INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

THE COMPARATIVE POWER OF THE INDEPENDENT-SAMPLES T-TEST AND WILCOXON RANK SUM TEST IN NON NORMAL DISTRIBUTIONS OF REAL DATA SETS IN EDUCATION AND PSYCHOLOGY

by

PATRICK DAVID BRIDGE

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

1996

MAJOR: EVALUATION AND RESEARCH

Approved by:

Shlomo Sawilowsky 1/17/96
Advisor Date

Richard C. Gallagher

Barry S. Markman

James L. Mackey

TABLE OF CONTENTS

DEDICATION

This dissertation is dedicated to my father David L. Bridge, my mother Pauline Ann Diroff and to my wife Tana J. Bridge. It was their support throughout the years that made this dissertation and other life accomplishments possible.

LIST OF FIGURES

CHAPTER ONE

Introduction

The foundation on which many parametric statistical tests are derived assumes that variables sampled are normally distributed and that they are independently and identically distributed (e.g., Zumbo & Zimmerman, 1993a). Yet, statisticians and researchers have questioned the frequency of occurrence of normally distributed data in real world problems (e.g., Pearson, 1895; Geary, 1947; Nunnally, 1978; Micceri, 1989; Pearson & Please, 1975; Tan, 1982). This, in turn, has generated controversy in the robustness and comparative power literature, and has directly influenced past and present applications of parametric and nonparametric statistical tests.

The independent samples t-test and its nonparametric counterpart, the Wilcoxon Rank Sum test, have been in the forefront of this controversy. The t-test is derived under the assumption of normality and is the Uniformally Most Powerful Unbiased test (UMPU) under normal curve theory. But, does it maintain its robustness and power properties when normal theory distribution assumptions are violated? Monte Carlo studies have been the traditional method in testing the robustness and comparative power properties of statistical tests. Prior to the middle 1980s, Monte Carlo studies for small samples comparing these two tests have been minimal. In

fact, Blair & Higgins (1985) pointed out "the researcher who searches the literature for guidance in choosing the more powerful/efficient of these two statistics in nonideal (i.e., applied) situations is almost certain to be disappointed or confused" (p.120). Since the middle 1980s there has been an increase in small samples Monte Carlo studies comparing these two tests. However, it is important to recognize that most small samples Monte Carlo studies have generally been restricted to mathematically "tame" or convenient distributions, (e.g., Micceri, 1989). Sawilowsky and Blair (1992), however, used real world data sets from education and psychology as the referent distribution. Examining the robustness and power of these two tests in real world distributions will aid educators and psychologists in obtaining statistical efficiency and accuracy in research practice.

The Purpose of this Study

The purpose of this study is to use Monte Carlo techniques in comparing the small samples power properties of the independent samples t-test to that of its nonparametric counterpart, the Wilcoxon Rank Sum test to violations of normality. The distributions include eight of the most prevalent shapes from the fields of education and psychology which are identified in a study by Micceri (1989), ranging from smooth symmetric to extreme asymmetry.

Definitions

In the stated purpose of the study, the following terms

are defined.

Monte Carlo Studies

Monte Carlo studies are computer simulations measuring the mathematical properties of statistical tests to violations of underlying assumptions (Harwell, 1990).

Power

According to Bradley (1968) "the power of a test is the probability of its rejecting a specified false null hypothesis" (p.56). Power is defined 1-ß, where ß is defined as Beta error or Type II error (Cohen, 1988). For example, 1-ß, where ß=.51, establishes a power level of .49.

Power Efficiency

Power efficiency is defined as the smallest sample size necessary to detect a true treatment difference and/or locate a false null hypothesis (Sawilowsky, 1990).

Relative Efficiency (RE)

The relative efficiency is a comparison of the power properties of two statistical tests. According to Bradley (1968) "the relative efficiency of test A with respect to test B is generally defined as b/a, where "a" is the number of observations required by test A to equal the power of test B based on "b" observations, when both statistic test the same null hypothesis against the same alternative hypothesis at the same significance levels" (p.57).

Asymptotic Relative Efficiency (ARE)

The Asymptotic Relative Efficiency (also known as the Pitman Efficiency) compares the relative efficiency of two

statistical tests with large samples and small treatment effects (Sawilowsky, 1990). Blair and Higgins (1985) defined ARE as the "limiting value of b/a as "a" is allowed to vary in such a way as to give Test A the same power as Test B while "b" approaches infinity and the treatment effect approaches zero" (p.120).

Robustness

Hunter and May (1993) defined robustness of a statistical test as "the extent that violating its assumptions does not appreciably affect the probability of its Type 1 error" (p.386). Sawilowsky (1990) stated, "the robustness is related not only to Type I error, but also to Type II error, the compliment of the power of a statistical test" (p.98).

The Research Problem and Relevance to Education and Psychology

In education and psychology the most common two sample test utilized to measure for shift in location is the t-test (Blair & Higgins, 1980). However, when the underlying assumptions are not met, the t-test may not be the best test. Scheffe' (1959) commented,

> The question of whether F-tests preserve against nonnormal alternatives, the power calculated under normal theory should not be confused with that of their efficiency against such alternatives relative to other kinds of tests (p.351).

In addition, previous small samples Monte Carlo studies comparing the power and robustness properties of the independent samples t-test and Wilcoxon Rank Sum test have been questioned. These comparisons not only lack the use of

real world data sets, but many of the outcomes from these studies are often conflicting or lack the necessary data to support the results. Micceri (1989), referring to past Monte Carlo studies commented:

> One disturbing finding of this research was a general lack of data availability. Only 25% of the authors to whom requests were sent reported the ability to produce simple frequency distributions for data reported in their studies. Many different reasons for this inability were noted; however, no matter what the reasons, the situation is somewhat disquieting (p.165).

Gibbons (1991), also discussing the availability of data in past studies, commented:

> "A problem with interpreting the results of some simulation studies reported in the literature is often the lack of details provided by the authors. These details are necessary for ensuring accuracy and, perhaps more important, for the purpose of future replication in order to obtain an independent assessment of the conclusions" (p.260).

Therefore, concerns arise as to the validity of past Monte Carlo studies, and the appropriate application of statistical tests in past and present research practices.

This study will investigate the relative power properties of the independent samples t-test to that of its nonparametric alternative, the Wilcoxon Rank Sum test to violations of normality. Using Monte Carlo techniques of repeated sampling from real data sets found in the field of education and psychology will assist statisticians and researchers in determining the validity of past studies and appropriate test applications for future studies.

Limitations of the Study

The limitations of the study is as follows.

