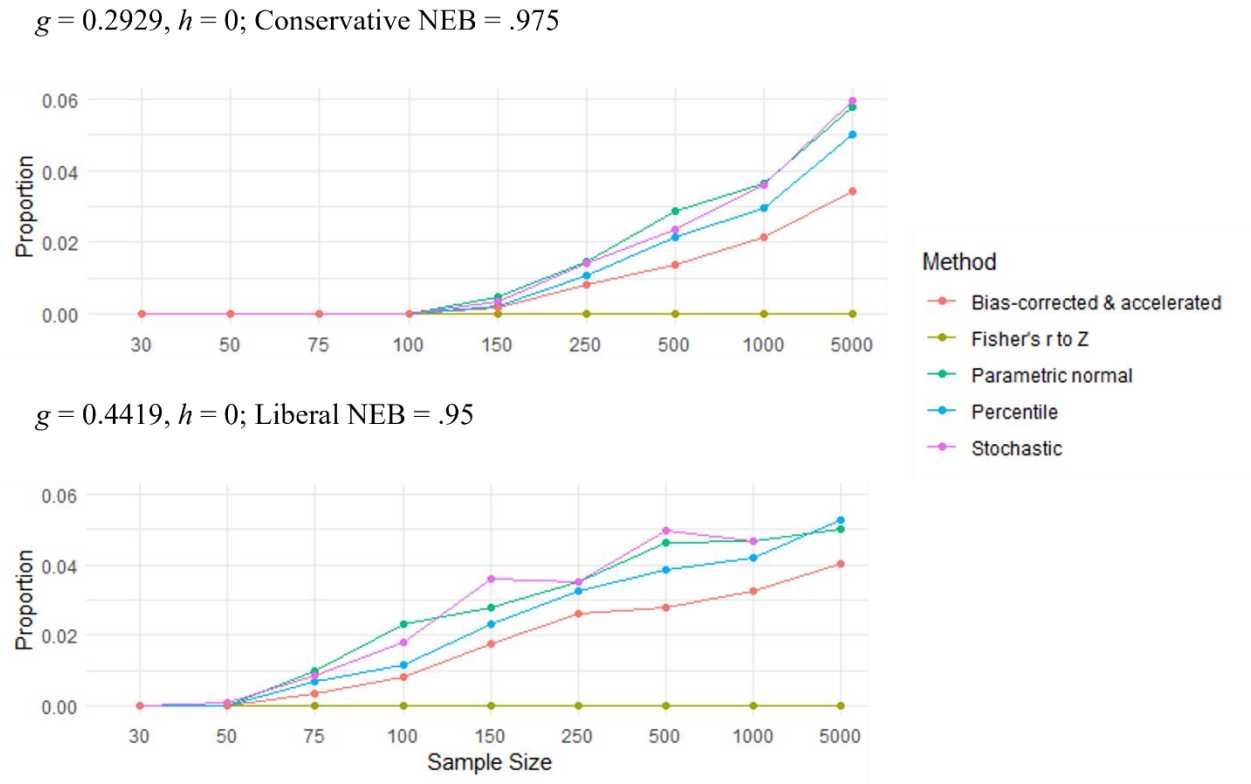


**Figure 11b**

*Proportion of Conclusions of Normality Across Confidence Interval Estimation Methods for Non-Negligibly Different from Normal Condition (*g *= 0.5,* h *= 0.1) with Conservative Negligible Effect Bound (NEB) of .975 and Liberal NEB of .95*

