A COMPARISON OF THE EFFECTS OF NON-NORMAL DISTRIBUTIONS ON TESTS OF EQUIVALENCE

by

LINDA F. ELLINGTON

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2011

MAJOR: EVALUATION AND RESEARCH

Approved by:

_____

Advisor Date

_____

_____

_____

_____

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

DEDICATION

To my beloved parents, Isaiah and Katie (Loveless) Ellington, whose love, faith and devotion are the foundation to my life. Thank you both for all that you have been, and continue to be, your daughter, Linda.

ACKNOWLEDGMENTS

It is with the deepest appreciation and respect that I would like to thank the following committee members for their support and guidance:

My major advisor, Dr. Shlomo Sawilowsky, who with deft patience and humor, kept me focused (pretty much) throughout each phase of the dissertation process. Dr. Gail Fahoome, an excellent teacher who walked me through the Monte Carlo simulation design process. I am indebted to Dr. Michael Addonizio and Dr. Allan Goodman, who agreed to join my committee. A final ‘thank you’ to the late Dr. Donald Marcotte; a kind and brilliant man who taught statistics as the beautiful language that it is.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Statistical theory and its application provide the foundation to modern systematic inquiry in the behavioral, physical and social sciences disciplines (Fisher, 1958; Wilcox, 1996). It provides the tools for scholars and researchers to operationalize constructs, describe populations, and measure and interpret the relations between populations and variables (Weinbach & Grinnell, 1997; Wilcox, 1996).Tests of equivalence, for example, are uniquely suited to research where the objective is to demonstrate that two populations are equivalent on a particular measure (Cribbie, Gruman & Arpin-Cribbie, 2004; Gruman, Cribbie & Arpin-Cribbie, 2007).

Equivalency testing provides behavioral and social sciences researchers the necessary tools to conduct analyses that evaluate the degree to which different conditions produce similar results. The most commonly used equivalency testing approaches, symmetrical confidence intervals and interval hypothesis testing, assume data normality (Berger & Hsu, 1996; Johnson & Duke, 2008). Such an assumption poses particular concern to behavioral and social sciences researchers; behavioral and social sciences data sets rarely follow normal distribution patterns (Keseleman, et al 1998; Micerri, 1989; Pearson & Please, 1975).

Study Significance

In developing an empirical study design, it is necessary that the underlying assumptions of the statistical methods employed in the study be adequately understood. Although usage of equivalency tests has increased amongst behavioral and social sciences researchers, several authors (Mutke and Holm-Mueller, 2004; Kristofersson and Navrud, 2005) have reported that the majority of documented studies on equivalency testing are conducted without acknowledgment of the normality assumption, or on the extent to which non-normality may exist in the data sets. As will be discussed in the literature review section of this study, the four most commonly used equivalency tests rely on the assumption of normality. Given that the majority of real data analysis in the behavioral and social sciences is comprised of non-normally distributed data, it is important that researchers be aware of the effects of non-normal data sets on the probability of detecting equivalence between populations.

Problem Statement

Determining equivalence between two populations requires the investigator acknowledge the underlying assumptions and limitations of the various statistical approaches, evaluate the appropriateness of their data sets, and select the approach that is most suitable for optimal results. To date, the number of published studies on the probability of detecting equivalency when data is non-normally distributed is limited (Jones, Jarvis, Lewis & Ebutt, 1996).

Aim of Study

The aim of this study is two-fold, it will (1) examine the effects and management of non-normally distributed data on equivalency tests under varied conditions for a two-sample design; and (2) compare the properties of showing equivalence between populations at the smallest effect sizes (.001 δ selected for this study).The present study has the following research objectives:

1. To assess the impact of data non-normality on three traditional equivalency tests commonly used by behavioral and social sciences researchers: Schuirmann's two one-sided *t*-test; Anderson and Hauck's nonequivalence null hypothesis; and Patel and Gupta's procedure.
2. To assess the impact of sample size under varying degrees of normality and non-normality.
3. To assess the impact of data non-normality on Type I error rate performance.
4. To provide recommendations based on the findings of the above.

Limitations to the Study

The following limitations to the study are presented below:

1. Is limited to addressing the underlying assumption of normality, and excludes underlying assumptions of heteroscedasticity.
2. Is limited to detecting non-equivalence at the smallest effect size, and excludes a comparative power analysis. A statistical power analysis may be either retrospective (post hoc) or prospective (a priori). A prospective analysis may be used to determine a required sample size to achieve

target statistical power, and a retrospective power analysis computes the statistical power of a test given sample size and effect size (Park, 2008). The data sets will be computer simulated from one theoretical distribution (Gaussian) and two real world data sets: Smooth Symmetric and Extreme Asymmetry as described by Micerri (1989).

3. Is limited to three traditional equivalency tests most commonly used by behavioral and social sciences researchers: Schuirmann's two one-sided *t*-test; Anderson and Hauck's nonequivalence null hypothesis; and Patel and Gupta's procedure.

Human Subjects

Human subjects will not be employed in this study. The Behavioral Protocol Summary Form was submitted to the Wayne State University Behavioral Investigation Committee; exemption was granted on September 14, 2009.

Identification of Variables

For the stated purpose of this study, the following variables are defined as follows in Table 1.

Table I. Monte Carlo Study Variables

Variable	Variable Function
% of rejection rates	Dependent
Alpha level (α)	Independent
Length of Equivalence Interval	Independent
Sample size (n)	Independent
Sampling Distribution	Independent

Definition of Statistical Terms

For the stated purpose of this study, the following terms are defined as follows:

Alpha level. The pre-specified level of significance used in selecting the critical value, and refers to the probability of making a Type I error if H_o is rejected (Hinkle, Weirsman & Jurs, 1998).

Confidence interval. An interval between two numbers with an associated probability *p* which is generated from a random sample of an underlying population, such that if the sampling was repeated numerous and the interval recalculated from each sample according to the same method, a proportion of *p* of the intervals would contain the population parameter in question (Cohen & Cohen, 1983).

Critical mean difference: Any difference smaller than would be considered meaningless within the framework of the study (Cribbie, Gruman & Arpin-Cribbie, 2004).

Distribution: The arrangement of values/outcomes that demonstrates observed frequency (Hinkle, Weirsman & Jurs, 1998).

Distribution-free tests: Tests hypotheses without relying on underlying assumptions about population parameters (Sawilowsky & Fahoome, 2003).

Equivalency interval: Primarily dependent on subjective “level of confidence” with which to declare two (or more) populations equivalent (Cribbie, Gruman & Arpin-Cribbie, 2004).

Equivalence tests: Statistical methods for determining if populations are equivalent on a specific dependent variable (Schuirmann, 1987).

Kurtosis: A measure of the degree to which a distribution is peaked (Wilcox, 1996).

Monte Carlo simulation: Computer simulations that involve statistical sampling and allow for the measuring of mathematical properties of statistical tests (Harwell, 1990).

Normal distribution: A bell-shaped curve with a skewness value of 0 and a kurtosis value of 0 (Wilcox, 1996).

Power: The probability of rejecting the null hypothesis when it is false also known as Type II error (Hinkle, Weirsman & Jurs, 1998). The statistical power is the ability of a test to detect an effect, if the effect actually exists (*Ibid*).

Robustness: (1) Pertains to statistical test and the extent that violating its assumptions does not affect the probability of its Type I error (Hunter & May, 1993). (2) Pertains to Type II error and the compliment of the power of the statistical test (Sawilowsky, 1990).

Sample size: The number of scores in the subset of the population (Wilcox, 1996).

Skew: The lack of symmetry of a distribution of scores, elongation of either the left or right tails (Wilcox, 1996)

CHAPTER 2

LITERATURE REVIEW

Overview of Equivalency Tests

Tests of equivalence are the appropriate techniques where the research objective is to demonstrate that two populations are equivalent on a particular measure. Equivalency tests are designed to assess if population or treatment differences are acceptably small to rule in favor of equivalence (Anderson & Hauck, 1983; Schuirmann, 1987; Selwyn & Hall, 1984; Westlake, 1976). Barker, Luman, McCauley & Chu (2002) have described the procedure as follows:

> In equivalence testing, the null hypothesis is "a difference of Δ or more." Thus, α is the probability of concluding that the populations differ by less than Δ when, in fact, the difference is Δ or more. Similarly, β is the probability that the populations' coverage will be found to differ by at least Δ when the true difference is less than Δ" (p. 1058).

Testing for equivalence requires that the investigator (1) specify an equivalency interval and (2) determine if the difference between the population means or medians is within the specified equivalence interval. The null hypothesis (H_0) is stated such that the statistical test is evidence of non-equivalence: the populations or groups differ by more than a tolerably small amount, designated asΔ. The alternative hypothesis (H_1) are the populations or groups differing by less thanΔ, or that they are similar or equivalent on the dependent variable or measure (Anderson & Hauck, 1983; Rouanet, 1996; Schuirmann, 1987; Selwyn & Hall, 1984; Westlake, 1976):

$$H_{01}: \mu_T - \mu_C \leq \Delta_L \text{ versus } H_{11}: \mu_T - \mu_C > \Delta_L, \quad \text{and}$$

$$H_{02}: \mu_T - \mu_C \geq \Delta_L \text{ versus } H_{12}: \mu_T - \mu_C < \Delta_L.$$

Equivalency Tests and Applicable Areas of Research

Long-standing convention amongst researchers in the behavioral and social sciences has been to study and determine statistically significant differences between populations or programs. Due to lack of familiarity with equivalency tests, many researchers continue to use statistically significant difference tests when in fact the study aim is to determine similarity or equivalence between populations or treatments (Cribbie, Gruman & Arpin-Cribbie, 2004; Gruman, Cribbie & Arpin-Cribbie, 2007). Nevertheless, equivalency tests have been demonstrated to be of significant theoretical and practical relevance in empirical research scenarios located in varied research and cost analysis scenarios (Rogers, Howard & Vessey, 1993; Cribbie, Gruman & Arpin-Cribbie, 2004; Kristofersson and Navrud, 2005).

Behavioral/Psychometrics Research Scenarios: Where the research objective might be to demonstrate parallelism between two forms, the correlation of test form A with a criterion which may be construed as a measure of validity, may be compared to the correlation of test form B to the same criterion. It is known that the shorter of the two test forms, B, is less prohibitive in time and cost to administer than test form A. Equivalency testing would be used to compare the correlation coefficients for the longer test form A and the alternative short parallel test form B based on Lord and Norvick (1968) mental test theories. Behavioral studies conducted by Rogers, et al drew on examples of baseline equivalence assessment and the assessment of equivalency in efficacy between cognitive and behavioral interventions (Rogers, Howard & Vessey, 1993).

Cost Benefit Analyses: In organizational management, the efficacy of different therapeutic programs designed to treat alcoholic employees may be compared. If for example, the efficacy is the same for cognitive- and behavioral-based intervention programs; management may recommend and implement the intervention that is least cost and time-prohibitive. Among education economists and policy makers, the program objective might be to demonstrate that early intervention programs are of equal benefit to rural and metropolitan children. Two potentially under-utilized settings for equivalency testing are Forensic Science (e.g., ballistics matching) and Criminal Justice. For the 38 states that allow capital punishment for specific offenses, the related human cost of Type I error is clearly catastrophic (M. Addonizio, personal communication, August 18, 2009).

Health Care Research Scenarios: Demonstrating equivalency is of great importance in medical and allied health care research, most notably pharmaceutical research and manufacturing. With the explosive growth in the manufacture and selling of generic drugs, federal regulatory agencies demand that pharmaceutical manufacturers demonstrate in clinical trial studies, the equivalency between proposed generic drugs and the more expensive, but established, referent drug (Brown, Hwang & Munk, 1997). Within the current U.S. health care system model, there exists a vast array of surgical procedures that effect physician decision making. Amongst surgeons and health insurance companies, the question posed is not “which procedure is superior?”, but “are they bioequivalent with regards to patient outcomes?” If in clinical trials bioequivalence has been established, determinations will be made based on

tradeoffs in costs, access and utilization, and risk incurred by patients (Breslow and Day, 1980).

Equivalency Test Models

Gruman, et al (2007) has stated: "…researchers require a statistical technique designed specifically to test the degree to which different conditions produce similar results. Tests of equivalence serve this purpose" (p 134). Amongst behavioral and social sciences researchers, the four most commonly used tests of equivalence are Westlake's symmetric confidence interval (Westlake, 1976); Schuirmann two one-sided *t*-test (Schuirmann, 1987; 1979); Anderson and Hauck's nonequivalence null hypothesis (1983); and Patel and Gupta (1984). The Westlake model is an example of the confidence interval approach to equivalency testing; the latter three are examples of the interval hypothesis testing approach.

Westlake Symmetric Confidence.