1. The study is limited to the following real distributions identified by Micerri (1989): 1) Discrete mass at zero with gap (psychometric), 2) Mass at zero (achievement), 3) Extreme asymmetry (psychometric), 4) Extreme asymmetry (achievement), 5) Extreme bimodality (psychometric), 6) Multimodality and lumpy (achievement), 7) Digit preference (achievement), 8) Smooth Symmetric (achievement).
2. The study is limited to a sample size selection of (n1,n2)= (10,10), (5,15), (30,30) and (15,45).
3. This study is limited to addressing the normality issue, excluding other underlying assumptions such as homogeneous variances.
4. Treatments are not derived from real world data sets.
5. The real data sets include only eight of the most prevalent identified in the field of education and psychology and may differ from those in other disciplines.
6. This study is limited to the independent samples t-test and Wilcoxon Rank Sum test and does not compare the dependent samples t-test and Wilcoxon Sign Rank test.

CHAPTER TWO

THEORETICAL FOUNDATIONS AND LITERATURE REVIEW

Conceptualizing Parametric and Nonparametric Statistics

A fundamental purpose of statistical methods is to study variables. Understanding the characteristics of these variables is dependent on a distribution function which determines the probability of the observed variables having the same values as the population from which it was drawn. From these probability distribution characteristics, statistical models are derived. These statistical models, also known as parametric tests (e.g., t-test, F test) are based on sets of assumptions which traditionally thought, if violated, would alter the results. Although, nonparametric statistics (e.g., Pearson's Chi-Square, Wilcoxon test) also have underlying assumptions, these assumptions are considered less frequent and much weaker than their parametric counterpart (Siegal & Castellan, 1988). It is the violation of these distributional characteristics (specifically normality) which has generated controversy in the application of parametric and nonparametric statistics.

Gibbons (1985) defined parametric statistics as "those techniques that require data measured on at least an interval scale, and require specific population distribution assumptions, and relate to inferences concerning parameters" (p. 23). According to Hunter and May (1993, p.384) "the

assumptions underlying the best-known and most frequently used parametric statistics include:

1. All observations are randomly and independently sampled from their parent populations.
2. The population distributions from which samples are selected are normal.
3. All populations have the same variance.
4. The data are measured on at least an interval scale.

According to Gibbons (1985, p.23) the term nonparametric can be broken down in two categories.

1. Those inferences that are not concerned with the value of one or more parameters (termed nonparametric).
2. Those inferences whose validity does not rest on a specific probability model in the population (termed distribution free).

Both criteria are used interchangeably and can be labeled as nonparametric. As Mansfield (1986) stated, "The hallmark of these tests is that they avoid the assumption of normality" (p. 383).

Nonparametric statistics can be divided into three main groups: categorical, sign, or rank (Sawilowsky, 1990). Choosing the appropriate family of nonparametric tests is based on the type of data and the process in which the data is used. For example, nonparametric rank tests (e.g., Wilcoxon Sign-Rank and Rank Sum tests) have the ability to analyze original rank data (ordinal scale) in addition to transforming higher level variables (interval or ratio) to order or rank data. Whereas, parametric tests focus on the original values of the data and are restricted to interval/ratio scale

variables (Siegel & Castellan, 1988). Conover and Iman (1981) have proposed a rank transformation process for using parametric tests, although it has generated much criticism (for further references see, Sawilowsky et al., 1989; Blair et al., 1987; Thompson, 1991; Akritas, 1991).

Historically, the use of nonparametric statistics gained popularity in the 1950s, plateaued in the 1960s, and then began to steadily decline (Sawilowsky, 1990; Kelly, 1994). Sawilowsky (1990, p. 92) summarized three reasons for this decline.

> First, it is usually asserted that parametric statistics are extremely robust with respect to the assumption of population normality (Boneau, 1960; Box, 1954; Glass et al., 1972; Lindquist, 1953), precluding the need to consider alternative tests. Second, it is assumed that nonparametric tests are less powerful than their parametric counterparts (Kerlinger, 1964, 1973; Nunnally, 1975), apparently regardless of the shape of the population from which the data were sampled. Third, there has been a paucity of nonparametric tests for the more complicated research designs (Bradley, 1968).

As previously discussed, the robustness and comparative power properties of parametric and nonparametric tests to populations sampled from nonnormal distributions have been in the forefront of this controversy. The most prevalent example being the independent samples t-test and Wilcoxon Rank Sum test, and is the primary focus of this study. (For further information on nonparamteric tests for the more complicated research designs see Sawilowsky, 1990; Kelly, 1994).

Rank Test Theory and the Wilcoxon Rank Sum and Sign-Rank Tests

Rank tests are derived from the family of permutation tests

and were developed "to provide exact tests for wide (nonparametric) hypothesis, similar to those developed for parametric models in the small sample theory" (Hajek & Sidak, 1967, p. 11). A permutation can be defined as "a spatial arrangement or temporal sequence of objects or events with respect to one another" (Bradley, 1968, p. 63). Permutation tests have been considered extremely robust with respect to Type I error to departures from normality under shift in location parameters, although found impractical to use with having to calculate all possible permutations. Rank tests became a substitute for permutation tests, because they "maintain the properties of the parent permutation test in being nonparametric exact tests, and yet these procedures are often easy to compute" (Sawilowsky, 1990, p. 94). In addition, studies showing the high asymptotic efficiency of rank tests (e.g., Hodges & Lehman, 1956) increased their reputation and use (Hajek & Sidak, 1967).

According to Pratt and Gibbons (1981, p.249),

> Kruskal (1957) has traced the history of rank sum tests as far back as Deuchler (1914). However, the publications initiating the tests in modern terms are Wilcoxon (1945), Mann and Whitney (1947), and also Festinger (1946).

These tests are best known as the Wilcoxon Rank Sum, Mann-Whitney U test, or the Mann-Whitney-Wilcoxon test (Pratt and Gibbons, 1981). For the purpose of this study the Wilcoxon Rank Sum test (WRS) will be the test of reference.

The WRS test is a randomized two group design, where groups or observations are independently sampled and put in a single

array in ascending order within their respective group. A rank is then assigned to the groups or arrays with the smallest value given a 1, the next smallest a 2, etc. The ranks in each group are summed and the group with the smaller sum (x) is used to test for significance. The test criteria is based on the magnitude of (x) and sample size (n). This test is the nonparametric counterpart to the independent samples t-test and can be measured at an ordinal scale or when interval/ratio scale data are present. The WRS test can be considered one of the most powerful nonparametric tests (Runyon & Haber, 1991). Bradley (1968) commented, "In comparisons with other distribution-free statistics, the Wilcoxon test typically either ranks first or, when the set of tests being compared includes the optimum test for the conditions of the comparison, ranks a close second" (p.109-110).

Although not applicable to the analysis in this study it is important to conceptualize the dependent samples test. The Wilcoxon Sign-Rank Test (WSR) can be defined as a "nonparametric statistical test for ordinally scaled variables used with matched or correlated samples" (Runyon & Haber, 1991, p. 500). By having matched or correlated samples, the researcher reduces error by enhancing between group variance, allowing the isolation and identification of true treatment effects.