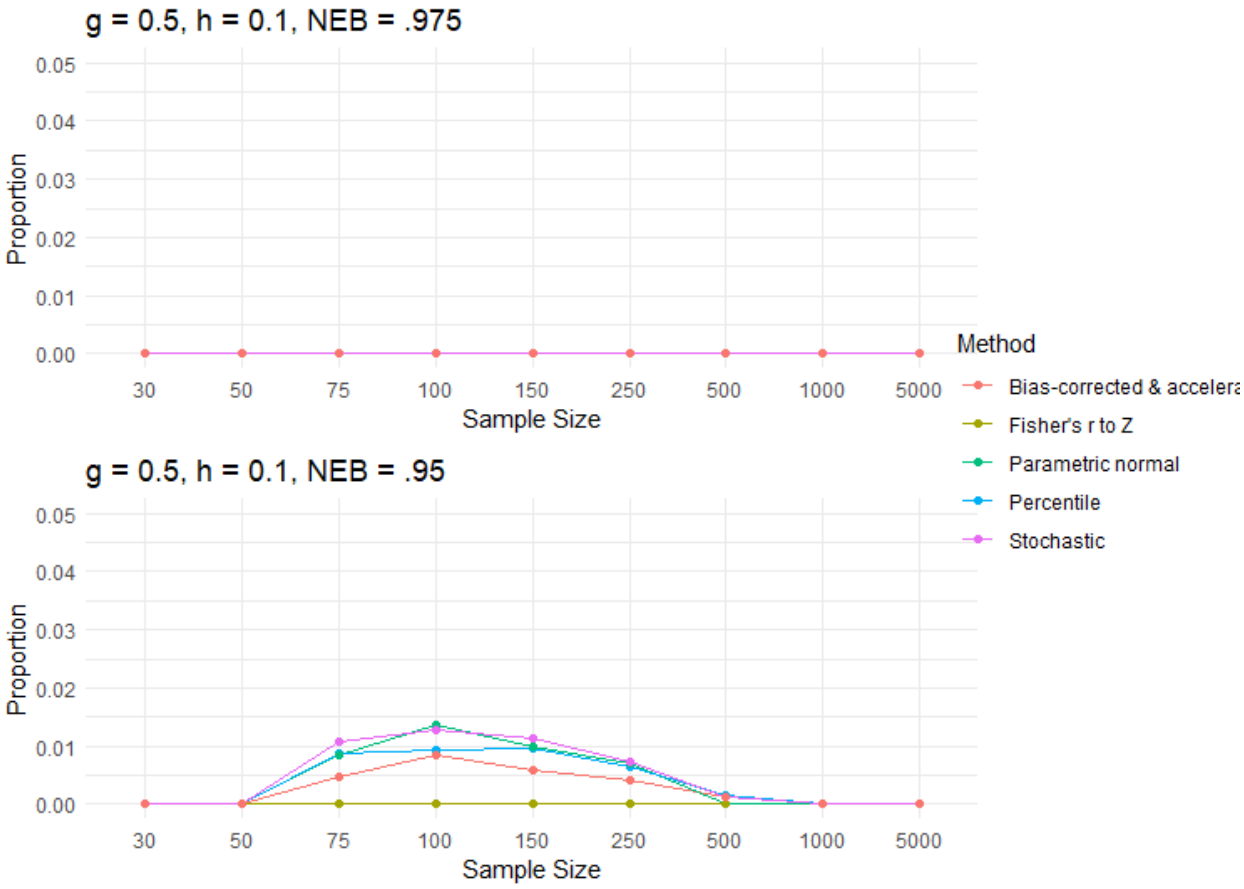


**Figure 11c**

*Proportion of Conclusions of Normality Across Confidence Interval Estimation Methods for Non-Negligibly Different from Normal Condition (*g *= 0.6,* h *= 0) with Conservative Negligible Effect Bound (NEB) of .975 and Liberal NEB of .95*