The confidence interval approach uses experimental data to formulate confidence intervals mean or median differences (or ratios). The researcher constructs a confidence interval that is compared to the limits of the equivalence interval, Δ_1 and Δ_2, which is selected *a priori,* or predetermined, by the researcher or regulatory agency (Seaman & Serlin, 1998; Stegner, Bostrom & Greenfield, 1996). Should the entire constructed interval fall within the upper and lower limits, the two populations or groups are considered equivalent, if not, equivalence between populations or groups is rejected (Seaman & Serlin, 1998).

Westlake (1972) and Metzler (1974) proposed the use of confidence intervals as a statistical method to test for equivalence in place of the inappropriate use of the null hypothesis tests of statistically significant differences. “Testing the null hypothesis of no difference is inappropriate for studies in which the primary objective is to demonstrate that two populations are equivalent, rather than different on a dependent measure” (Gruman, Cribbie & Arpin-Cribbie, 2007, p 133). Later, after observing, “most clinicians tend to make their equivalence statements in a symmetrical manner” (p 741, 1976), Westlake proposed a confidence interval adjusted to be symmetric about zero for the mean difference or one for ratios and proportions:

> The conventional method of setting confidence intervals for the difference of the means of two normal populations gives an interval which is not, in general, symmetrical about zero. A modification of the conventional method leads to symmetry about zero is discussed and is recommended as particularly appropriate for use in bioequivalence trials. This modification has the effect of decreasing the "effective" length of the confidence interval, on which the decision concerning bioequivalence is based, while increasing the confidence coefficient (abstract, 1976).

The confidence interval of mean difference formulated by Westlake (1972) and Metzler (1976) is given as follows:

$$\mu_c - \Delta < \mu_T < \mu_c + \Delta . \quad (1)$$

where μ_c denotes the population mean of the control or referent group

where μ_T denotes the population mean of the experimental treatment group.

Westlake and Metzler’s method may be understood as the construction of a confidence interval for μ_T that is symmetric about μ_c. Re-arrangement of the *T*-statistic yields:

$$\mu_C + k_U S_{AT-AC} - (\overline{A}_T - \overline{A}_C) < \mu_T < \mu_C + k_L S_{AT-AC} - (\overline{A}_T - \overline{A}_C), \quad (2)$$

where $\overline{A}_T$ denotes the sample mean for the experimental treatment group

where $\overline{A}_C$ denotes the sample mean of the control or referent group

where $k_L = t_\alpha$ denotes the lower or 5th percentile

where $k_U = t_{1-\alpha}$ denotes the upper or 95th percentile, and

where k_L and k_U are selected so that $\int_{k_L}^{k_U} Tdf = 1 - \alpha$.

Thus, the probability of T between k_L and k_U based on a central t-distribution with $n_T + n_C - 2df$ or 2 degrees of freedom, is equal to $1 - \alpha$, where n_T denotes size of experimental group and n_C denotes size of control or referent group. To assure symmetry of the confidence interval, the following statements must prove valid:

$$\Delta = k_L S_{AT-AC} - (\overline{A}_T - \overline{A}_C) = k_U S_{AT-AC} + (\overline{A}_T - \overline{A}_C) \quad (3)$$

This result indicates that:

$$(k_L + k_U) S_{AT-AC} = 2(\overline{A}_T - \overline{A}_C). \quad (4)$$

Westlake (1989) estimated the probability of establishing equivalence would increase with the use of the symmetric confidence interval at approximately zero for mean differences (approximately unity for the ratio of means).

Critique of the Westlake symmetric confidence interval.

Chow & Liu (1992) and Frick (1987) have noted that Westlake's symmetric confidence interval has at minimum $1 - \alpha$ coverage probability and is conservative in the sense that the real Type I error rate might be at most the nominal α level. Mantel (1977) and Mandallaz & Mau (1981) identified two significant limitations to the model. Firstly, the upper and lower limits are

artificially constructed such that the direction of the difference is not obvious. A conventional confidence limit of $93 < \mu_1 - \mu_2 < 128$ reflects that the mean for group 1 is higher than the mean for group 2. The Westlake symmetric confidence interval does not provide information on location.

Secondly, the tail probabilities are not symmetric: as the difference between means increases, the confidence interval shifts from two-sided to one-sided. This becomes a significant disadvantage as probabilities result in a confidence interval with the shortest length (Kendall & Stuart, 1961). Metzler (1988) advised that the symmetrical confidence interval be retained as a statistical method for decision-making, but not for estimation or testing. Serlin (1993) found difficulty with symmetric confidence interval method because it is not related to the research hypothesis of equivalence.

Due to the above stated reasons, the Westlake symmetric confidence interval will not be included in this study. However, other confidence interval methods have been suggested for testing of equivalence. Lock (1984) proposed a procedure for constructing a confidence interval for the ratio of means based on the Fieller theorem (Fieller, 1954). Chow & Shao (1990) put forward a joint confidence region for assessing equivalence.

The Schuirmann Two One-Sided *t* -Test.

The determination of equivalence between populations or groups is based on the inspection of differences in the parameter of interest between two populations, such as the mean or median (Schuirmann, 1981; 1987; Anderson & Hauck, 1883). However, it is noted that no two groups or treatments have

precisely the same mean or median. Based on this supposition, two groups or treatments that differ by a clinically unimportant difference in either direction may still be accepted as equivalent. These clinically unimportant differences must be determined *a priori* by the researcher, and, based on them, interval null hypotheses are formulated (Schuirmann, 1987; Welleck, 2003). Based on the interval hypothesis approach, several hypothesis tests for equivalence were formulated. Lehmann (1986) described the common approach to testing range null hypotheses. The domain of well known and frequently used interval hypothesis testing methods for equivalence include the Schuirmann two one-sided *t*-test (1981; 1987); the Anderson and Hauck nonequivalence null hypothesis (1983); and the Patel and Gupta procedure (1984).

Schuirmann (1981; 1987) first introduced the use of an interval hypothesis for assessing equivalence, and is the most widely used by behavioral and social science researchers when the research objective is to determine equivalency between populations (Hsu et al, 1994; Berger & Hsu, 1996; Gruman, Cribbie & Arpin-Cribbie, 2007).The popularity of the test may be attributed to its bounded Type I error rate, good power (≥0.80), and a well-behaved rejection region (Hsu et al, 1994). Rogers, Howard & Vessey (1993) are credited with introducing the Schuirmann two one-sided *t*-test with examples of its application to behavioral and social science research literature. Gruman, Cribbie & Arpin-Cribbie (2007) summarized the Schuirmann model as follows:

> The first step in conducting Schuirman's test of equivalence is to establish a critical mean difference for declaring two population means equivalent (*D*). Any mean difference smaller than D would be considered meaningless within the framework of the experiment. The selection of an equivalency interval (*D*) is an important aspect of equivalence testing that is primarily dependent on a subjective "level of confidence" with

which to declare two (or more) populations equivalent. This level of confidence can take on many different forms including a raw value (e.g., mean test scores different by 10 points), a percentage difference (e.g. +/- 10%), a percentage of the pooled standard deviation difference, etc (p.134, 2007).

The interval hypothesis test for equivalence is formulated as follows

$$H_0 : \mu_T - \mu_C \leq \Delta_L \text{ or } \mu_T - \mu_C \geq \Delta_U \quad (5) \text{ versus}$$

$$H_1 : \Delta_L < \mu_T - \mu_C < \Delta_U . \quad (6)$$

It is assumed that the samples meet the underlying assumptions of being randomly and independently selected from normally distributed populations with equal variance (Berger & Hsu, 1997). Two one-sided hypothesis tests can be used to establish equivalence as the null hypothesis relates to the nonequivalence of the population means and can be expressed as two sets of one-sided hypotheses (Rogers, Howard & Vessey, 1993; Seaman & Serlin, 1998) $H_{01} : \mu_T - \mu_C \leq \Delta_L$ versus $H_{11} : \mu_T - \mu_C > \Delta_L$, (7) and

$$H_{02} : \mu_T - \mu_C \geq \Delta_L \text{ versus } H_{12} : \mu_T - \mu_C < \Delta_L. \quad (8)$$

The first set of hypotheses is intended to verify that the difference between the population means is not too small, while the second set of hypotheses is intended to confirm that the difference between population means is not too large. The two sets of one-sided hypotheses are tested by the following set of statistics:

$$T_L = \frac{\overline{A}_T - \overline{A}_C - \Delta_L}{S_{AT-AC}} \quad (9) \quad \text{and}$$

$$T_U = \frac{\overline{A}_T - \overline{A}_C - \Delta_U}{S_{AT-AC}} \quad (10)$$

for the second set of hypotheses. Under the normality assumption, T_L and T_U

follow a *t*-distribution with $n_T + n_c - 2df$ (2 degrees of freedom), equivalence is established only if both H_{01} and H_{02} are simultaneously rejected:

$$T_L - t(n_T + n_c - 2, \alpha) \text{ and } T_U - t(n_T + n_c - 2, \alpha)$$

where t $t(n_T + n_c - 2, \alpha)$ is the 100(1- α) percentile of the *t*-distribution with $n_T + n_c - 2df$.

To establish equivalency, it is noted that only the test that yields a larger ρ-value is required and sufficient for decision-making. If the test with a larger ρ-value results in a rejection of the null hypothesis for a given α, it follows then, that the test with a smaller ρ-value must yield a rejection as well (Wang, DasGupta & Hwang, 1996). The conclusion of equivalency is established on the simultaneous rejection of both tests; if the ρ-values of the two tests are the same, both tests lead to the same conclusion (Schuirmann, 1987). Based on the above rationale, only the test with the larger ρ-value is necessary for the assessment of equivalency. Furthermore, because the result from the test with the smaller ρ-value is pre-empted by the test of the larger ρ-value, the Type I error rate is equal to that assigned to the test with the larger ρ-value (Schuirmann, 1987; Berger & Hsu, 1997).

A Type I error can only be committed when this hypothesis is rejected. This is because H_{01} and H_{02} is mutually exclusive and only one of them can be true. Schuirmann (1987) demonstrated that the same conclusion would be reached using his two one-sided *t*-test method at Type I error rate of α and the conventional 100(1-2α) % confidence interval. From this perspective, the two

one-sided *t*-test method and the traditional confidence interval approach are noted to be operationally equivalent (Schuirmann, 1987; Berger & Hsu, 1997).

Critique of the Schuirmann Two One-Sided *t*-Test.

Chow and Liu (1992) found in a small simulation study that the 1-2α confidence interval does not guarantee that, over time, the chance of the 1-2α confidence interval being within the acceptance limits is at least 1-2α. Only 91.5%, 43.9% and 7.5% of confidence intervals were within the equivalence limits for intra-subject variability of 20%, 30% and 40%, respectively. In addition, the two one-sided *t*-test method was found to be conservative in terms of Type I error rate (Chow & Liu, 1992). Cribbie, Gruman & Arpin-Cribbie (2004) found the Schuirmann two one-sided *t*-test "To be more effective than Student's *t*-test at detecting population mean equivalence with large samples sizes (n=25); however, Schuirmann's test of equivalence performs poorly relative to Student's *t*-test with small sample sizes and /or inflated variances "(p.1, 2004).

Anderson and Hauck's Nonequivalence Null Hypothesis Procedure.

Instead of using T_L and T_U defined previously to assess H_{01} and H_{02}, respectively, Anderson and Hauck (1983) proposed a technically simple procedure for evaluating H_0: $\mu_T - \mu_C \leq \Delta_L$ *or* $\mu_T - \mu_C \geq \Delta_U$ directly. For the Anderson and Hauck procedure, the test statistic is given as

$$T_{AH} = \frac{\overline{A}_T - \overline{A}_C - (\Delta_L + \Delta_U)/2}{S_{AT-AC}} \quad (11)$$

Under the assumption of normality of the population distributions, the test statistic T_{AH} follows a non-central *t*-distribution with non-centrality parameter

$$\delta = \frac{\overline{A}_T - \overline{A}_C - (\Delta_L + \Delta_U)}{\delta\sqrt{1/n_T + 1/n_C}} \qquad (12)$$

The test will reject H_0 in favor of equivalence if T_{AH} falls between two critical values C_L and C_U, which satisfy

$$P(C_L < T_{AH} < C_U \mid \mu_T - \mu_C = \Delta_L, \delta^2) =$$
$$P(C_L < T_{AH} < C_U \mid \mu_T - \mu_C = \Delta_U, \delta^2) = \alpha . \qquad (13)$$

However, Anderson and Hauck (1983) demonstrated that only a single critical value $C = C_U = C_T$ is required. The critical value C may be obtained by solving the following:

$$P(\mid T_{AH} \mid < C \mid \mu_T - \mu_C = \Delta_L, \delta^2) =$$
$$P(\mid T_{AH} \mid < C \mid \mu_T - \mu_C = \Delta_L, \delta^2) = \alpha \qquad (14)$$

The decision about equivalence can also be based on the ρ-value. With the observed data, the empirical ρ-value can be calculated under the null hypothesis. If the non-centrality parameter is known, the ρ-value is given by

$$\rho = P(\mid T_{AH} \mid < t_{AH} \mid \mu_T - \mu_C = \Delta_U, \delta^2), \qquad (15)$$

where t_{AH} is the observed value of T_{AH}. If $\rho \le \alpha$, the null hypothesis is rejected and equivalence is then concluded. On the other hand, if the non-centrality parameter is unknown, approximation to the $\rho-$value is used. Anderson and Hauck (1983) considered three approximations based on the non-central *t-,* central *t-* and normal distributions. Among the three approximations, the central *t*-distribution approximation was found to be the best in terms of power (Anderson & Hauck, 1983).