The WSR is the nonparametric counterpart to the paired samples t-test and can be broken down in two categories: sign and rank. The sign test also utilizes ordinally scaled

variables with matched or correlated samples and focuses on the direction of the difference between the pairs or groups of scores by assigning a positive or negative difference (Runyon & Haber, 1991). The differences are summed (x), critical values are established based on the sample size (n) and difference (x), and a decision on the null hypothesis is made. The WSR test finds the difference between initial scores, and ranks the differences (ignoring the sign) by assigning the smallest rank difference a 1, the next smallest a 2, etc. The signs are restored to the ranks and are summed and grouped according to their respective sign. The testing criteria is the sum of the smallest negative or positive paired values. The WSR test is more powerful than the simple sign test since it quantifies the magnitude of the actual differences between pairs of scores.

Measurement Issues

Rank tests as well as other nonparametric tests have been criticized by many authors who question their robustness and comparative power properties, claiming they lose valuable information in the transformation process from a higher order scale to a lower level scale. (For references see Sawilowsky, 1990, p. 94). As Meddis (1984) stated, "In certain circumstances the original values may contain crucial information with respect to the hypothesis being tested. Conversion to ranks may cause this information to be lost and a false research decision may be made" (p. 61).

There has also been an ongoing debate as to the relevance

of Stevens' (1946) scales of measurement (nominal, ordinal, interval, and ratio) as a determining factor in using parametric or nonparametric tests in testing hypothesis (Zumbo & Zimmerman, 1991). Many authors (e.g., Stevens, 1946; Siegel, 1956; Senders, 1958) viewed the scales of measurement as necessary criteria in determining the appropriate use of statistical tests. Others (e.g., Lord, 1953; Anderson, 1961; Gaito, 1986) find these scales as "irrelevant for the selection of statistics" (Zumbo & Zimmerman, 1991, p. 3). As Zumbo & Zimmerman (1991) stated, "In discussions of data analysis and significance testing in mathematical statistics, on the contrary, levels of measurement are not prominent" (p. 3). The relevant question for this study is whether transforming higher order variables (interval, ratio) to ranks (ordinal scale) changes the probability estimates (Baker, 1966). As Sawilowsky (1990) stated,

> The notion of converting higher level variables to a lower rank scale has concerned many text book authors. They view tests with fewer assumptions as testing more general, and hence, weaker hypothesis, and the ranking of interval and ratio scale measures as a process that throws away valuable information. A consequence that is assumed to occur by the deliberate reduction of information is that the rank tests lack statistical power (p. 94).

It is this statement questioning the power of rank tests (e.g., Wilcoxon Rank Sum test) which transforms data, as compared to parametric statistics that utilize original data (e.g., t-test, F test) that is the primary purpose of this study and will be reviewed in subsequent sections. For further review of the scales of measurement controversy, the reader is

referred to Stevens (1946, 1951); Siegel (1956); Senders (1958); Anderson (1961); Boneau (1961); Baker et al. (1966); Zumbo & Zimmerman (1991); and Sawilowsky (1993).

The t-test

The t-test can be defined as "a test statistic for determining the significance of a difference between means (for a two sample case)" (Runyon & Haber, 1991, p.337) and "whose distribution is equal to the square root of the F distribution with one degree of freedom in the numerator" (Sawilowsky, 1990, p. 99).

Gosset's (Student, 1908) t-test was developed at a time when the concept of normality in real world distributions was under scrutiny (Micceri, 1989). This test was designed with mandatory underlying assumptions. These assumptions include,

1. All observations are randomly and independently sampled from their parent populations.
2. The population distributions from which samples are selected are normal.
3. All populations have the same variance.
4. The data are measured on at least an interval scale (Hunter & May, 1993, p.384).

As previously stated, there has been an ongoing controversy concerning the robustness and power of the t-test to violations of these assumptions. When these assumptions are not met many authors state the t-test maintains acceptable robustness and power properties. Bradley (1978, p.145) summarized the views of various authors toward the robustness of the t-test and other parametric tests from population

normality.

> " 'this assumption may be violated almost with impunity ... may be safely ignored' (Hays, 1963, pp. 322, 380); 'nearly immune to violation ...invulnerability ... functionally a distribution-free test' (Boneau, 1960, pp. 50, 51, 60); 'assumption may be waived' (Dinham, 1976, p.174); 'assumption whose failure does not much matter' (Wright, 1976, p. 392); 'this remarkable property of robustness to nonnormality' (Box, 1953, p.318); remarkably robust' (Walker & Lev, 1969, p.286); amazingly insensitive...extremely gratifying' (Linquist, 1953, pp. 81, 86)".

In addition, the comparative power properties and robustness of the independent samples t-test and Wilcoxon Rank Sum test under the assumption of homogeneous variances has also been an on going debate in the literature (e.g., Zimmerman, 1987, 1991; Gibbons and Chakraborti, 1991, 1992). Subsequent sections will review the robustness and comparative power of the t-test to its nonparametric counterpart to violations from normality.

Normal Theory Statistics in Education and Psychology

Normal theory statistics such as the t-test, analysis of variance, and regression are dependent upon the probabilities under the normal distribution. In education and psychology, normal theory tests have been used more extensively than nonparametric statistics (Harwell, 1990; Hunter & May, 1993; Miccerri, 1989). But, in these fields, how common are data found to be normally distributed? Zumbo and Zimmerman (1993a) commented,

> First, it has been documented extensively that strict normality (ie., the exact Gaussian without any deviations at all) seldomly occurs in practice. Also, everyone believes in the strict normal distribution: experimenters because they think it is a mathematical theorem and mathematicians because they think it is an

experimental fact (p.382).

According to Nunnally (1978), tests of mental ability are developed so test items have a positive correlation, therefore "average correlations as high as .40 would produce a distribution flatter than the normal distribution" (p.160). It has also been argued that many of the variables studied in psychology have distributions with long or heavier tails (e.g., Hoaglin, Mosteller, & Tukey, 1983; Zumbo and Zimmerman, 1993b). Micceri (1989) examined 440 psychometric and ability achievement measures and found that all of the data sets were nonnormal according to the Kolomogorov-Smirnov test for normality at .01 alpha level. Only 7% (31) had tail weights and symmetry similar to the normal distribution, and only 4.3% (19) were found to be relatively smooth.

Researchers in the disciplines of education and psychology have been slow in recognizing the frequency of normality in real world data sets. In turn, this has generated concerns as to the validity of past research. As discussed by Micceri (1989), "Today's literature suggest a trend toward distrust of normality; however, this attitude frequently bypasses psychometricians and educators" (p.156). The question becomes; when do we use alternatives to normal theory? Zumbo (1993b) commented,

> It is true that we may be losing a little efficiency, and perhaps gaining some, when we use alternatives to normal theory. That is, in applied data analytic settings we never know what is the true distribution. Therefore, we never know the unique optimal or best test (p.441).