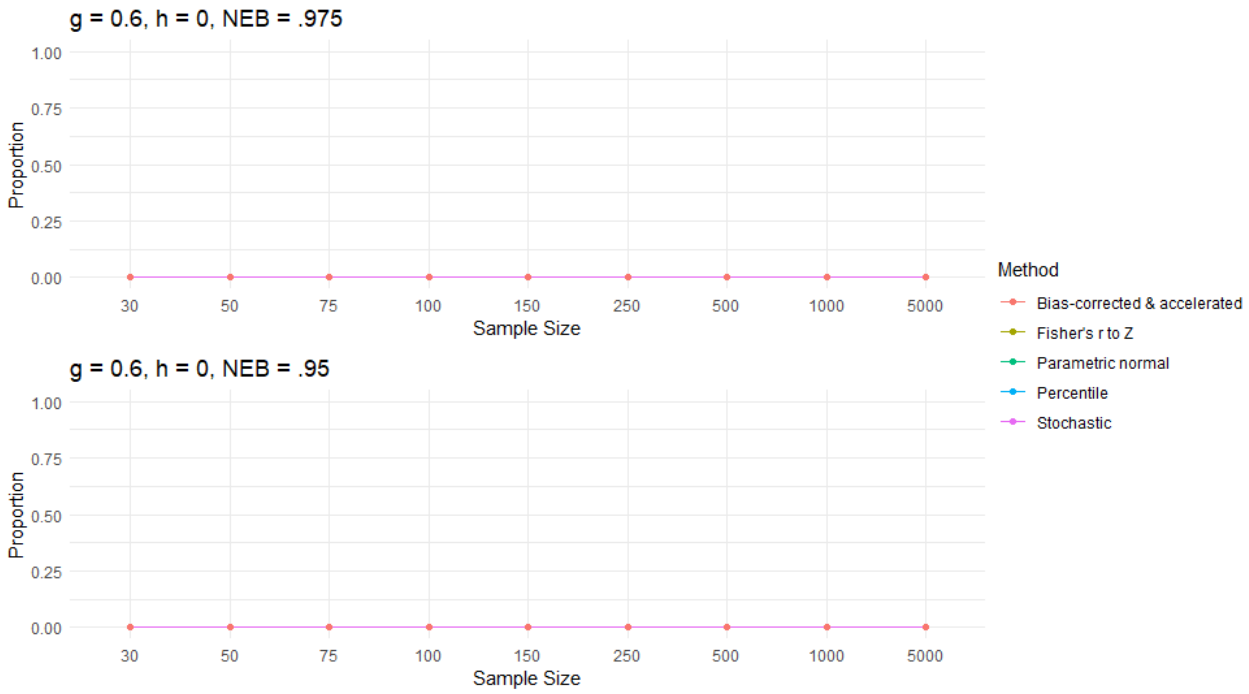


**Figure 11d**

*Proportion of Conclusions of Normality Across Confidence Interval Estimation Methods for a Normal Distribution (g = 0, h = 0) with Conservative Negligible Effect Bound (NEB) of .975 and Liberal NEB of .95*