Since the sample standard deviation, s_ρ, is a consistent estimator of the population standard deviation δ, the non-centrality parameter at the limit of the equivalence interval, for example, $\mu_T - \mu_C = \Delta_U$, can be estimated by

$$\delta = \frac{\Delta_L + \Delta_U}{2S_{AT-AC}}, \qquad (16)$$

and therefore, the statistic

$$T_{AH} - \delta = \frac{\overline{A}_T - \overline{A}_C - (\Delta_L + \Delta_U)/2}{S_{AT-AC}} \qquad (17)$$

$$- \frac{\Delta_L + \Delta_U}{2S_{AT-AC}}$$

$$= \frac{\overline{A}_T - \overline{A}_C - \Delta_U}{S_{AT-AC}}$$

approximately follows a central *t*-distribution with $n_T + n_c - 2df$.

Assuming $T_{AH} > 0_H$, it is noted that at the upper limit of the equivalence interval, Δ_U, $T_{AH} - \delta$ is equal to T_U of the Schuirmann two one-sided *t*-test. Similarly, at the lower limit, Δ_L, - $T_{AH} - \delta$ is equal to T_U of the Schuirmann two one-sided *t*-test. However, Chow & Liu (1992, p 92) observed that Anderson and Hauck's test is always more powerful than the Schuirmann two one-sided *t*-test. In addition, Anderson and Hauck (1983) demonstrated in a simulation study that the power of their method always exceeds the power of both the Schuirmann and Westlake methods.

Patel - Gupta's Procedure.

The Patel - Gupta procedure is similar to the Anderson & Hauck test: a single test is used to evaluate the null hypothesis involving a pre-specified difference (Δ). However, unlike Anderson & Hauck (1983), which employed the central t- distribution as approximation, Patel-Gupta (1984) utilized non-central *F*-distributions to test

$$H_0 : | \mu_1 - \mu_2 | \geq \Delta$$

against the alternative given by

$$H_0 : | \mu_1 - \mu_2 | < \Delta$$

where Δ is some pre-determined clinically important difference. Patel-Gupta's test statistic is given by

$$F\gamma = \frac{n_1(\overline{A}_1 - \overline{A})^2 + n_2(\overline{A}_2 + \overline{A})}{S^2} \qquad (18)$$

Under the null hypothesis it is distributed as a non-central F-distribution with 1-degree of freedom and $n_T + n_c - 2$, and approximate non-centrality parameter

$$\gamma = \frac{(n_1 n_2)}{n} \frac{\Delta^2}{\delta^2} \qquad (19)$$

with the usual notation for the mean, sample size and estimated standard deviation. The null hypothesis is rejected if $F_{1,n-2,\gamma} \leq c$ where

$$P(F_{1,n-2,\gamma} \leq c \mid H_0) = \alpha \qquad (20)$$

One significant drawback to both the methods are that both are slightly liberal in the sense that the real Type I error rate might exceed the nominal level (Chou & Liu, 1992; Frick, 1990; Shuirmann, 1987).

There is no one optimal test to be found for the purpose of establishing equivalency: it has been determined that there are tradeoffs in Type I error rate, statistical power, and shape of the rejection region (Chow & Liu, 1992; Berger & Hsu, 1996; Perlman & Wu, 1999). The α level for the Westlake and Schuirmann procedures can both be slightly conservative in the sense that the real Type I error rate might be at most the nominal α level. The Anderson and Hauck and Patel and Gupta procedures can both be slightly liberal in the sense that the real Type I error rate might exceed the nominal α level. Berger & Hsu (1996, p 289) commented on the continued popularity of the Schuirmann procedure: “Although not the most powerful version of equivalence testing available, the ‘simplicity and intuitive appeal’ of the two one-sided *t*-test has led to its widespread use and acceptance”.

Equivalency Tests and Violations to Normality

When considering statistical methods to detect the degree to which equivalence may exist, it is important that researchers understand the statistical properties of these tests under conditions that may or may not meet underlying assumptions. All four equivalency tests are predicated on underlying assumptions that samples are randomly and independently selected from normally distributed populations with equal variances (Gruman, Cribbie & Arpin-Cribbie, 2007). However, various studies in pharmaceutical (Metzler & Hung, 1983; Zhou, He & Yuan, 2004), behavioral and social sciences research (Keselman, et al, 1998; Micerri, 1989; Pearson & Please, 1975) have demonstrated that the underlying assumption of normal distribution of real data

is frequently violated.

In an investigation of 440 distributions taken from education and psychology studies, Micceri found that none of the data sets followed normal distribution patterns and just 3% were identified as relatively symmetric with light tails (Micceri, 1989). Studies conducted by Bridge & Sawilowsky (1999) and Barber & Thompson (1998) found distributions in medicine often display extreme skewness. It has also been determined that non-normally distributed data sets greatly affect hypothesis tests incorporating the *t* and *F*-statistics (Bradley, 1968; Fahoome & Sawilowsky, 2000; Kerlinger & Lee, 2000; Zimmerman, 1998). The Schuirmann two one-sided *t*-test assumes samples are drawn from a normal distribution. However if normality has been violated, then "tests such as the two one-sided *t*-test which is based on the Student *t*-distribution is inappropriate" (Berger & Hsu, 1996; p 287). Therefore, if the presence of non-normal distributions is a rule rather than an exception, researchers must take a close look at the shape of their data and the tests they are applying.

Characteristics of Non-Normally Distributed Data

Researchers and statisticians have been concerned with non-normally distributed data reaching back to the early nineteenth century (Pearson & Please, 1975; Stigler, 1973). Non-normally distributed data is common in practice, and has been documented in many applied studies (Keselman, et al 1998; Micceri, 1996; Pearson & Please, 1975). Non-normally distributed data is of concern to researchers as it has an effect on statistical procedures such as summary statistics and hypothesis tests. Micceri (1989) identified several factors

that might contribute to violation of normal distribution of data:

Other factors that might contribute to a non-Gaussian error distribution in the population of interest include but are not limited to (a) the existence of undefined subpopulations within a target population having different abilities or attitudes, (b) ceiling or floor effects, (c) variability in the difficulty of items within a measure, and (c) treatment effects that change not only the location parameter and variability but also the shape of a distribution (p 157).

Detecting Departure from Normality

Sawilowsky & Fahoome (2003) identified the characteristics of a normal distribution: a mean, μ, of 0.00, a standard deviation, σ, of 1.00, skewness of 0.00 and kurtosis of 3.00 (Sawilowsky & Fahoome, 2003). In characterizing and summarizing data (i.e., measures of central tendency, dispersion), the main concern has been robustness of the statistical procedure to normality. The term robust, when used to describe a statistical procedure, refers to the insensitivity of parametric statistics to violations of their assumptions. Non-normality of data may occur due to a variety of reasons including growth or decay in which the underlying distribution is exponential, multimodal lumpy (Micceri, 1989), mass at zero with gap (Sawilowksy & Hillman, 1992) or some non-Gaussian shape.

Examples of non-normal distributions include contaminated distributions that are Gaussian in shape, but contaminated by the presence of outliers (Wilcox, 1997). Outliers are values occurring in the dataset and are significantly larger or smaller than other values; thus creating bias toward measures of central tendency (location) and dispersion: the sample mean and variance (*Ibid*,

1997). Wilcox (1997) made several observations on the characteristics of contaminated distributions: (1) measures of central tendency do not fall in the same location of the tail, and (2) exhibited variance that is larger than the normal distribution.

The mean of a contaminated distribution will demonstrate bias in the direction of the skewed tail of the distribution and is not a robust estimator of location (Sawilowsky & Fahoome, 2003). The median, which is much less sensitive to the presence of outliers in the distribution, is a more robust estimate of the center of the distribution (*Ibid*, 2003). Lastly, Wilcox was noted that "Outliers and heavy-tailed distributions are serious practical problems because they inflate the standard error of the sample mean…Modern robust methods provide an effective way of dealing with this problem" (p 2, 1997).

CHAPTER 3

METHODS

The aim of this study is examine the effects and management of non-normally distributed data on equivalency tests under varied conditions for a two-sample design; and to compare the properties of showing equivalence between populations at the smallest effect size, $.001\delta$, selected for this study. A Monte Carlo simulation study is designed to address the following research questions:

1. Which, if any, of the tests examined in this study control Type I error?
2. If the Type I error rate is not controlled, under what conditions are tests liberal or conservative?
3. Is there an overall best test to recommend for the management of non-normally distributed data?
4. Are there specific circumstances that dictate which of the three models is most appropriate under conditions in which normality is violated?

This study is limited to three of the four traditional equivalency tests discussed in the literature review: Schuirmann's two one-sided *t*-test; Anderson and Hauck's nonequivalence null hypothesis; and Patel and Gupta's procedure. As discussed, the Westlake symmetric confidence interval approach is found to be rather conservative in terms of Type I error rate: it will not be included in the study.

Monte Carlo Design

Harwell (1990) defined the Monte Carlo simulation study as a series of computer simulations capable of measuring the mathematical properties of a

given statistical test achieved by allowing for the simulation and control of all variables under investigation:

> In the typical MC study of a given statistical test the following process is repeated for a large number of samples: data are simulated which reflect a specific relationship among variables... The values of the statistical test provide information on its properties (e.g., the proportion of the "significant" values on the test). If the underlying assumption of the test were satisfied, exact statistical theory would guarantee that the test would have a specified type I error rate and would permit the probability of rejecting a false statistical hypothesis to be computed. Monte Carlo studies permit these characteristics to be examined when underlying assumptions are violated (p.4).

This study will employ Monte Carlo simulation techniques using Dell DIM 4600, Dell XPS 210 and Essential Lahey Fortran 90 v. 4 software (Lahey Computer Systems, 1995-2000). A program will be written and compiled using Essential Lahey Fortran 90 v.4 that will compare the Type I error rate of three statistical tests under conditions of both normal and non-normally distributed data sets. Data will be generated using pseudo-random number generator, provided through the Essential Lahey Fortran 90 v.4 software. Sub-routines will be derived from BFRA, a Fortran module, developed by Blair (1987) and updated by Dr. G. Fahoome, Department of Educational Evaluation and Research, at Wayne State University.

Methodology

Utilizing Monte Carlo simulation techniques, Schuirmann's two one-sided *t*-test; Anderson and Hauck's nonequivalence null hypothesis; and Patel and Gupta's procedure will be compared for the probability of detecting equivalence under conditions in which underlying assumptions of normality are violated. The three equivalency tests will be compared with regards to percentage of rejection rates.

Study Parameters

Three variables were manipulated in this study including nominal α level; sample size, sampling distribution, and length of the equivalence interval. The critical mean difference for establishing equivalence with the Schuirmann two one-sided *t*-test was 1 throughout all conditions.

Sample Size and Nominal Alpha.

"One of the primary motivations for utilizing tests of equivalence is that as sample size increases, the probability of finding even trivial mean differences statistically significant becomes larger." (Cribbie, Gruman & Arpin-Cribbie, 2004). The sample size balance and imbalance were selected based on their representation of real world datasets often used in behavioral, health and social sciences research studies (Keselman, et al 1998). For the case of equal numbers of observations per group, one million repetitions were conducted for the sample size combinations n_1, n_2 = (10, 10); (20, 20); (40, 40); and (60, 60) using nominal alpha levels of .001, .01, and .05. For the case of unequal numbers of observations per group, one million repetitions were conducted for sample size combinations n_1, n_2 = (10, 20); (10, 40); (10, 60); (20, 40); (20, 60); and (40, 60) using nominal alpha levels of .001, .01 and .05.

Length of Equivalence Interval.