Yet, even when distribution characteristics are known,

controversy still remains as to the most efficient test to use under various research conditions. As discussed by Anderson (1961), and Blair (1985), Sawilowsky (1990) summarizes, "robustness, power, and versatility are traditional areas of comparison between parametric and nonparametric tests" (p. 96).

Robustness

Rey (1983), referencing Kendall and Buckland (1981), describes robustness in the following manner.

> Many test procedures involving probability levels depend for their exactitude on assumptions concerning the generating mechanism, e.g. that the parent variation is Normal (Gaussian). If the inferences are little affected by departure from those assumptions, e.g. if the significance points of a test vary little if the population departs quite substantially from the Normality the test on the inferences are said to be robust. In a rather more general sense, a statistical procedure is described as robust if it is not very sensitive to departures from the assumptions on which it depends (p.1).

As previously discussed, distribution characteristics are frequently unknown, leading to inconclusive violations of the underlying assumption of the statistic. When these violations occur the distribution of the statistic may be altered, effecting the probability of Type I and Type II error rates, and leading to an incorrect research decision. As discussed by Glass, Peckham, and Sanders (1972), the issue is not the strict adherence to the violation of assumptions, but the effects on the validity of the probability statements.

The steady decline in the use of nonparametric statistics can be attributed to the belief in the robustness of

parametric statistics. Boneau (1960, p.50) commented, "There is however, evidence that the ordinary t and F tests are nearly immune to violation of assumptions or can easily be made so if precautions are taken (Pearson, 1931; Bartlett, 1935; Welch, 1937; Daniels, 1938, Quensel, 1947; Gayen, 1950a, 1950b; David & Johnson, 1951; Jorsnell, 1953; Box, 1954a, 1954b; Box & Anderson, 1955)". In addition, past studies on robustness have lacked agreed upon definitions and quantification, leaving robustness results open to subjective explanations. Bradley (1978) summarizes, "Not only is there no generally accepted, and therefore standard, quantitative definition of what constitutes robustness, but, worse, claims of robustness are rarely accompanied by any quantitative indication of what the claimer means by the term" (p. 145). The influence of these two factors has led to many unsubstantiated robustness studies being over-generalized and in some cases under generalized, therefore, neglecting influences of higher order interactions (e.g., sample size, distribution shapes, tail region) and negatively influencing past and present statistical practices.

According to Bradley (1978) "In order to provide a quantitative definition of robustness (of significance level) you would have to state for a given alpha value the range of p values for which the test would be regarded as robust (p.146)". Bradley further identifies a liberal and stringent definition of robustness. The liberal criterion can be defined as 0.5 alpha $\leq \pi \leq$ 1.5 alpha, therefore a nominal alpha level

=.05 would generate a p value range of .025 to .075 and for a one tailed test the range of the p values would be .0125 to .0375. The stringent definition of robustness is as follows; 0.9 alpha $\leq \pi \leq$ 1.1 alpha, thus a nominal alpha level = .05 would represent a p range of .045 to .055. For a one tailed test the range of p values would be .0225 to .0275. For the purpose of this study both definitions of robustness will be assessed.

Other influential studies such as Rider (1929); Cochran (1947); Hack (1958); Norton (cited in Linquist, 1953); Scheffe (1959); Boneau (1960); Hsu & Feldt (1969); Glass, Peckham, and Sanders (1972); Andrews et al. (1972); and Ito (1980), also contributed to the robustness literature favoring parametric statistics.

According to Glass, Peckham, and Sanders (1972), Bradley (1963, 1966) evoked controversy regarding the robustness of the t-test and F-test, arguing that under many nonnormal distributions with extreme tails (<.01), parametric tests did not perform adequately. Glass, Peckham, and Sanders (1972, p.526) responded,

> His point is well taken that it is risky to generalize the results of a few studies of alpha to any specific distribution and set of experimental conditions. However, we are unsympathetic to dramatizations of the lack of robustness of the ANOVA by appeal to small alpha's. Statements of significance at levels beyond .001 ought not be taken too literally since minor violations of assumptions could easily distort the nominal .001 level into the .002 level (a 100 percent distortion).

In further studies, Bradley (1977, 1978, 1980a, 1980b, 1980c,

1982) continued to discuss the prominence of mixed-normal distributions in real world settings in education, psychology, and other science disciplines (Sawilowsky, 1990). Bradley soon gathered support from other influential researchers and statisticians (e.g., Blair, 1980, 1981; Micceri, 1989; Still & White, 1981; and Tan, 1982). This widespread belief in normality identified the need to re-establish the robustness and comparative power properties of the t-test and Wilcoxon test, using real world data.

Until the study by Micceri (1989), there were few robustness studies that attempted to use real world data (e.g., Hill & Dixon, 1982; Stigler, 1977; Tapia & Thompson, 1978). Most of the robustness studies previously discussed have been limited to statistical comparisons using mathematical distributions, attempting to mirror real data (e.g., Cauchy, Chi-square, Exponential, Uniform). Sawilowsky and Blair (1992) referencing Type I error properties of the independent samples t-test, summarized the conclusions of these studies.

> The prevailing view seems to be that the independent-samples t test is reasonably robust, insofar as Type 1 errors are concerned, to non-Gaussian population shape so long as (a) sample sizes are equal or nearly so, (b) sample sizes are fairly large (Boneau, 1960 mentions sample sizes of 25-30), and (c) tests are two-tailed rather than one tailed (p.352).

Bradley (1968, 1977, 1982) objected to these conclusions, claiming that distribution shapes using real world data are much more extreme than those distributions used in past robustness studies (Sawilowsky and Blair, 1992).

Micceri (1989) in his study of 440 psychometric and ability achievement measures also refuted the results of these studies, claiming they were a poor representation of data found in education and psychology. Sawilowsky and Blair (1992), referring to Micerri's (1989) results, commented.

> Micceri (1989) noted little overlap in the types of distributions that occur with real-world data and those selected for study in the classical study by Boneau (1960) demonstrating the robustness of the t test to nonnormality. Micceri concluded that the robustness issue remains unresolved, because "almost none of these comparisons occur in real life (p. 97).

It is important to recognize that this does not invalidate past studies, but raises concerns to the validity of these studies when comparing them to real world data (Micerri, 1989).

Using the eight real distributions identified by Micceri (1989), Sawilowsky and Blair (1992) found the independent samples t-test to be robust (Type 1 error) to violations of normality in such distributions as smooth symmetric achievement, digit preference achievement, and multimodality and lumpy achievement. Many of the nonrobust results were found in distributions with extreme skews, with the exceptions being when the aforementioned t-test criteria was used. As Sawilowsky and Blair (1992) stated,

> The degree of nonrobustness seen in these instances was at times more severe than has been previously reported. Having said this, however, we must note that the results obtained from these distributions do not change, in any fundamental fashion, the conclusions reached on the basis of studies that focused on populations modeled by well-known mathematical functions. That is to say, this study showed the t-test to be reasonably robust under the

> conditions outlined in the introduction to this article: when sample sizes are equal or nearly so, sample sizes are fairly large (25-30), and tests are two-tailed rather than one tailed (p.359).