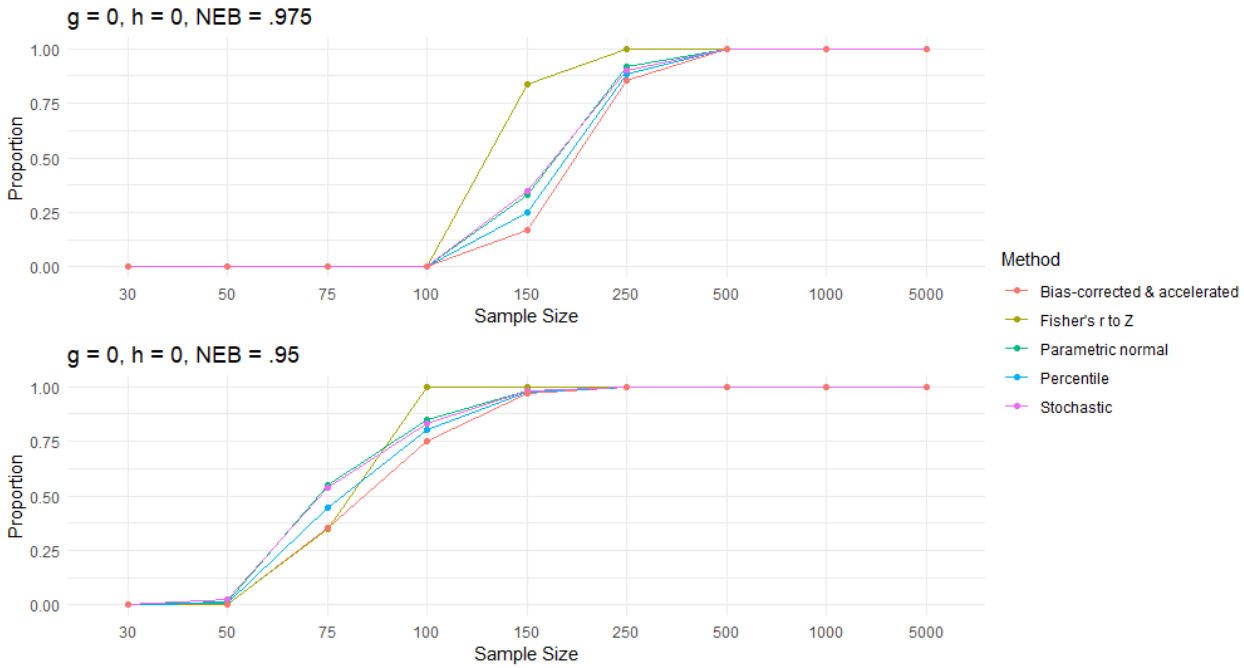


**Figure 11e**

*Proportion of Conclusions of Normality Across Confidence Interval Estimation Methods for a Normal Distribution (g = 0.2, h = 0) with Conservative Negligible Effect Bound (NEB) of .975 and Liberal NEB of .95*

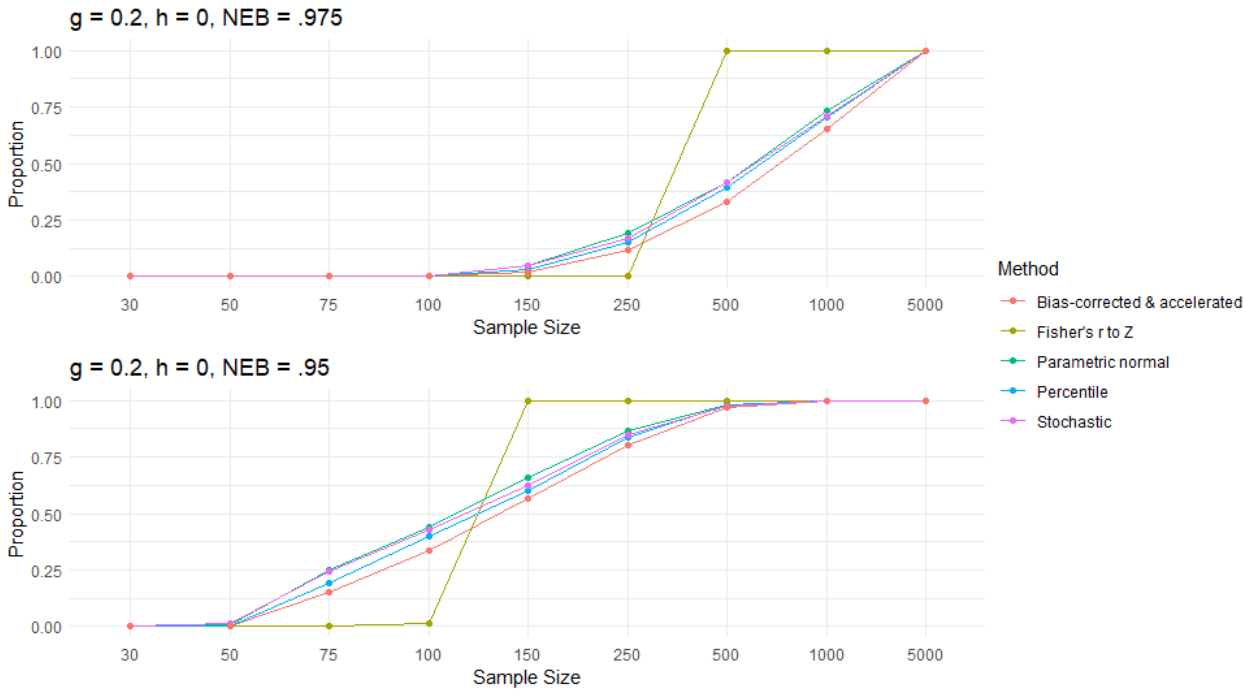


**Figure 11f**

*Proportion of Conclusions of Normality Across Confidence Interval Estimation Methods for a Normal Distribution (g = 0, h = -0.1) with Conservative Negligible Effect Bound (NEB) of .975 and Liberal NEB of .95*