Three levels of length of equivalence interval in standard deviation units are used. The length of the equivalence interval denoted by Δ is $.001\delta$, $.005\delta$, and $.01\delta$. The standard deviation was set at unity; this is because the length of the equivalency interval is a function of the standard deviation. The purpose of

the study is not to conduct a comparative power analysis: it is a comparison of the properties of showing equivalence. As such, recommended effect sizes should be very small with the point being the reverse of the typical power study. In other words, which of the competitors shows non-equivalence at the smallest effect size? (S. Sawilowsky, personal communication, June 2, 2009).

Sampling Distributions.

For points of comparison, three population distributions have been selected. The selected distributions are the Gaussian (normal) and two identified by Micceri (1989) as possessing “real world data” characteristics representative of education and psychology data sets: the Smooth Symmetric, and Extreme Asymmetry, Achievement. The theoretical variate values generated from the standard normal distribution provide the baseline for comparison with the ‘real world’ variate values generated by the Smooth Symmetric and Extreme Asymmetry, Achievement data sets.

1. Gaussian (Normal) Distribution.

This bell shaped distribution has equally weighted tails and distributions of scores. The mean and median = 0.00, standard deviation = 1.00, skew= 0.00, and kurtosis = 3.00 (Sawilowsky & Blair, 1992).

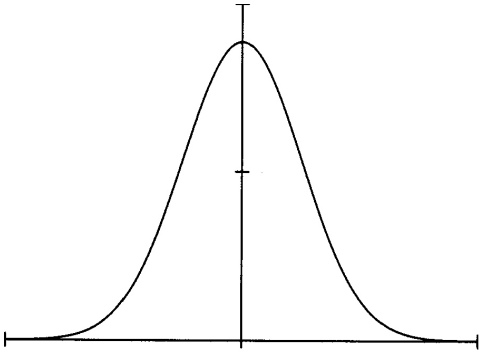

Figure1. Gaussian (Normal) Distribution (Sawilowsky & Fahoome, 2003)

2. Smooth Symmetric.

The Smooth Symmetric data set is similar to the normal distribution however it is distinguished by a light skew and a small variance in kurtosis from the normal distribution. It has a mean = 13.91, median =13.00, standard deviation = 4.91, skew = 0.01 and kurtosis = 2.66. The Smooth Symmetric demonstrates an 11.3% variance from normal kurtosis, thus slightly playkurtic (*Ibid*, 1992).

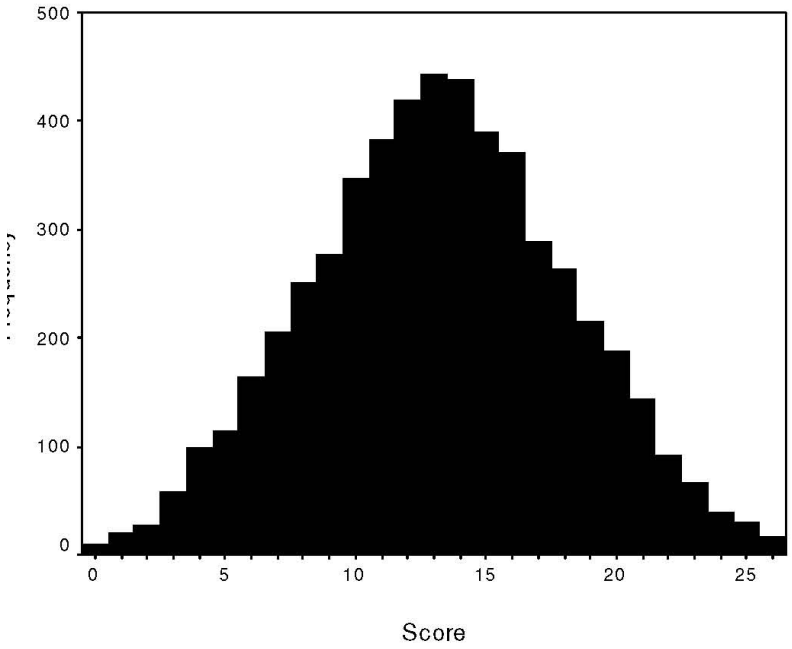

Figure 2. Achievement: Smooth Symmetric (Sawilowsky & Fahoome, 2003)

3. Extreme Asymmetry, Achievement.

The Extreme Asymmetry, Achievement data set has a mean = 24.5, median =27.00, standard deviation = 5.79, skew =1.64, and kurtosis = 4.11. The Extreme Asymmetry, Achievement data set demonstrates a 37% variance (leptokurtic) from normal kurtosis (*Ibid*, 1992).

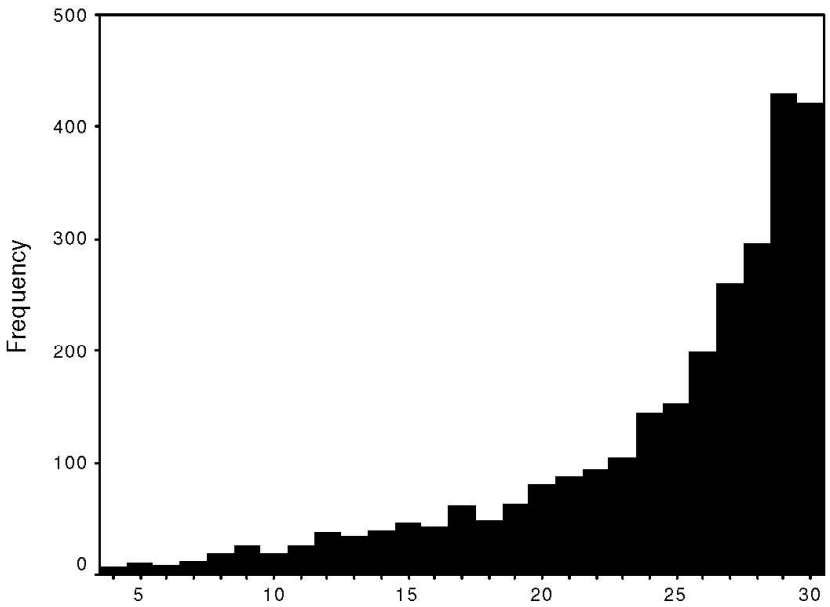

Figure 3. Achievement: Extreme Asymmetry, Growth (Sawilowsky & Fahoome, 2003)

Study Design

For the purposes of this study only the two-sample case was considered. The conditions placed on the generated observations are varied and include nominal α level; sample size, sampling distribution, and length of the equivalence interval. The levels of these experimental conditions are selected so as to reflect test usage in applied studies and to yield reasonable levels of statistical power (≥0.80).

For the purpose of examining the Type I error rate for the combinations previously outlined, 1,000,000 repetitions per condition are simulated. For each repetition, two independent samples are randomly generated based on the given condition. Each of the three tests examined in this study will be applied to the generated samples, and failure to reject/rejection of the null hypothesis will be recorded. Simulated Type I error rates for a test can then be obtained on completion of the 1,000,000 repetitions by dividing the number of times that the

test exceeded the associated critical values by 1,000,000. The simulated Type I error rates for each test will be tabulated and summarized for each condition, then compared with regards to percentage of rejection rates.

The robustness of each statistical test with respect to Type I error rate, will be assessed using the Bradley (1978) liberal criterion test. According to Bradley, a statistical test is determined robust with respect to Type I error rate if the empirical rate of Type I error falls within the range of +/- $.5\alpha$. Specifically, the upper and lower range of robustness is 0.00105 and 0.00095 at nominal α=0.001; 0.0105 and 0.005 at α=0.01; and 0.0525 and 0.0475 at nominal α=0.05. Simulated values above the upper robustness limit will be recorded as liberal (L); simulated values below the lower robustness limit will be recorded as conservative (C). Given the null hypothesis and alternative hypothesis for equivalence:

H_0: The difference between means falls above or below the limits of the equivalency interval; the means are found to be non-equivalent.

H_1: The mean difference falls within the limits of the interval; the means are found to be equivalent.

Within the above paradigm, conservative' suggests that the test declares that the means are equivalent, less often than at the desired Type I error rate, when they are in fact not equivalent. Conversely, liberal suggests that it is concluded that the means are equivalent, more often than at the desired Type I error rate (ibid, 1978).

CHAPTER 4

RESULTS

The findings are presented from the Monte Carlo simulation study on the effects of non-normal distributions on the performance of Schuirmann's two one-sided t-test, Anderson & Hauck's non-equivalence null hypothesis, and Patel-Gupta's procedure, under conditions of small effect sizes. The first section presents findings as they relate to the effects of population variability on the performance of each test, as measured by percent of rejection rate. The second section presents a comparison of the properties of showing equivalence, for the purpose of determining which equivalency test showed non-equivalence at the smallest effect size, .001δ selected for this study.

Non-Normal Distribution Effects

Table 2 displays the average rejection rates under each condition (Gaussian/normal, Smooth Symmetric & Extreme Asymmetry) for the nominal α levels of .001, .01, and .05 at different levels of sample size combinations. The average rejection rates under the normal distribution, represents the benchmark for comparison with average rejection rates under the Smooth Symmetric and Extreme Asymmetric data sets. For the nominal α =.001 level and sample sizes n_1, n_2=10, 10 through n_1, n_2=20, 40, the tests were applied to non-normally distributed sample sets where they rejected at statistically significantly higher rates in comparison to tests applied to normally distributed sample sets. As nominal α level increased to .01, tests applied to non-normally distributed

sample sets n_1, n_2=10, 60 through n_1, n_2=60, 60 rejected at slightly lower rates in comparison to normally distributed sample sets.

Table 2. Percentage of Rejection Rates (Average) for nominal α at .001 δ

n	α	NORM	SSYM	EXASY
10, 10	.001	.0%	.00067%	.00067%
	.01	.00067%	.00067%	.00067%
	.05	.00096%	.00096%	.00096%
10, 20	.001	.0%	.00060%	.00060%
	.01	.00060%	.00060%	.00060%
	.05	.00087%	.00087%	.00087%
10, 40	.001	.0%	.00053%	.00053%
	.01	.00053%	.00053%	.00053%
	.05	.00077%	.00077%	.00077%
10, 60	.001	.0%	.00047%	.00047%
	.01	.00063%	.00047%	.00047%
	.05	.00067%	.00050%	.00050%
20, 20	.001	.0%	.00040%	.00040%
	.01	.00057%	.00040%	.00040%
	.05	.00057%	.00043%	.00043%
20, 40	.001	.0%	.00033%	.00033%
	.01	.00050%	.00033%	.00033%
	.05	.00040%	.00033%	.00033%
20, 60	.001	.00027%	.00027%	.00027%
	.01	.00040%	.00027%	.00027%
	.05	.00030%	.00030%	.00030%
40, 40	.001	.00030%	.00020%	.00020%
	.01	.00030%	.00020%	.00020%
	.05	.00027%	.00023%	.00023%
40, 60	.001	.00013%	.00013%	.00013%
	.01	.00020%	.00013%	.00013%
	.05	.00020%	.00020%	.00020%
60, 60	.001	.00010%	.000060%	.000060%
	.01	.00010%	.000060%	.000060%
	.05	.00010%	.000060%	.000060%

Gaussian

Under the Gaussian distribution, it was determined that as the equivalency length interval increased from $.001\delta$, to $.005\delta$, to $.01\delta$, respectively; no changes in percentage of rejection rates occurred. Given these findings, only outcomes reported at the equivalency length interval, $.001\delta$, will be presented. Tables for outcomes at equivalency length intervals $.005\delta$ and $.01\delta$, are presented in Appendix A and B, respectively. Further investigation determined statistically significantly lower (ρ<.05) rejection rates at nominal alpha level α =.001and sample sizes n_1, n_2=10, 10 through n_1, n_2=20, 40 in comparison with rejection rates at nominal α =.01 and .05, respectively. For the above conditions, it was determined that all three tests showed non-equivalence (see Table 3). For nominal *a*=.01, for sample sizes n_1, n_2=10, 10 and n_1, n_2=10, 40; the Schuirmann t-test rejected at a statistically significantly lower rate (ρ<.05) in comparison to the Anderson & Hauck and Patel-Gupta tests.

Under the normal distribution, the Schuirmann t-test showed non-equivalence in approximately 90% of conditions (n_1, n_2=10, 10 *through* n_1, n_2=20, 60), in comparison with the Anderson & Hauck (60%; n_1, n_2=10, 10 through n_1, n_2=20, 40), and Patel-Gupta tests (60%; n_1, n_2=10, 10 through n_1, n_2=20, 40), respectively. Overall, under normal conditions, all three traditional equivalency tests showed non-equivalence under the smallest equivalency interval, for the lowest nominal *a* level, for samples sizes $N \leq 60$.