Further analysis on Type II error demonstrated the independent samples t-test to be reasonably robust with power rates comparable to those under normal theory. Although, the skewed distributions did demonstrate a little more power than obtained under the normal distribution.

Sawilowsky and Blair (1992) demonstrated the robustness of the independent samples t-test (Type I & Type II errors) under various conditions of real world data. Yet, readers should not be hurried to use the t-test when in fact a nonparametric counterpart may be a more powerful statistic. As Sawilowsky (1993) stated, "The real issue of the effects of nonnormality, as indicated by Sawilowsky and Blair (1992), is on the comparative power, not robustness of the t-test" (p. 432).

Errors in Hypothesis Testing and Power Analysis

When testing hypotheses there is danger in making two types of errors. The first and more recognized Type I error (alpha) occurs when the null hypothesis is rejected, when in fact it is true. The second, Type II error (beta) occurs when failing to reject a null hypothesis when it is actually false. Both are based on the inverse of the other. For example, as alpha becomes smaller (.05 to .01) beta inflates, increasing the probability of committing a Type II error. As alpha becomes larger (.05 to .10) beta decreases, increasing the probability of committing a Type I error. Reaching a delicate balance

between alpha and beta generates a need in understanding the "power" of a statistical test.

According to Bradley (1968) "the power of a test is the probability of its rejecting a specified false null hypothesis" (p. 56). Mathematically power is defined as P=1-ß, where ß is defined as Beta error or Type II error (Cohen, 1988). For example, 1-ß, where ß=.45, gives a power level of .55. Therefore, the research undertaken may not yield significant results 45% of the time when in fact there may be treatment differences among the groups.

The power of a statistical test relies upon three sets of criteria: the significance level (alpha), the sample size or the reliability of the sample results, and the effect size or the degree to which the null hypothesis is actually false (Cohen, 1988). Recognizing these parameters and establishing power levels prior to beginning a study is imperative. First, power reflects the ability to detect true treatment differences if they exist. Second, it allows others to replicate the study under the same conditions, increasing validity. Third, power analysis can eliminate the wasting of resources through the use of efficient statistical tests.

Power analysis has been a neglected methodological tool in research. For example, Cohen (1962) canvased a (1960) volume in the Journal of Abnormal and Social Psychology and found the average power levels to be .48. In other similar studies, Brewer (1972), Friedman et al. (1978), Sedlmeier and Gigerenzer (1989), echoed similar results. The failure to

recognize the need for power analysis has mystified many. Kraemer and Thiemann (1987) attribute the neglect to overtraining in dealing with significant levels and under training in the use of power. Cohen (1992) stated, "One possible reason for the continued neglect of statistical power analysis in research in the behavioral sciences is the inaccessibility of or difficulty with the standard material" (p. 155). With Cohens (1988, 1992) articles the failure to conduct a power analysis to determine a level of accuracy is no longer acceptable.

There is not a universally agreed upon power standard among researchers and statisticians. But, for the behavioral sciences many researchers are beginning to agree that a power level of .80 is sufficient, or a 4:1 ratio of Type II error to Type I error when alpha is set at .05. The need for higher power levels calls for extreme increases in sample size and associated research costs. Cohen (1988) comments, "The behavioral scientist must set desired power values as well as desired significance criteria on the basis of the consideration of the seriousness of the consequences of the two kinds of errors and the cost of obtaining data" (p. 56). As Sawilowsky (1990) stated, "The smaller the sample necessary to detect a treatment, the more efficient, or powerful, is the statistic" (p. 93).

Comparative Power of the t-test and Wilcoxon Test

When choosing the appropriate application of statistical tests, a researcher must consider the efficiency of these

tests. Hunter and May (1993) defined efficiency as,

> a relative term comparing the power of one test to another when both are used to test the same null hypothesis, and the relative efficiency of one test with respect to another is the ratio of the sample sizes needed for both tests to achieve the same power (p.385).

As previously discussed, the second reason for the decline in the use of nonparametric tests, is that "it is assumed that nonparametric tests are less powerful than their parametric counterparts, regardless of the shape of the population" (Sawilowsky, 1990, p. 92). Therefore, these tests are considered less efficient.

Early comparative power studies between the t-test and Wilcoxon test have shown that the t-test holds only modest power advantages under normality (e.g., Dixon, 1954; Hodges and Lehmann, 1956; Lehmann, 1975; Neave and Granger, 1968) Yet, under nonnormality, conditions under which nonparametric tests were designed, there are many authors who question the power of nonparametric statistics (e.g., Glass et al. 1978; Kerlinger 1964, 1973; Nunnally, 1975). As Blair (1981) stated,

> One might assume that because the t-test is the uniformly most powerful (UMP) unbiased test under normal theory, it will naturally be more powerful than other tests in the non-normal situation, provided that its normal theory power is preserved. But this is fallacious reasoning because the optimal power properties associated with the t-test under normal theory are no longer in force once the normality stipulation has been abandoned (p. 500).

Studies by Hoeffding (1952), and Lehman and Stein (1959) have shown that a nonparametric randomization test is as efficient as the t-test under normal conditions. But, the failure of researchers to recognize and utilize this nonparametric

distinction has questioned the validity of many studies.

Historically, there has been a variety of methods of measuring the comparative power properties of the t-test and Wilcoxon test. The asymptotic relative efficiency (A.R.E), also known as the Pitman Efficiency, compares tests under standardized conditions and is a generalized predictor of power when sample sizes are large. In theory, the further the A.R.E deviates from 1.00, the more powerful or less powerful the test is. Hodges and Lehmann (1956) compared the t-test and Wilcoxon test and found that when samples where drawn from a normally distributed population, "the asymptotic relative efficiency of the Wilcoxon test relative to the t-test can be as high as infinity, it can never be lower than .864" (Blair and Higgins, 1980, p.311). In addition, Blair (1981) comments,

> The A.R.E of the Wilcoxon test relative to the t-test is .995 under normality and homogeneity. This indicates that when power is compared by means of this method, the t-test shows only a slight advantage over the Wilcoxon test when the former statistic's assumptions are perfectly met (p. 500).

The properties of the A.R.E method for power comparisons are based on infinitely large sample sizes, criteria that are viewed as unrealistic in many research settings (Bradley, 1968). Boneau (1962), in a computer generated small samples (n=5) Monte Carlo study, questioned the asymptotic results of Hodges and Lehman (1956), claiming such results cannot be compared when sample sizes are finite. Further, he demonstrated that when samples are from a normal, rectangular,

or exponential distribution, the t-test is more powerful than the Mann-Whitney U test in nonnormal conditions.