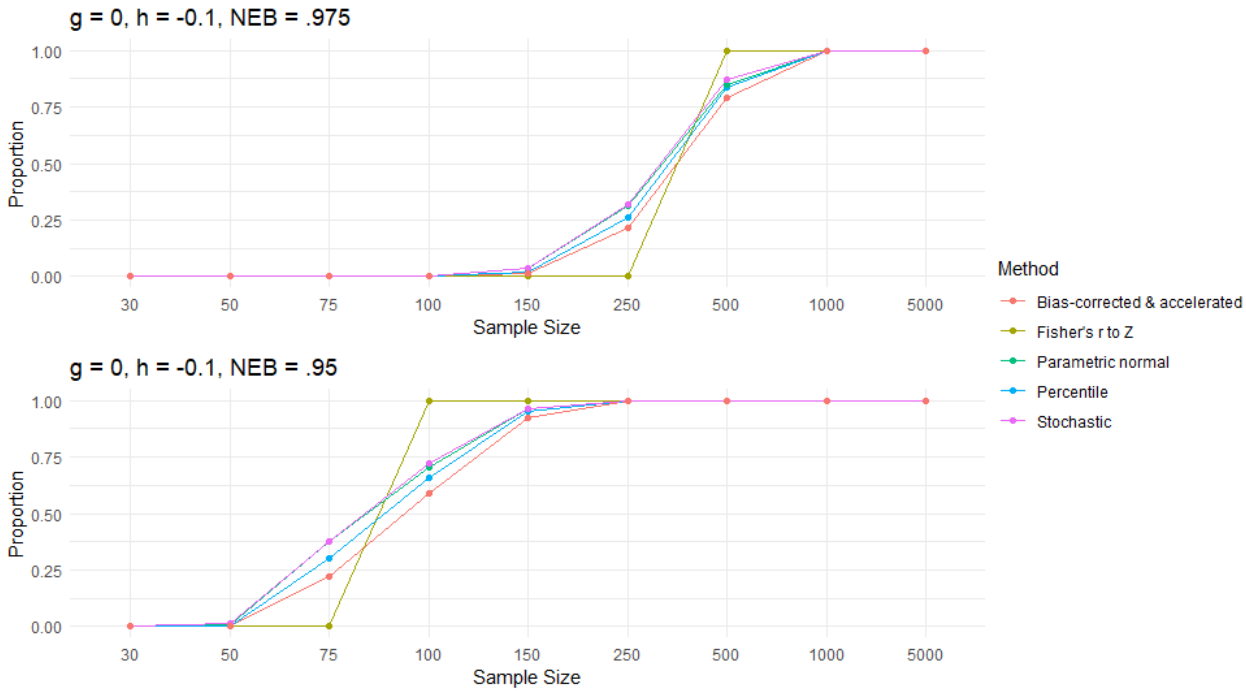
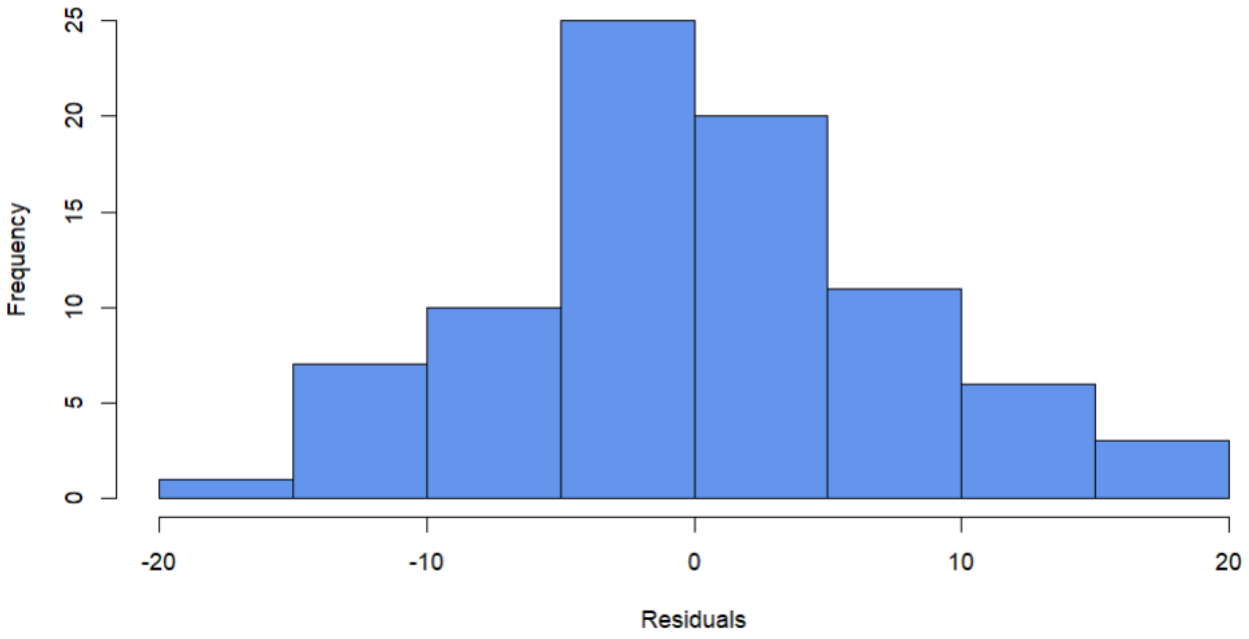