Table 3. Gaussian: Percentage of Rejection Rates for nominal α at .001 δ

n	α	SCHUI	A&H	P-G
10, 10	.001	.0%	.0%	.0%
	.01	.0%	.001%	.001%
	.05	.0009%	.001%	.001%
10, 20	.001	.0%	.0%	.0%
	.01	.0%	.0009%	.0009%
	.05	.0008%	.0009%	.0009%
10, 40	.001	.0%	.0%	.0%
	.01	.0%	.0008%	.0008%
	.05	.0007%	.0008%	.0008%
10, 60	.001	.0%	.0%	.0%
	.01	.0005%	.0007%	.0007%
	.05	.0006%	.0007%	.0007%
20, 20	.001	.0%	.0%	.0%
	.01	.0005%	.0006%	.0006%
	.05	.0005%	.0006%	.0006%
20, 40	.001	.0%	.0%	.0%
	.01	.0005%	.0005%	.0005%
	.05	.0004%	.0004%	.0004%
20, 60	.001	.0%	.0004%	.0004%
	.01	.0004%	.0004%	.0004%
	.05	.0003%	.0004%	.0004%
40, 40	.001	.0003%	.0003%	.0003%
	.01	.0003%	.0003%	.0003%
	.05	.0002%	.0003%	.0003%
40, 60	.001	.0%	.0002%	.0002%
	.01	.0002%	.0002%	.0002%
	.05	.0002%	.0002%	.0002%
60, 60	.001	.0001%	.0001%	.0001%
	.01	.0001%	.0001%	.0001%
	.05	.0001%	.0001%	.0001%

Smooth Symmetric, Achievement

The Smooth Symmetric data set is most similar in behavior to the standard normal distribution, and is the population set identified most closely with the Gaussian (Micceri, 1989). Comparison of rejection rates generated under the Smooth Symmetric data set revealed both similarities (inner test-consistency), and differences (rate of rejections) to outcomes produced under the normal distribution. For nominal α =.001 and sample sizes n_1, n_2=10, 10 through n_1, n_2=20, 40; the Anderson & Hauck and Patel-Gupta tests rejected at statistically significantly (ρ<.05) higher rates in comparison with rejection rates reported by both tests under the normal distribution. For all three nominal α levels, the Anderson & Hauck and Patel-Gupta tests rejected at approximately the same rates. In contrast, for the nominal α =.001, the rejection rate of the Schuirmann t-test decreased from 30% (normal distribution) to .0% (Smooth Symmetric). Furthermore, as the nominal alpha level increased from nominal α =.001 to nominal α =.01, the rejection rate of the test decreased from 70% (normal) to .0% (Smooth Symmetric).

Table 4. Smooth Symmetric: Percentage of Rejection Rates for nominal α at .001 δ

n	α	SCHUI	A&H	P-G
10, 10	.001	.0%	.001%	.001%
	.01	.0%	.001%	.001%
	.05	.0009%	.001%	.001%
10, 20	.001	.0%	.0009%	.0009%
	.01	.0%	.0009%	.0009%
	.05	.0008%	.0009%	.0009%
10, 40	.001	.0%	.0008%	.0008%
	.01	.0%	.0008%	.0008%
	.05	.0007%	.0008%	.0008%
10, 60	.001	.0%	.0007%	.0007%
	.01	.0%	.0007%	.0007%
	.05	.0001%	.0007%	.0007%
20, 20	.001	.0%	.0006%	.0006%
	.01	.0%	.0006%	.0006%
	.05	.0001%	.0006%	.0006%
20, 40	.001	.0%	.0005%	.0005%
	.01	.0%	.0005%	.0005%
	.05	.0001%	.0005%	.0005%
20, 60	.001	.0%	.0004%	.0004%
	.01	.0%	.0004%	.0004%
	.05	.0001%	.0004%	.0004%
40, 40	.001	.0%	.0003%	.0003%
	.01	.0%	.0003%	.0003%
	.05	.0001%	.0003%	.0003%
40, 60	.001	.0%	.0002%	.0002%
	.01	.0%	.0002%	.0002%
	.05	.0002%	.0002%	.0002%
60, 60	.001	.0%	.0001%	.0001%
	.01	.0%	.0001%	.0001%
	.05	.0001%	.0001%	.0001%

Extreme Asymmetry, Achievement

The Extreme Asymmetry data set is distinguished by its extreme negative skew, in comparison to the Gaussian distribution. For the three nominal α levels, the performance of the Anderson & Hauck and Patel-Gupta tests mirrored the performance under the Smooth Symmetric data set. In contrast, for the nominal α =.05, the rejection rate of the Schuirmann t-test decreased from 100% (normal) to 80% (Extreme Asymmetry). For nominal α =.05, the Schuirmann t-test showed non-equivalence for sample sizes n_1n_2 =40, 60 and n_1n_2 =60, 60.

Table 5. Extreme Asymmetry: Percentage of Rejection Rates for nominal α at .001 δ

n	α	SCHUI	A&H	P-G
10, 10	.001	.0%	.001%	.001%
	.01	.0%	.001%	.001%
	.05	.0001%	.001%	.001%
10, 20	.001	.0%	.0009%	.0009%
	.01	.0%	.0009%	.0009%
	.05	.0001%	.0009%	.0009%
10, 40	.001	.0%	.0008%	.0008%
	.01	.0%	.0008%	.0008%
	.05	.0001%	.0008%	.0008%
10, 60	.001	.0%	.0007%	.0007%
	.01	.0%	.0007%	.0007%
	.05	.0001%	.0007%	.0007%
20, 20	.001	.0%	.0006%	.0006%
	.01	.0%	.0006%	.0006%
	.05	.0001%	.0006%	.0006%
20, 40	.001	.0%	.0005%	.0005%
	.01	.0%	.0005%	.0005%
	.05	.0001%	.0005%	.0005%
20, 60	.001	.0%	.0004%	.0004%
	.01	.0%	.0004%	.0004%
	.05	.0001%	.0004%	.0004%
40, 40	.001	.0%	.0003%	.0003%
	.01	.0%	.0003%	.0003%
	.05	.0001%	.0003%	.0003%
40, 60	.001	.0%	.0002%	.0002%
	.01	.0%	.0002%	.0002%
	.05	.0%	.0002%	.0002%
60, 60	.001	.0%	.0001%	.0001%
	.01	.0%	.0001%	.0001%
	.05	.0%	.0001%	.0001%

Performance of Equivalency Tests

The outcomes obtained in the Monte Carlo study supported previously published findings of the general performance of each of the above tests under standard normal conditions. Additional insight specific to (a) comparisons of the properties of showing equivalence, and (b) which of the above competitors showed non-equivalence at the smallest effect size, .001, selected for this study.

Table 6. Type I Error Rate of Equal Sample Sizes for nominal α =.001 at .001 δ

	Schuirmann	Anderson & Hauk	Patel & Gupta
Normal			
n_1, n_2=10,10	.000000	.000000	.000000
n_1, n_2=20,20	.000000	.000000	.000000
n_1, n_2=40,40	.000003	.000003	.000003
n_1, n_2=60,60	.000001	.000001	.000001
Smooth Symmetric			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000000	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001
Extreme Asymmetry			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000000	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001

Table 7. Type I Error Rate of Equal Sample Sizes for nominal α =.01 at .001 δ

	Schuirmann	Anderson & Hauk	Patel & Gupta
Normal			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000005	.000006	.000006
n_1, n_2=40,40	.000003	.000003	.000003
n_1, n_2=60,60	.000001	.000001	.000001
Smooth Symmetric			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000000	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001
Extreme Asymmetry			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000000	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001

Table 8. Type I Error Rate of Equal Sample Sizes for nominal α =.05 at .001 δ

	Schuirmann	Anderson & Hauk	Patel & Gupta
Normal			
n_1, n_2=10,10	.000009	.000010	.000010
n_1, n_2=20,20	.000005	.000006	.000006
n_1, n_2=40,40	.000002	.000003	.000003
n_1, n_2=60,60	.000001	.000001	.000001
Smooth Symmetric			
n_1, n_2=10,10	.000009	.000010	.000010
n_1, n_2=20,20	.000001	.000006	.000006
n_1, n_2=40,40	.000001	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001
Extreme Asymmetry			
n_1, n_2=10,10	.000001	.000010	.000010
n_1, n_2=20,20	.000001	.000006	.000006
n_1, n_2=40,40	.000001	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001

Table 9. Type I Error Rate of Unequal Sample Sizes for nominal α =.001 at .001 δ

	Schuirmann	Anderson & Hauk	Patel & Gupta
Normal			
n_1, n_2=10, 20	.000000	.000000	.000000
n_1, n_2=10, 40	.000000	.000000	.000000
n_1, n_2=10, 60	.000000	.000000	.000000
n_1, n_2=20, 40	.000005	.000000	.000005
n_1, n_2=20, 60	.000000	.000004	.000004
n_1, n_2=40, 60	.000002	.000002	.000002
Smooth Symmetric			
n_1, n_2=10, 20	.000000	.000009	.000009
n_1, n_2=10, 40	.000000	.000008	.000008
n_1, n_2=10, 60	.000000	.000007	.000007
n_1, n_2=20, 40	.000000	.000005	.000005
n_1, n_2=20, 60	.000000	.000004	.000004
n_1, n_2=40, 60	.000000	.000002	.000002
Extreme Asymmetry			
n_1, n_2=10, 20	.000000	.000009	.000009
n_1, n_2=10, 40	.000000	.000008	.000008
n_1, n_2=10, 60	.000000	.000007	.000007
n_1, n_2=20, 40	.000000	.000005	.000005
n_1, n_2=20, 60	.000000	.000004	.000004
n_1, n_2=40, 60	.000000	.000002	.000002

Table 10. Type I Error Rate of Unequal Sample Sizes for nominal α =.01 at .001 δ

	Schuirmann	Anderson & Hauk	Patel & Gupta
Normal			
n_1, n_2=10, 20	.000000	.000009	.000009
n_1, n_2=10, 40	.000000	.000008	.000008
n_1, n_2=10, 60	.000005	.000007	.000007
n_1, n_2=20, 40	.000005	.000005	.000005
n_1, n_2=20, 60	.000004	.000004	.000004
n_1, n_2=40, 60	.000002	.000002	.000002
Smooth Symmetric			
n_1, n_2=10, 20	.000000	.000009	.000009
n_1, n_2=10, 40	.000000	.000008	.000008
n_1, n_2=10, 60	.000000	.000007	.000007
n_1, n_2=20, 40	.000000	.000005	.000005
n_1, n_2=20, 60	.000000	.000004	.000004
n_1, n_2=40, 60	.000000	.000002	.000002
Extreme Asymmetry			
n_1, n_2=10, 20	.000000	.000009	.000009
n_1, n_2=10, 40	.000000	.000008	.000008
n_1, n_2=10, 60	.000000	.000007	.000007
n_1, n_2=20, 40	.000000	.000005	.000005
n_1, n_2=20, 60	.000000	.000004	.000004
n_1, n_2=40, 60	.000000	.000002	.000002

Table 11. Type I Error Rate of Unequal Sample Sizes for nominal α =.05 at .001 δ

	Schuirmann	Anderson & Hauk	Patel & Gupta
Normal			
n_1, n_2=10, 20	.000008	.000009	.000009
n_1, n_2=10, 40	.000007	.000008	.000008
n_1, n_2=10, 60	.000006	.000007	.000007
n_1, n_2=20, 40	.000004	.000005	.000005
n_1, n_2=20, 60	.000003	.000004	.000004
n_1, n_2=40, 60	.000002	.000002	.000002
Smooth Symmetric			
n_1, n_2=10, 20	.000008	.000009	.000009
n_1, n_2=10, 40	.000007	.000008	.000008
n_1, n_2=10, 60	.000001	.000007	.000007
n_1, n_2=20, 40	.000001	.000005	.000005
n_1, n_2=20, 60	.000001	.000004	.000004
n_1, n_2=40, 60	.000002	.000002	.000002
Extreme Asymmetry			
n_1, n_2=10, 20	.000001	.000009	.000009
n_1, n_2=10, 40	.000001	.000008	.000008
n_1, n_2=10, 60	.000001	.000007	.000007
n_1, n_2=20, 40	.000001	.000005	.000005
n_1, n_2=20, 60	.000001	.000004	.000004
n_1, n_2=40, 60	.000000	.000002	.000002

Schuirmann Two One-Sided t-test

Table 12 displays the percentage of rejection rates for the Schuirmann t-test for the three nominal α levels for all sample size combinations. The Schuirmann t-test was the most conservative, rejecting at statistically significant lower rates (ρ=.05) in comparison to the more liberal, Anderson & Hauk and Patel-Gupta procedures. When data was sampled from Smooth Symmetric data set, the Schuirmann t-test, demonstrated non-equivalence at a statistically significantly higher rate (ρ=.05) in comparison to results produced under the normal distribution for nominal α levels .001 and .01, for all sample size combinations.