In opposition, Blair, Higgins, and Smitley (1980), using computer simulations reexamined Boneau's (1962) study, focusing on the exponential distribution with a broader range of sample sizes (3,9), (6,6), (9,27), (18,18), (27,81), (54,54). Their results showed the Mann-Whitney U test having large power advantages over the t-test. In fact, for a one tail test with alpha at .05, the percent difference in the rejection of the null hypothesis was as high as 31% for equal sample sizes and as high as 33% for unequal sample sizes. For a two tail test the percent rejection was 31% and 37% respectively. In both one and two tail tests the Mann-Whitney U showed large power advantages when sample sizes were small. Blair, Higgins, and Smitley (1980) commented,

> The first (and most important) point to note is that Boneau (1962) investigated extremely small sample sizes. This is important for two reasons: (1) actual educational and psychological research usually involves much larger samples than those investigated by Boneau (1962), and (2) the asymptotic properties of a statistic may not be manifested in such small samples (Ramsey, 1971; Conover et al. 1978" (p. 115).

"Boneau (1962) apparently believed that his small sample results would be maintained when more realistic sample sizes were considered-an assumption that was simply not true" (Blair, Higgins, and Smitley, 1980, p.118).

Further studies by Dixon (1954); Chernoff and Savage (1958); Neave & Granger (1968); Lehman (1975); Randles and Wolfe (1979); and Blair & Higgins (1980, 1981), found that

under various nonnormal distributions (e.g., half normal, Laplace, exponential, mixed normal) the Wilcoxon test was more powerful than the t-test. For example, Blair & Higgins (1980) using computer generated Monte Carlo studies compared the power of the Wilcoxon Rank Sum test (WRS) to the one tailed independent samples t-test under a uniform, Laplace, half normal, exponential, mixed-normal, and mixed uniform distributions. Sample sizes of (3,9), (6,6), (9,27), (18,18), (27,81), and (54,54) were used, with alpha measured at .005, .010, .025, and .05. Tests of significance were carried out on both statistics and a reject or fail to reject response to the null hypothesis was recorded. Any rejection of the null was an indication of the power properties of the particular test. Power advantages were calculated by obtaining the proportion of null hypothesis rejected by the less powerful statistic and subtracting its sum from the proportion of rejections of the more powerful statistic (Blair and Higgins, 1980). The results were mixed, showing with small sample sizes (n1 + n2)=12, the t-test held moderate power advantages under certain distributions and sample sizes, but the WRS also maintained small samples power advantages in some situations. In contrast the WRS showed power advantages with most moderate sample sizes. Blair and Higgins (1980) commented,

> From the results outlined above, it can be concluded that the results obtained from small sample studies that compare the power of the two statistics in question do not, typically, generalize to situations involving samples of moderate sizes. In fact, conclusions reached on the basis of small sample studies are oftentimes in direct opposition to those reached on the basis of

moderate sample size studies (p.332).

Exceptions were the uniform distribution where the t-test obtained power superiority in both small and moderate sample sizes.

Even though the t-test held power advantages in many small sample sizes and some moderate sample sizes, the magnitude of these advantages were rather small. Whereas, when the WRS held power advantages, the magnitude was large. For example, in only 19% of the power comparisons the WRS never held an advantage, while 46% of the WRS power advantages exceeded .10 and 27% of the time the showed power advantages larger than .20. Comparably, 53% of the time the t-test never held an advantage, and only 8% of the time did the t-test exceed a power of .10. In addition, the t-test never showed power advantages greater than .20.

These studies demonstrated the comparative power of the t-test and its nonparametric alternative the Wilcoxon test under mathematical distributions. As previously discussed, Micceri (1989) found that these were a poor representation of data found in education and psychology. Sawilowsky (1992) conducted a Monte Carlo study using the real world data sets from Micceri's (1989) study and compared the independent samples t-test to the Wilcoxon Rank Sum. A Sample size of (5,15) was drawn from an extreme asymmetric (psychometric) distribution with alpha set at .05 and treatment effects of .20 and .50. Results showed the power of WRS with an effect size of .20 was .395, as compared to .139 for the t-test. When the effect size

was .50, the power advantage of the WSR increased to .723 compared to .495 for the t-test.

Sawilowsky (1992) demonstrated the power advantages of the WRS over the independent t-test under one extreme nonnormal distribution. As previously discussed, the purpose of this study is to analyze the comparative power properties of the independent samples t-test to that of the Wilcoxon Rank Sum test using eight real world data sets identified by Micceri (1989). In turn, this will expand the study by Sawilowsky (1992) and contribute new research to the field of education and psychology by comparing the power properties of these tests using these real world data sets.

CHAPTER THREE

METHODOLOGY

Micceri (1989) demonstrated the prevalence of nonnormal data in education and psychology. For researchers and statisticians in these fields, this insight has aided in generating "profession" specific information, and further identifies the need for additional robustness and comparative power studies. Historically, when conducting these studies, the methods, available resources, and continuing technological advances have varied, generating controversy as to the validity of many past studies. According to Harwell (1990),

> Whenever possible exact statistical theory is used to determine the mathematical properties a test will have when its underlying assumptions are not tenable. This is seldom possible since most exact statistical theory requires normality of the population distribution of scores, an assumption which educational data rarely satisfies (p.3).

Monte Carlo techniques is a useful method for analyzing the comparative power and robustness properties of statistical tests. Harwell (1990) defined Monte Carlo studies as computer simulations measuring the mathematical properties of statistical tests to violations of underlying assumptions. Until recently, the scope of this definition has been limited to measuring the properties of mathematical distributions. The failure to have readily agreed upon real data sets in education and psychology has restricted the application of this technique. Micceri's (1989) study identified multiple nonnormal distributions of real data, expanding the parameters

of Monte Carlo techniques. Today, these techniques can be defined as a computer generated method, sampling real or mathematical distributions with or without replacement, for the purpose of measuring the mathematical properties of statistical tests to violations of underlying assumptions. Monte Carlo study is conducted in the following manner.

> 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 specified relationship among variables (but which do not usually conform to the assumptions required for correct application of the test), the statistical test is computed for the data, and the value of the statistical test is recorded. The collection of values of the statistical test provide information on its properties (e.g., the proportion of "significant" values of the test). If the underlying assumptions 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; MC studies permit these characteristics to be examined when underlying assumptions are violated (Harwell, 1990, p.4).

Using a Gateway 2000 4DX2-66V and Microsoft Fortran 5.1 program, computer generated small samples Monte Carlo techniques were used to analyze the comparative power properties of the independent samples t-test to the Wilcoxon Rank Sum test. Based on the limitations in obtaining critical values for the Wilcoxon Rank Sum and the ease of computation, the Wilcoxon Rank Sum test will be conducted by applying the independent samples t-test on the ranks of the scores. Zimmerman and Zumbo (1990); Conover, (1980); Conover and Iman, (1976, 1981) and others, have shown that the Wilcoxon Rank Sum test on original scores is equivalent to conducting an

independent samples t-test on the ranks of those scores.