**Figure 12**

*Model Residuals from Multiple Regression of CESD Depression Predicted from Socially-Prescribed Perfection and Beck Anxiety Inventory Scores*



## Appendix A

**Table A1**

**Proportion of Conclusions of Normality using Fisher's *r* to *z* when NEB = .975, *g* = 0, *h* = 0.**

| Sample size | 140 | 141 | 142 | 143 | 144 | 145 | 146 | 147 | 148 | 149 | 150 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 0 | .003 | .006 | .026 | .069 | .182 | .329 | .517 | .699 | .8354 |

*Note*. 5000 simulations, 1000 bootstrap samples.

**Table A2**

**Proportion of Conclusions of Normality using Fisher's *r* to *z* when NEB = .95, *g* = 0, *h* = 0.**

| Sample size | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 |
|---|---|---|---|---|---|---|---|---|
| | 0 | .009 | .084 | .348 | .716 | .938 | .993 | 1 |

*Note*. 5000 simulations, 1000 bootstrap samples.

**Table A3**

**Proportion of Conclusions of Normality using Fisher's *r* to *z* when NEB = .975, *g* = 0.2, *h* = 0.**

| Sample size | 310 | 320 | 330 | 340 | 350 | 400 |
|---|---|---|---|---|---|---|
| | .007 | .052 | .213 | .513 | .806 | 1 |

*Note*. 5000 simulations, 1000 bootstrap samples.

**Table A4**

**Proportion of Conclusions of Normality using Fisher's *r* to *z* when NEB = .95, *g* = 0.2, *h* = 0.**

| Sample size | 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 | 109 | 110 | 120 |
|---|---|---|---|---|---|---|---|---|---|---|---|

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| .0386 | .0998 | .1844 | .302 | .4652 | .6314 | .7554 | .8654 | .9274 | .965 | 1 |

*Note*. 5000 simulations, 1000 bootstrap samples.

**Table A5**

**Proportion of Conclusions of Normality using Fisher's *r* to *z* when NEB = .975, *g* = 0, *h* = 0.1.**

| Sample size | 253 | 254 | 255 | 256 | 257 | 258 | 259 | 260 | 261 | 262 | 263 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | .001 | .003 | .005 | .008 | .017 | .027 | .033 | .051 | .076 | .117 | .158 |
| Sample size | 264 | 265 | 266 | 267 | 268 | 269 | 270 | 271 | 272 | 273 | 274 |
| | .196 | .260 | .317 | .395 | .469 | .542 | .614 | .679 | .746 | .799 | .847 |
| Sample size | 275 | 276 | 277 | 278 | 279 | 280 | 281 | 282 | 283 | 284 | 285 |
| | .888 | .923 | .938 | .957 | .970 | .982 | .991 | .992 | .997 | .997 | .998 |
| Sample size | 286 | 287 | 288 | | | | | | | | |
| | .999 | .999 | 1 | | | | | | | | |

*Note*. 5000 simulations, 1000 bootstrap samples.

**Table A6. Proportion of Conclusions of Normality using Fisher's *r* to *z* when NEB = .95, *g* = 0, *h* = -0.1.**

| Sample size | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 |
|---|---|---|---|---|---|---|---|---|---|---|
| | .0018 | .0194 | .0994 | .3092 | .5922 | .8364 | .959 | .9932 | .9992 | 1 |

*Note*. 5000 simulations, 1000 bootstrap samples.