With data sampled from the Extreme Asymmetry data set, the Schuirmann t-test mirrored its performance under the Smooth Symmetric. In addition, for nominal α = .05, sample size combinations $n_1 n_2$ (40, 60; 60, 60), this test demonstrated non-equivalence at a statistically significantly higher rate (ρ=.05) in comparison with results produced under the normal distribution. The Schuirmann t-test, for all three nominal α levels, for all sample size combinations, rejected (.0%) at a statistically significantly lower rate (ρ=.001) in comparison to results produced under the normal distribution.

Table 12. Schuirmann: Percentage of Rejection Rates for nominal α at .001δ

n	α	Norm	SS	EA
10, 10	.001	.0%	.0%	.0%
	.01	.0%	.0%	.0%
	.05	.0009%	.0001%	.0001%
10, 20	.001	.0%	.0%	.0%
	.01	.0%	.0%	.0%
	.05	.0008%	.0008%	.0001%
10, 40	.001	.0%	.0%	.0%
	.01	.0%	.0%	.0%
	.05	.0007%	.0007%	.0001%
10, 60	.001	.0%	.0%	.0%
	.01	.0005%	.0%	.0%
	.05	.0006%	.0001%	.0001%
20, 20	.001	.0%	.0%	.0%
	.01	.0005%	.0%	.0%
	.05	.0009%	.0001%	.0001%
20, 40	.001	.0%	.0%	.0%
	.01	.0005%	.0%	.0%
	.05	.0004%	.0001%	.0001%
20, 60	.001	.0%	.0%	.0%
	.01	.0004%	.0%	.0%
	.05	.0003%	.0001%	.0001%
40, 40	.001	.0003%	.0%	.0%
	.01	.0003%	.0%	.0%
	.05	.0002%	.0001%	.0001%
40, 60	.001	.0%	.0%	.0%
	.01	.0002%	.0%	.0%
	.05	.0002%	.0002%	.0%
60, 60	.001	.0001%	.0%	.0%
	.01	.0001%	.0%	.0%
	.05	.0001%	.0001%	.0%

Anderson and Hauck's Nonequivalence Null Hypothesis

Table 13 displays the percentage of rejection rates for the Anderson & Hauck nonequivalence null hypothesis procedure for the three nominal α levels for all sample size combinations. The Anderson & Hauk is one of two liberal equivalency test examined in this study (Frick, 1990). For data sampled from the Smooth Symmetric and Extreme Asymmetry data sets, with nominal α =.001, small n_1n_2 (10, 10; 10, 20) to medium n_1n_2 (20, 20; 10, 40) size samples, the Anderson & Hauck procedure rejected at statistically significantly higher rates in comparison with results produced under the normal distribution. For data sampled from Extreme Asymmetry data set, for nominal α =.01, equal sample sizes n_1n_2 (10, 10; 20, 20; 40, 40; 60, 60), this procedure rejected at statistically significantly higher rates in comparison with results produced under the normal distribution.

Table 13. Anderson & Hauck: Percentage of Rejection Rates for nominal α at .001 δ

n	α	Norm	SS	EA
10, 10	.001	.0%	.001%	.001%
	.01	.001%	.001%	.001%
	.05	.001%	.001%	.001%
10, 20	.001	.0%	.0009%	.0009%
	.01	.0009%	.0009%	.0009%
	.05	.0009%	.0009%	.0009%
10, 40	.001	.0%	.0008%	.0008%
	.01	.0008%	.0008%	.0008%
	.05	.0008%	.0008%	.0008%
10, 60	.001	.0%	.0007%	.0007%
	.01	.0007%	.0007%	.0007%
	.05	.0007%	.0007%	.0007%
20, 20	.001	.0%	.0006%	.0003%
	.01	.0006%	.0006%	.001%
	.05	.001%	.0006%	.0003%
20, 40	.001	.0%	.0005%	.0005%
	.01	.0005%	.0005%	.0005%
	.05	.0004%	.0005%	.0005%
20, 60	.001	.0004%	.0004%	.0004%
	.01	.0004%	.0004%	.0004%
	.05	.0004%	.0004%	.0004%
40, 40	.001	.0003%	.0003%	.0003%
	.01	.0003%	.0003%	.001%
	.05	.0003%	.0003%	.0003%
40, 60	.001	.0002%	.0002%	.0002%
	.01	.0002%	.0002%	.0002%
	.05	.0002%	.0002%	.0002%
60, 60	.001	.0001%	.0001%	.0001%
	.01	.0001%	.0001%	.0006%
	.05	.0001%	.0001%	.0001%

Patel-Gupta Procedure

Table 14 displays the percentage of rejection rates for the Patel-Gupta procedure for the three nominal α levels for all sample size combinations. The Patel-Gupta procedure is considered to be a liberal equivalency test rejecting at rates comparable with the Anderson & Hauck procedure (Frick, 1990). Similar to the Anderson & Hauck, under conditions of the Smooth Symmetric and Extreme data sets, nominal α =.001, small $n_1 n_2$ (10, 10; 10, 20) to medium $n_1 n_2$ (20, 20; 10, 40) size samples, the Patel-Gupta procedure rejected at statistically significantly higher rates in comparison to results produced under the normal distribution. For nominal α =.01 and .05 levels, the Patel-Gupta procedure rejected at comparable rates to the Anderson & Hauck procedure under all data sets.

Table 14. Patel-Gupta: Percentage of Rejection Rates for nominal α at .001 δ

n	α	Norm	SS	EA
10, 10	.001	.0%	.001%	.001%
	.01	.001%	.001%	.001%
	.05	.001%	.001%	.001%
10, 20	.001	.0%	.0009%	.0009%
	.01	.0009%	.0009%	.0009%
	.05	.0009%	.0009%	.0009%
10, 40	.001	.0%	.0008%	.0008%
	.01	.0008%	.0008%	.0008%
	.05	.0008%	.0008%	.0008%
10, 60	.001	.0%	.0007%	.0007%
	.01	.0007%	.0007%	.0007%
	.05	.0007%	.0007%	.0007%
20, 20	.001	.0%	.0006%	.0003%
	.01	.0006%	.0006%	.001%
	.05	.001%	.0006%	.0003%
20, 40	.001	.0%	.0005%	.0005%
	.01	.0005%	.0005%	.0005%
	.05	.0004%	.0005%	.0005%
20, 60	.001	.0004%	.0004%	.0004%
	.01	.0004%	.0004%	.0004%
	.05	.0004%	.0004%	.0004%
40, 40	.001	.0003%	.0003%	.0003%
	.01	.0003%	.0003%	.0003%
	.05	.0003%	.0003%	.0003%
40, 60	.001	.0002%	.0002%	.0002%
	.01	.0002%	.0002%	.0002%
	.05	.0002%	.0002%	.0002%
60, 60	.001	.0001%	.0001%	.0001%
	.01	.0001%	.0001%	0001%
	.05	.0001%	.0001%	.0001%

CHAPTER 5

DISCUSSION

The present study examined the effects and management of non-normally distributed data on equivalency tests under varied conditions for a two-sample design; and compared the properties of showing equivalence between populations at the smallest effect sizes. An increasing body of literature has evolved within the social sciences research paradigms (clinical psychology, management operations) where the question of interest is whether the difference between two treatment means is large enough to be considered statistically significantly meaningful. Articles appearing in both health and social science literature (e.g., Brown, Hwang & Munk, 1997; Breslow & Day, 1980) have increased both the availability and the popularity of these procedures. However, little research into the statistical properties of these procedures under conditions of non-normally distributed data, and small effect sizes has been conducted.

Summary of Tests' Performance

The findings indicated that under conditions where sample sets were non-normally distributed, the differences in the statistical properties of the three equivalency tests became most pronounced at the lowest nominal α =.001 for small to medium sample sizes. As defined previously in Chapter One of this study, statistical power is the probability of detecting an effect, given that the effect is actually present (Hinkle, Weirsman & Jurs, 1998). Overall, all three tests demonstrated low power (≤0.80) due to the relatively small sample size

combinations paired with small effect sizes(.001δ), and failed to control Type I error. The Anderson & Hauck and Patel-Gupta tests reported rejection rates that remained relatively stable without regard to sample size or alpha level. In contrast, the rate of rejection reported for the typically conservative Schuirmann t-test, decreased steadily without regard to increase in both sample size and alpha level.

Schuirmann Two One-Sided t-test

Findings for this simulation study supported previous studies, showing the Schuirmann t-test to be extremely conservative, specifically for small (10, 10; 10, 20) to medium (20, 20; 10, 40) sample size combinations under the Gaussian distribution. With increased sample size, the t-test improved its rejection rate, although the test never controlled the Type I error rate. With the introduction of non-normally distributed sample sizes, the t-test failed to reject (.0%) for all sample sizes, as the nominal α=.001 rate increased to α=.01.

For data sampled from the data set which most closely resembles the normal distribution, the Smooth Symmetric; the test expanded its range for showing non-equivalence to include all sample sizes paired with nominal α=.001 and .01. For data sampled from the Extreme Asymmetry data set, the test expanded its range for showing non-equivalence to include sample sizes N≥100 paired with nominal α=.05. The above results are consistent with the manner in which the critical values for the Schuirmann t-test are determined: the smaller the interval width or the lower the nominal α level, the less room the t-test has for

determining critical values and the less the actual condition resembles the way that the critical values are determined.

Anderson & Hauck Non-Equivalence / Patel-Gupta Procedure

There were only two conditions where the Anderson & Hauck and Patel-Gupta procedures approached control of the Type I error rate: for the conditions of nominal α =.001, for sample sizes $n_1 n_2$ (10, 10) to $n_1 n_2$ (10, 20) drawn from the Smooth Symmetric and Extreme Asymmetry data sets. In terms of rejection rates, both procedures reported maximum rejection rates for all three nominal α levels, for non-equal small $n_1 n_2$ (10, 20) to medium $n_1 n_2$ (10, 40) size samples.

Recommendations for Further Research

In general, all three tests demonstrated low power (≤0.80) due to the relatively small sample size combinations paired with small effect sizes (.001 δ). Optimal performance in relation to detecting equivalence occurred for the nominal α =.001, for small sample sizes $n_1 n_2$ (10, 10; 10, 20) for data sampled under the Smooth Symmetric and Extreme Asymmetry data sets. However, the power properties of both tests were extremely low (≤0.80), and all three tests failed to control Type I error. Based on the findings of this study, none of the three tests are recommended as being superior to the other.

To more accurately understand the behavior of equivalency tests under conditions of small effect sizes, a more thorough study is recommended. Firstly, the sample sizes selected for this study did not result in sufficient power (≥0.80) for the three traditional equivalency tests. Pairing minimal sample size combinations of $n_1 n_2$=80,140 through $n_1 n_2$=120, 120 (10,000 repetitions), with

the effect sizes selected for this study would be expected to produce adequate power for the traditional tests. Secondly, this study did not consider the potential effects of variance heterogeneity on the standard error of the individual tests. Gruman, Cribbie & Arpin-Cribbie (2007), initiated work in this area by exploring modified tests of equivalence that incorporate heteroscedasticity error terms. Finally, future areas of research might include testing for non-equivalence under conditions where there are three or more groups, or where groups are dependent.