The observations were then randomly sampled with replacement, using the International Mathematical and Statistical Libraries (IMSL, 1987) RNSET, RNUND, RNKSM, and ERSET subroutines. Sample sizes (n1,n2)= (10,10), (15,25), and (30,30), were used, with nominal alpha set at .05. Eight treatment alternatives of varying sizes were analyzed for each distribution and sample size to measure shift in location parameters. The treatments were a function of a distributions standard deviation, multiplied by a constant of .25. A one-tailed independent samples t-test and Wilcoxon Rank Sum was then performed on the samples. Power advantages were obtained by comparing the actual alpha for each test statistic under each condition, to departures from nominal alpha and subtracting the result of the less powerful statistic by the more powerful statistic. Robustness of the two tests was also conducted under each distribution. Ten thousand repetitions were conducted for each distribution, sample size, and treatment effect interval.

The samples were taken from eight of the most prevalent educational and psychometric distributions identified in a study by Micceri (1989). Of the 440 distributions identified by Micceri (1989), 265 were derived from journal articles or research, 30 from national tests, 64 from statewide tests, 65 from districtwide tests, and 17 from the University of South Florida's GRE and college entrance files. There were four sets of separately sampled criteria, including; general

achievement/ ability tests, criterion/mastery tests, psychometric measures, and pretest and posttest measures (Micceri, 1989). The most prevalent of these distributions and their descriptive properties are as follows.

Normal Distribution

In order to compare Micceri's (1989) distributions to violations from normality, the characteristics of the normal distribution are presented below. The functional form of the normal (Gaussian) distribution is,

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

Skewness and kurtosis are common descriptive properties measuring violations from normality (Glass et al. 1978). Skewness can be defined as,

$$\sqrt{b_1} = \frac{E(X-\mu)^3}{\sigma^3}$$

In a normal distribution skewness = 0. A positive skew represents an elongated right tail. A negative skew depicts an elongated left tail.

Kurtosis represents the peak of the distribution and can be defined as:

$$b_2 = \frac{E(x-\mu)^4}{\sigma^4}$$

For a normal distribution, kurtosis = 3.00. As the value exceeds 3.00 the distribution demonstrates a larger peak (leptokurtic). As the value falls below 3.00 the flatter the distribution characteristics (playkurtic). For the purpose of this study, kurtosis has been scaled so a normal kurtosis =

3.00. Below is a summary of the eight distributions identified by Micceri (1989), and their descriptive properties (i.e., mean, median, standard deviation, skewness, kurtosis) which were summarized by Sawilowsky & Blair (1992) and used in this study.

Smooth Symmetric, Achievement

The smooth symmetric data set is an achievement measure with characteristics similar to digit preference with a light skew and variance in kurtosis from the normal distribution (see figure 1). This distribution has a mean = 13.91, median = 13.00, standard deviation = 4.91, skewness = 0.01, and kurtosis = 2.66. This distribution demonstrates an 11.3% variance (playkurtic) from normal kurtosis.

Digit Preference, Achievement

Digit preference is an achievement measure that also has not been considered in past robustness and comparative power studies, until Sawilowsky and Blair (1992). This distribution is relatively symmetric, with a mean = 536.95, median = 535, standard deviation = 37.64, skewness = -0.07, and kurtosis = 2.76 (see figure 2). This distribution demonstrates an 8% variance from normal kurtosis.

Mass at Zero, Achievement

Mass at zero is a relatively symmetric achievement measure with a small mass of scores (36 of 2429 = 1.48%) accumulating at zero (see figure 3). This distribution has a slight negative skew = -0.03 and kurtosis = 3.31. This distribution demonstrates a 10.33% variance (leptokurtic) from normal

kurtosis and has a mean = 12.92, median = 13.00, and a standard deviation = 4.42.

Multimodal and Lumpiness, Achievement

Multimodality and lumpy is an achievement data set that has not been considered in past robustness and comparative power studies, until Sawilowsky and Blair (1992). This distribution has a mean = 21.15, median = 18.00, standard deviation = 11.90,skewness = 0.19, and kurtosis = 1.80; demonstrating a 40% variance from normal kurtosis (see figure 4).

Extreme Bimodality, Psychometric

Extreme bimodality is a psychometric data set with a mean = 2.97, median = 4.00, standard deviation = 1.69, skewness = -0.08, and kurtosis = 1.30 (see figure 5). This distribution has a 57% variance (playkurtic) from normal kurtosis.

Extreme Asymmetry, Psychometric

The extreme asymmetry psychometric data set has characteristics similar to the discrete mass at zero with gap, psychometric measure, with a skew = 1.64 and 1.65 respectively (see figure 6). Although, with this data set, scores demonstrated 13.6% more kurtosis than the discrete mass at zero with gap, and 51% more kurtosis than the normal distribution (leptokurtic). This data set has a mean = 13.67, median = 11.00, standard deviation = 5.75 and kurtosis = 4.52.

Extreme Asymmetry, Achievement

The extreme asymmetry achievement data set has a skew = -1.33, which is opposite of the extreme asymmetry,

psychometric measure, having a skew = 1.64. This distribution has a mean = 24.5, median = 27.00, standard deviation = 5.79, and kurtosis = 4.11 (see figure 7). This distribution demonstrates a 37% variance (leptokurtic) from normal kurtosis.

Mass at Zero with Gap, Psychometric

Discrete mass at zero with gap is a psychometric measure with 80% of the scores (519 of 648) accumulating at zero, while the next interval of scores range between 8 and 11 (see figure 8). This data set has a mean = 1.85, median = 0, standard deviation = 3.80, skew = 1.65, and kurtosis = 3.98. This distribution has a 33% variance (leptokurtic) from normal kurtosis.