APPENDIX A. Individual Tests

Table 15. Schuirmann One-Sided T-Test Type I Error Rate by Sampling Distribution

	.001 δ								
Sample	**Gaussian**			**Smooth Symmetric**			**Extreme Asymmetry**		
	α =.05	α =.01	α =.001	α =.05	α =.01	α =.001	α =.05	α =.01	α =.001
(10,10)	.000009	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,20)	.000005	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(40,40)	.000002	.000003	.000003	.000001	.000000	.000000	.000001	.000000	.000000
(60,60)	.000001	.000001	.000001	.000000	.000000	.000000	.000000	.000000	.000000
(10,20)	.000008	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(10,40)	.000007	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(10,60)	.000006	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,40)	.000004	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,60)	.000003	.000004	.000004	.000001	.000000	.000000	.000001	.000000	.000000
(40,60)	.000002	.000002	.000002	.000000	.000000	.000000	.000000	.000000	.000000
	.005 δ								
(10,10)	.000009	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,20)	.000005	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(40,40)	.000002	.000003	.000003	.000001	.000000	.000000	.000001	.000000	.000000
(60,60)	.000001	.000001	.000001	.000000	.000000	.000000	.000000	.000000	.000000
(10,20)	.000008	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(10,40)	.000007	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(10,60)	.000006	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,40)	.000004	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,60)	.000003	.000004	.000004	.000001	.000000	.000000	.000001	.000000	.000000
(40,60)	.000002	.000002	.000002	.000000	.000000	.000000	.000000	.000000	.000000
	.01 δ								
(10,10)	.000009	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,20)	.000005	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(40,40)	.000002	.000003	.000003	.000001	.000000	.000000	.000001	.000000	.000000
(60,60)	.000001	.000001	.000001	.000000	.000000	.000000	.000000	.000000	.000000
(10,20)	.000008	.000000	.000000	.000000	.000000	.000000	.000001	.000000	.000000
(10,40)	.000007	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(10,60)	.000006	.000005	.000000	.000000	.000000	.000000	.000001	.000000	.000000
(20,40)	.000004	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,60)	.000003	.000004	.000004	.000001	.000000	.000000	.000001	.000000	.000000
(40,60)	.000002	.000002	.000002	.000000	.000000	.000000	.000000	.000000	.000000

Table 16. Anderson & Hauck Non-Equivalent Null Hypothesis Type I Error Rate by Sampling Distribution

	.001 δ								
Sample	**Gaussian**			**Smooth Symmetric**			**Extreme Asymmetry**		
	α =.05	α =.01	α =.001	α =.05	α =.01	α =.001	α =.05	α =.01	α =.001
(10,10)	.000009	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,20)	.000005	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(40,40)	.000002	.000003	.000003	.000001	.000000	.000000	.000001	.000000	.000000
(60,60)	.000001	.000001	.000001	.000000	.000000	.000000	.000000	.000000	.000000
(10,20)	.000008	.000000	.000000	.000000	.000000	.000000	.000001	.000000	.000000
(10,40)	.000007	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(10,60)	.000006	.000005	.000000	.000000	.000000	.000000	.000001	.000000	.000000
(20,40)	.000004	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,60)	.000003	.000004	.000004	.000001	.000000	.000000	.000001	.000000	.000000
(40,60)	.000002	.000002	.000002	.000000	.000000	.000000	.000000	.000000	.000000
	.005 δ								
(10,10)	.000009	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,20)	.000005	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(40,40)	.000002	.000003	.000003	.000001	.000000	.000000	.000001	.000000	.000000
(60,60)	.000001	.000001	.000001	.000000	.000000	.000000	.000000	.000000	.000000
(10,20)	.000008	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(10,40)	.000007	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(10,60)	.000006	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,40)	.000004	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,60)	.000003	.000004	.000004	.000001	.000000	.000000	.000001	.000000	.000000
(40,60)	.000002	.000002	.000002	.000000	.000000	.000000	.000000	.000000	.000000
	.01 δ								
(10,10)	.000009	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,20)	.000005	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(40,40)	.000002	.000003	.000003	.000001	.000000	.000000	.000001	.000000	.000000
(60,60)	.000001	.000001	.000001	.000000	.000000	.000000	.000000	.000000	.000000
(10,20)	.000008	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(10,40)	.000007	.000000	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(10,60)	.000006	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,40)	.000004	.000005	.000000	.000001	.000000	.000000	.000001	.000000	.000000
(20,60)	.000003	.000004	.000004	.000001	.000000	.000000	.000001	.000000	.000000
(40,60)	.000002	.000002	.000002	.000000	.000000	.000000	.000000	.000000	.000000

Table 17. Patel-Gupta Procedure Type I Error Rate by Sampling Distribution

	.001 δ								
Sample	**Gaussian**			**Smooth Symmetric**			**Extreme Asymmetry**		
	α =.05	α =.01	α =.001	α =.05	α =.01	α =.001	α =.05	α =.01	α =.001
(10,10)	.000010	.000010	.000000	.000010	.000010	.000010	.000010	.000010	.000010
(20,20)	.000006	.000006	.000000	.000006	.000006	.000006	.000006	.000006	.000006
(40,40)	.000003	.000000	.000000	.000003	.000003	.000003	.000003	.000003	.000003
(60,60)	.000001	.000001	.000000	.000001	.000001	.000001	.000001	.000001	.000001
(10,20)	.000009	.000009	.000000	.000009	.000009	.000009	.000009	.000009	.000009
(10,40)	.000008	.000008	.000000	.000008	.000008	.000008	.000008	.000008	.000008
(10,60)	.000007	.000007	.000000	.000007	.000007	.000007	.000007	.000007	.000007
(20,40)	.000005	.000005	.000000	.000005	.000005	.000005	.000005	.000005	.000005
(20,60)	.000004	.000004	.000000	.000004	.000004	.000004	.000004	.000004	.000004
(40,60)	.000002	.000002	.000000	.000002	.000002	.000002	.000002	.000002	.000002
	.005 δ								
(10,10)	.000010	.000010	.000000	.000010	.000010	.000010	.000010	.000010	.000010
(20,20)	.000006	.000006	.000000	.000006	.000006	.000006	.000006	.000006	.000006
(40,40)	.000003	.000000	.000000	.000003	.000003	.000003	.000003	.000003	.000003
(60,60)	.000001	.000001	.000000	.000001	.000001	.000001	.000001	.000001	.000001
(10,20)	.000009	.000009	.000000	.000009	.000009	.000009	.000009	.000009	.000009
(10,40)	.000008	.000008	.000000	.000008	.000008	.000008	.000008	.000008	.000008
(10,60)	.000007	.000007	.000000	.000007	.000007	.000007	.000007	.000007	.000007
(20,40)	.000005	.000005	.000000	.000005	.000005	.000005	.000005	.000005	.000005
(20,60)	.000004	.000004	.000000	.000004	.000004	.000004	.000004	.000004	.000004
(40,60)	.000002	.000002	.000000	.000002	.000002	.000002	.000002	.000002	.000002
	.01 δ								
(10,10)	.000010	.000010	.000000	.000010	.000010	.000010	.000010	.000010	.000010
(20,20)	.000006	.000006	.000000	.000006	.000006	.000006	.000006	.000006	.000006
(40,40)	.000003	.000000	.000000	.000003	.000003	.000003	.000003	.000003	.000003
(60,60)	.000001	.000001	.000000	.000001	.000001	.000001	.000001	.000001	.000001
(10,20)	.000009	.000009	.000000	.000009	.000009	.000009	.000009	.000009	.000009
(10,40)	.000008	.000008	.000000	.000008	.000008	.000008	.000008	.000008	.000008
(10,60)	.000007	.000007	.000000	.000007	.000007	.000007	.000007	.000007	.000007
(20,40)	.000005	.000005	.000000	.000005	.000005	.000005	.000005	.000005	.000005
(20,60)	.000004	.000004	.000000	.000004	.000004	.000004	.000004	.000004	.000004
(40,60)	.000002	.000002	.000000	.000002	.000002	.000002	.000002	.000002	.000002

APPENDIX B. Findings for .005 δ and .01 δ

Table 18. Comparative Type I Error Rates for Equal Size Samples at .005 δ

	α =.001 and .005 δ		
	Schuirmann	**Anderson & Hauk**	**Patel & Gupta**
Normal			
n_1, n_2=10,10	.000000	.000000	.000000
n_1, n_2=20,20	.000000	.000000	.000000
n_1, n_2=40,40	.000003	.000003	.000003
n_1, n_2=60,60	.000001	.000001	.000001
Smooth Symmetric			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000000	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001
Extreme Asymmetry			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000000	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001
	α =.01 and .005 δ		
Normal			
n_1, n_2=10,10	.000000	.000000	.000010
n_1, n_2=20,20	.000005	.000006	.000006
n_1, n_2=40,40	.000003	.000003	.000003
n_1, n_2=60,60	.000001	.000001	.000001
Smooth Symmetric			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000000	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001
Extreme Asymmetry			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000000	.000003	.000003
	α =.05 and 005 δ		
Normal			
n_1, n_2=10,10	.000009	.000010	.000010
n_1, n_2=20,20	.000005	.000006	.000006
n_1, n_2=40,40	.000002	.000003	.000003
n_1, n_2=60,60	.000001	.000001	.000001
Smooth Symmetric			
n_1, n_2=10,10	.000001	.000010	.000010
n_1, n_2=20,20	.000001	.000006	.000006
n_1, n_2=40,40	.000001	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001
Extreme Asymmetry			
n_1, n_2=10,10	.000001	.000010	.000010
n_1, n_2=20,20	.000001	.000006	.000006
n_1, n_2=40,40	.000001	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001

Table 19. Comparative Type I Error Rates for Equal Size Samples at .01 δ

	α =.001 and .01 δ		
	Schuirmann	**Anderson & Hauk**	**Patel & Gupta**
Normal			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000003	.000003	.000003
n_1, n_2=60,60	.000001	.000001	.000001
Smooth Symmetric			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000000
n_1, n_2=40,40	.000000	.000003	.000000
n_1, n_2=60,60	.000000	.000001	.000000
Extreme Asymmetry			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000000	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001
	α =.01 and .01 δ		
Normal			
n_1, n_2=10,10	.000000	.000000	.000010
n_1, n_2=20,20	.000005	.000000	.000006
n_1, n_2=40,40	.000003	.000000	.000003
n_1, n_2=60,60	.000001	.000001	.000001
Smooth Symmetric			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000000	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001
Extreme Asymmetry			
n_1, n_2=10,10	.000000	.000010	.000010
n_1, n_2=20,20	.000000	.000006	.000006
n_1, n_2=40,40	.000000	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001
	α =.05 and .01 δ		
Normal			
n_1, n_2=10,10	.000009	.000010	.000010
n_1, n_2=20,20	.000005	.000000	.000006
n_1, n_2=40,40	.000002	.000000	.000003
n_1, n_2=60,60	.000001	.000001	.000001
Smooth Symmetric			
n_1, n_2=10,10	.000001	.000010	.000010
n_1, n_2=20,20	.000001	.000006	.000006
n_1, n_2=40,40	.000001	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001
Extreme Asymmetry			
n_1, n_2=10,10	.000001	.000010	.000010
n_1, n_2=20,20	.000001	.000006	.000006
n_1, n_2=40,40	.000001	.000003	.000003
n_1, n_2=60,60	.000000	.000001	.000001

Table 20. Comparative Type I Error Rates for Un-Equal Size Samples at .005 δ

	Sample Size (n)	Schuirmann	Anderson & Hauk	Patel-Gupta
Normal (α=.001 and .005)	n_1, n_2=10, 20	.000000	.000000	.000000
	n_1, n_2=10, 40	.000000	.000000	.000000
	n_1, n_2=10, 60	.000000	.000000	.000000
	n_1, n_2=20, 40	.000000	.000000	.000005
	n_1, n_2=20, 60	.000004	.000004	.000004
	n_1, n_2=40, 60	.000002	.000002	.000002
Smooth Symmetric	n_1, n_2=10, 20	.000000	.000009	.000009
	n_1, n_2=10, 40	.000000	.000008	.000008
	n_1, n_2=10, 60	.000000	.000007	.000007
	n_1, n_2=20, 40	.000000	.000005	.000005
	n_1, n_2=20, 60	.000000	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002
Extreme Asymmetry	n_1, n_2=10, 20	.000000	.000009	.000009
	n_1, n_2=10, 40	.000000	.000008	.000008
	n_1, n_2=10, 60	.000000	.000007	.000007
	n_1, n_2=20, 40	.000000	.000005	.000005
	n_1, n_2=20, 60	.000000	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002
Normal **(α=.01 and .005)**	n_1, n_2=10, 20	.000000	.000009	.000009
	n_1, n_2=10, 40	.000000	.000008	.000008
	n_1, n_2=10, 60	.000005	.000007	.000007
	n_1, n_2=20, 40	.000005	.000005	.000005
	n_1, n_2=20, 60	.000004	.000004	.000004
	n_1, n_2=40, 60	.000002	.000002	.000002
Smooth Symmetric	n_1, n_2=10, 20	.000000	.000009	.000009
	n_1, n_2=10, 40	.000000	.000008	.000008
	n_1, n_2=10, 60	.000000	.000007	.000007
	n_1, n_2=20, 40	.000000	.000005	.000005
	n_1, n_2=20, 60	.000000	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002
Extreme Asymmetry	n_1, n_2=10, 20	.000000	.000009	.000009
	n_1, n_2=10, 40	.000000	.000008	.000008
	n_1, n_2=10, 60	.000000	.000007	.000007
	n_1, n_2=20, 40	.000000	.000005	.000005
	n_1, n_2=20, 60	.000000	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002
Normal **(α=.05 and .01)**	n_1, n_2=10, 20	.000008	.000009	.000009
	n_1, n_2=10, 40	.000007	.000008	.000008
	n_1, n_2=10, 60	.000006	.000007	.000007
	n_1, n_2=20, 40	.000004	.000005	.000005
	n_1, n_2=20, 60	.000003	.000004	.000004
	n_1, n_2=40, 60	.000002	.000002	.000002
Smooth Symmetric	n_1, n_2=10, 20	.000001	.000009	.000009
	n_1, n_2=10, 40	.000001	.000008	.000008
	n_1, n_2=10, 60	.000001	.000007	.000007
	n_1, n_2=20, 40	.000001	.000005	.000005
	n_1, n_2=20, 60	.000001	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002
Extreme Asymmetry	n_1, n_2=10, 20	.000001	.000009	.000009
	n_1, n_2=10, 40	.000001	.000008	.000008
	n_1, n_2=10, 60	.000001	.000007	.000007
	n_1, n_2=20, 40	.000001	.000005	.000005
	n_1, n_2=20, 60	.000001	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002

Table 21. Comparative Type I Error Rates for Un-Equal Size Samples at .01 δ

	Sample Size (n)	Schuirmann	Anderson & Hauk	Patel & Gupta
Normal (α=.001 and .01)	n_1, n_2=10, 20	.000000	.000000	.000000
	n_1, n_2=10, 40	.000000	.000000	.000000
	n_1, n_2=10, 60	.000000	.000000	.000000
	n_1, n_2=20, 40	.000000	.000000	.000005
	n_1, n_2=20, 60	.000004	.000004	.000004
	n_1, n_2=40, 60	.000002	.000002	.000002
Smooth Symmetric	n_1, n_2=10, 20	.000000	.000009	.000009
	n_1, n_2=10, 40	.000000	.000008	.000008
	n_1, n_2=10, 60	.000000	.000007	.000007
	n_1, n_2=20, 40	.000000	.000005	.000005
	n_1, n_2=20, 60	.000000	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002
Extreme Asymmetry	n_1, n_2=10, 20	.000000	.000009	.000009
	n_1, n_2=10, 40	.000000	.000008	.000008
	n_1, n_2=10, 60	.000000	.000007	.000007
	n_1, n_2=20, 40	.000000	.000005	.000005
	n_1, n_2=20, 60	.000000	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002
Normal (α=.01 and .01)	n_1, n_2=10, 20	.000000	.000009	.000009
	n_1, n_2=10, 40	.000000	.000008	.000008
	n_1, n_2=10, 60	.000005	.000007	.000007
	n_1, n_2=20, 40	.000005	.000005	.000005
	n_1, n_2=20, 60	.000004	.000004	.000004
	n_1, n_2=40, 60	.000002	.000002	.000002
Smooth Symmetric	n_1, n_2=10, 20	.000000	.000009	.000009
	n_1, n_2=10, 40	.000000	.000008	.000008
	n_1, n_2=10, 60	.000000	.000007	.000007
	n_1, n_2=20, 40	.000000	000005	.000005
	n_1, n_2=20, 60	.000000	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002
Extreme Asymmetry	n_1, n_2=10, 20	.000000	.000009	.000009
	n_1, n_2=10, 40	.000000	.000008	.000008
	n_1, n_2=10, 60	.000000	.000007	.000007
	n_1, n_2=20, 40	.000000	.000005	.000005
	n_1, n_2=20, 60	.000000	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002
Normal (α=.05 and .01)	n_1, n_2=10, 20	.000008	.000009	.000009
	n_1, n_2=10, 40	.000007	.000000	.000008
	n_1, n_2=10, 60	.000006	.000007	.000007
	n_1, n_2=20, 40	.000004	.000005	.000005
	n_1, n_2=20, 60	.000003	.000004	.000004
	n_1, n_2=40, 60	.000002	.000002	.000002
Smooth Symmetric	n_1, n_2=10, 20	.000001	.000009	.000009
	n_1, n_2=10, 40	.000001	.000008	.000008
	n_1, n_2=10, 60	.000001	.000007	.000007
	n_1, n_2=20, 40	.000001	.000005	.000005
	n_1, n_2=20, 60	.000001	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002
Extreme Asymmetry	n_1, n_2=10, 20	.000001	.000009	.000009
	n_1, n_2=10, 40	.000001	.000008	.000008
	n_1, n_2=10, 60	.000001	.000007	.000007
	n_1, n_2=20, 40	.000001	.000005	.000005
	n_1, n_2=20, 60	.000001	.000004	.000004
	n_1, n_2=40, 60	.000000	.000002	.000002

REFERENCES

Alkhadher, O., Clarke, D. D., & Anderson, N. (1998). Equivalence and predictive validity of paper-and-pencil and computerized adaptive formats of the differential aptitude tests. *Journal of Occupational and Organizational Psychology, 71*, 205-217.

Anderson, S.A., & Hauck, W.W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics: Theory and Methods*, 12, 2663–2692

Barker, L.E., Luman, E.T., McCauley, M.M, & Chu, S.Y. (2002). Assessing Equivalence: An Alternative to the Use of Difference Tests for Measuring Disparities in Vaccination Coverage. *American Journal of Epidemiology*, Vol. 156, No. 11

Baugh, F., & Thompson, B. (2001). Using effect sizes in social science research: New APA and journal mandates for improved methodology practices. *Journal of Research in Education*, 11, 120–129

Cannon, D.S., Bell, W.E., Fowler, D.R., Penk, W.E., & Finkelstein, A.S. (1990). MMPI differences between alcoholics and drug abusers: Effects of age and race. Psychological Assessment: *A Journal of Clinical and Consulting Psychology*, 2, 51–55.

Brown, L.D., Hwang, J.T.G., & Munk, A. (1997). An unbiased test for the bioequivalence problem. *The Annals of Statistics*, Vol. 25, No. 6, 2345{2367.

Green, W.L., Concato, J., & Feinstein, A.R. (2000). Claims of Equivalence in Medical Research: Are They Supported by the Evidence? *Ann Intern Med.* 2000; 132:715-722.

Gruman, J. A., Cribbie, R. A., & Arpin-Cribbie, C. A. (2007). The Effects of Heteroscedasticity on Tests of Equivalence. *Journal of Modern Applied Statistical Methods*, May, 2007, Vol. 6, No. 1, 133-140

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology, 60*, 1-10.

Dannenberg, O., Dett, H., & Munk, A. (1994). An extension of Welch's approximate t solution to comparative bioequivalence trials. *Biometrika, 81*, 91-101.

Epstein, J., Klinkenberg, W. D., Wiley, D., & McKinley, L. (2001). Insuring sample equivalence across internet and paper-and-pencil assessments. *Computers in Human Behavior, 17*, 339-346.

Frick, H. (1987). On level and power of Anderson and Hauck's procedure for testing equivalence in comparative bioavailability. *Communications in Statistics-Theory and Methods*, 16, 2771-2778.

Frick, H. (1991). On a special feature of the Patel-Gupta equivalence test. *Biometrical Journal*, 33, 221-224.

Greene, W. L., Concato, J., & Feinstein, A. R. (2000). Claims of equivalence in medical research: Are they supported by the evidence? *Annals of Internal Medicine, 132*, 715-722.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68*, 155-165.

Hauschke, D., Steinijans, V. W., & Hothorn, L. A. (1996). A note on Welch's approximate 't'-solution to bioequivalence assessment. *Biometrika, 83*, 236-237.

Hotelling, H., Bartky, W., Deming, W. E., Friedman, M., & Hoel, P. (1948). The teaching of statistics. *Annals of Mathematical Statistics, 19*, 95-115.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J.C. & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

Metzler, C.M. (1974). Bioavailability: a problem in equivalence. *Biometrics*, 30, 309-317.

Metzler, C.M. (1988). Statistical methods for deciding bioequivalence of formulations. In A. Yacobi & E. Halperin-Walega (Eds.), *Oral Substained Released Formulations: design and evaluation*. (pp.217-238). NY: Pergamon Press.

Metzler, C.M., & Huang, D.C. (1983). Statistical methods for bioavailability and bioequivalence. *Clinical Research Practices and Drug Regulation Affairs*, 1. 109-112.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Park, H.M. (2008). *Hypothesis Testing and Statistical Power of a Test*. Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University. http://www.indiana.edu/~statmath/stat/all/power/index.html.

Patel, H.L., & Gupta, G.D. (1984). A problem of equivalence in clinical trials. *Biometrical Journal*, 26, 471-474.

Patnaik, P.B. (1949). The non-central chi-square and F-distributions and their applications. *Biometrika*, 36, 202-232.

Rogers, J. L., Howard, K. I. & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*, 553-565.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110–114.

Sawilowsky, S.S. (1990). Nonparametric test of interaction in experimental design. *Review of Educational Research, 60*(1), 91-126.

Sawilowsky, S.S. (1993). Comments on using alternatives to normal theory statistics in social and behavioral science. *Canadian Psychology, 34(4),* 432-439.

Sawilowsky, S.S. & Blair, R.C. (1992). A more realistic look at the robustness and type II error properties of the t test departures from population normality. *Psychological Bulletin, 111*(2), 352-360.

Sawilowsky, S.S. & Fahoome, G.F. (2003). *Statistics through Monte Carlo Simulation with Fortran*. Michigan: JMASM, Inc.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*, 657-680.

Seaman, M. A. & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3*, 403-411.

Scheffe, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association, 65*, 1501-1508.

Serlin, R.C. (2000). Testing for robustness in Monte Carlo Studies. *Psychological Methods*, 5, 230-240.

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Education and Program Planning, 19*(3), 193-198.

Tyron, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371-386.

Welch, B. L. (1938). The significance of the difference between two means when population variances are unequal. *Biometrika, 29*, 350-362.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 37*, 589-594.

Wilcox, R.R. (1996). *Statistics for the Social Sciences*. San Diego: Academic Press.

Wilkinson, L., and the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604.

ABSTRACT

A COMPARISON OF THE EFFECTS OF NON-NORMAL DISTRIBUTIONS ON TESTS OF EQUIVALENCE

by

LINDA F. ELLINGTON

December 2011

Advisor: Dr. Shlomo Sawilowsky

Major: Evaluation and Research

Degree: Doctor of Philosophy

Statistical theory and its application provide the foundation to modern systematic inquiry in the behavioral, physical and social sciences disciplines (Fisher, 1958; Wilcox, 1996). It provides the tools for scholars and researchers to operationalize constructs, describe populations, and measure and interpret the relations between populations and variables (Weinbach & Grinnell, 1997; Wilcox, 1996). Given that the majority of real data analysis in the behavioral and social sciences is comprised of non-normally distributed data, it is important that researchers be aware of the effects of non-normal distributions on the probability of detecting equivalence between populations.

The present study examined the effects and management of non-normally distributed data on equivalency tests under varied conditions for a two-sample design; and compared the properties of showing equivalence between populations at the smallest effect sizes. The findings for this report indicated that under conditions where sample sets were non-normally distributed, the differences in the statistical properties of the three equivalency tests became

most pronounced at the lowest nominal α =.001 for small to medium sample sizes. Optimal performance in relation to detecting equivalence occurred for the nominal α =.001, for small sample sizes $n_1 n_2$ (10, 10; 10, 20) under the Smooth Symmetric and Extreme Asymmetry distributions. Overall, all three tests demonstrated low power due to the relatively small sample size combinations paired with small effect sizes, and failed to control Type I error. Based on the findings of this study, none of the three tests were recommended as superior to the other.

AUTOBIOGRAPHICAL STATEMENT

LINDA F. ELLINGTON

WAYNE STATE UNIVERSITY– DETROIT, MI 2008 to Present
CENTER FOR URBAN STUDIES

Project Manager/Research Analyst

- Report findings in summary and formal reports to Federal, State and local agencies.
    - Develop summary of indicators findings for the purposes of federal reporting.
    - Develop annual comprehensive reports with analyses of indicators with demographic information.
    - Develop summary of indicator findings at local and district level for public reporting.
- Design and analyze complex short term and long term studies relevant to data-based program improvement:
    - Building local and state capacity for data analysis and use.
    - Linkages between quality practices and outcomes measurement results.
    - Components in outcomes measurement systems.
- Manage process and summative evaluation activities:
    - Supervise construction and maintenance of databases (electronic and web-based)
    - Conduct longitudinal and panel analyses comparing key indicator outcomes
    - Conduct appropriate statistical analyses and interpret findings for Federal reporting.

SCHOOL OF MEDICINE 2001 to 2008

Research Assistant

- Managed planning and implementation of online evaluation and assessment programs:
    - Consulted with faculty to develop online learning and assessment modules
    - Recommended policy and procedural changes impacting distance learning programs
    - Prepared statistical reports mapping student outcomes and delivery of instruction

OFFICE FOR TEACHING AND LEARNING 1999 to 2001

Graduate Research Assistant

- Delivered research and analytical support to tenure-track faculty:
    - Consulted on research and data collection methodologies
    - Developed evaluation and assessment plans for on-line programs

WAYNE STATE UNIVERSITY–DETROIT, MI
Doctor in Philosophy in Evaluation and Research (PhD), 2011
Masters in Human Performance Technology (M.ED), 2001

UNIVERSITY OF MICHIGAN– ANN ARBOR, MI
Bachelors in Psychology (B.G.S), 1990