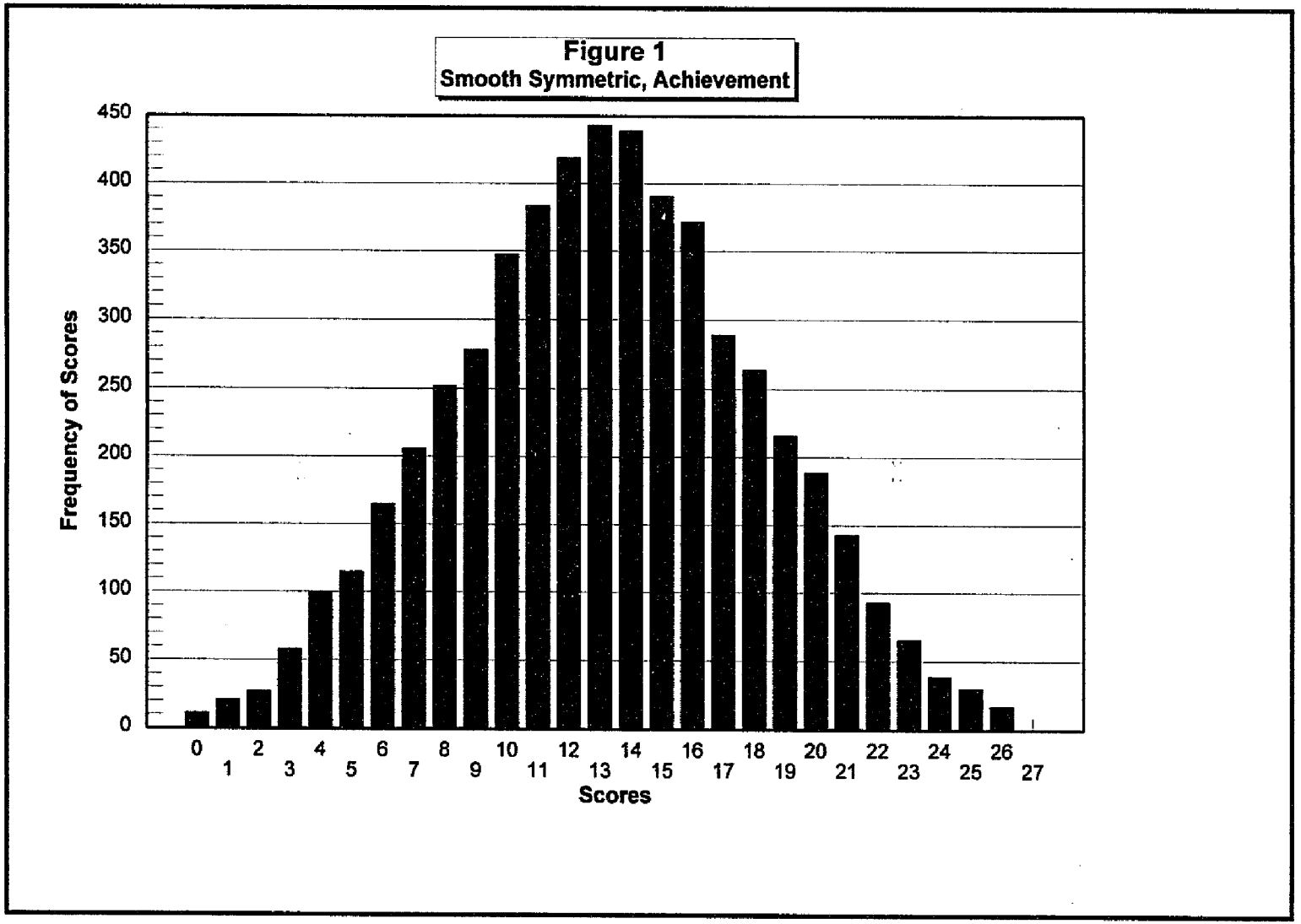


Figure 1
Smooth Symmetric, Achievement
Frequency of Scores
450
400
350
300
250
200
150
100
50
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
Scores

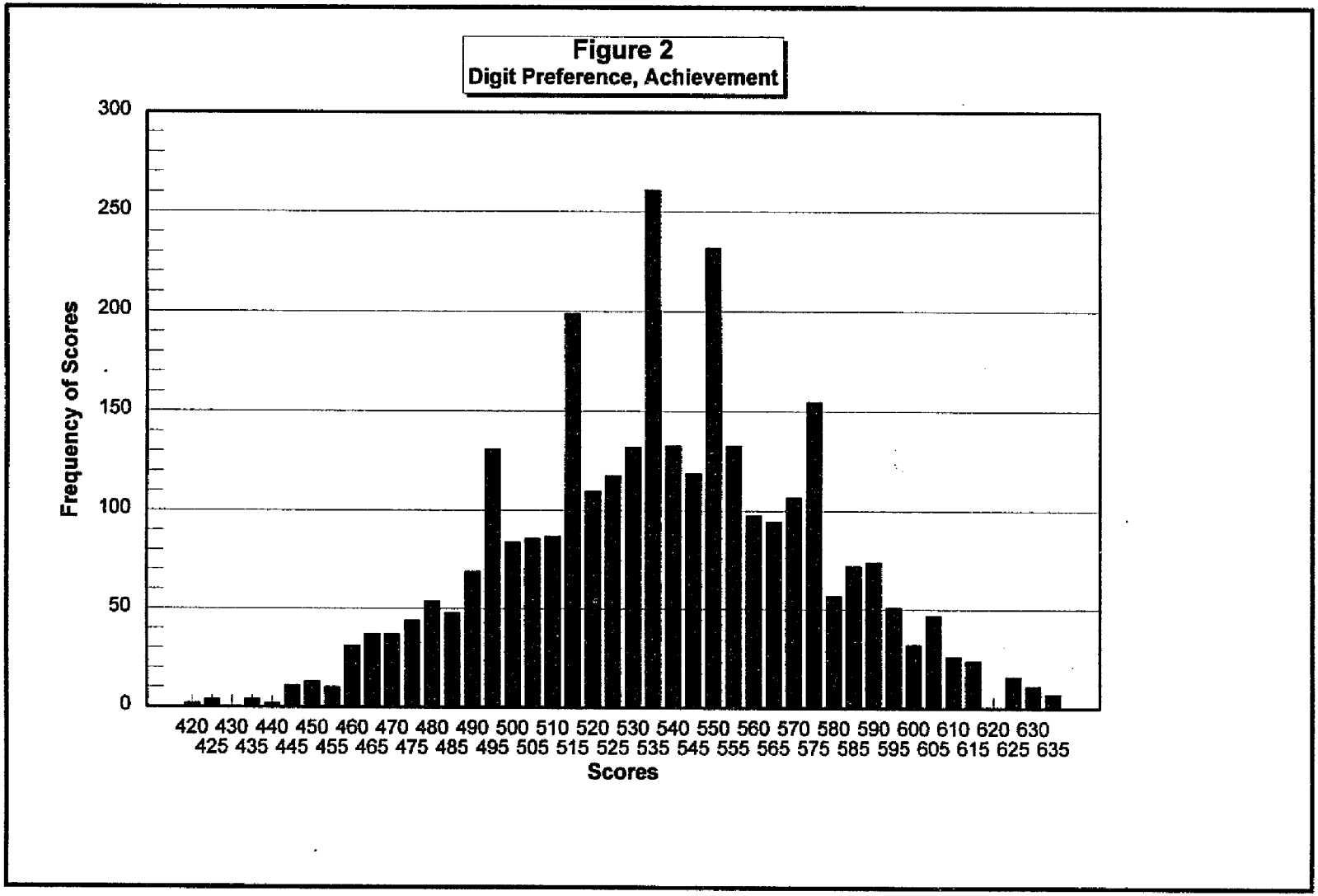


Figure 2
Digit Preference, Achievement
Frequency of Scores
300
250
200
150
100
50
0
420 430 440 450 460 470 480 490 500 510 520 530 540 550 560 570 580 590 600 610 620 630
425 435 445 455 465 475 485 495 505 515 525 535 545 555 565 575 585 595 605 615 625 635
Scores

Figure 3
Mass At Zero, Achievement

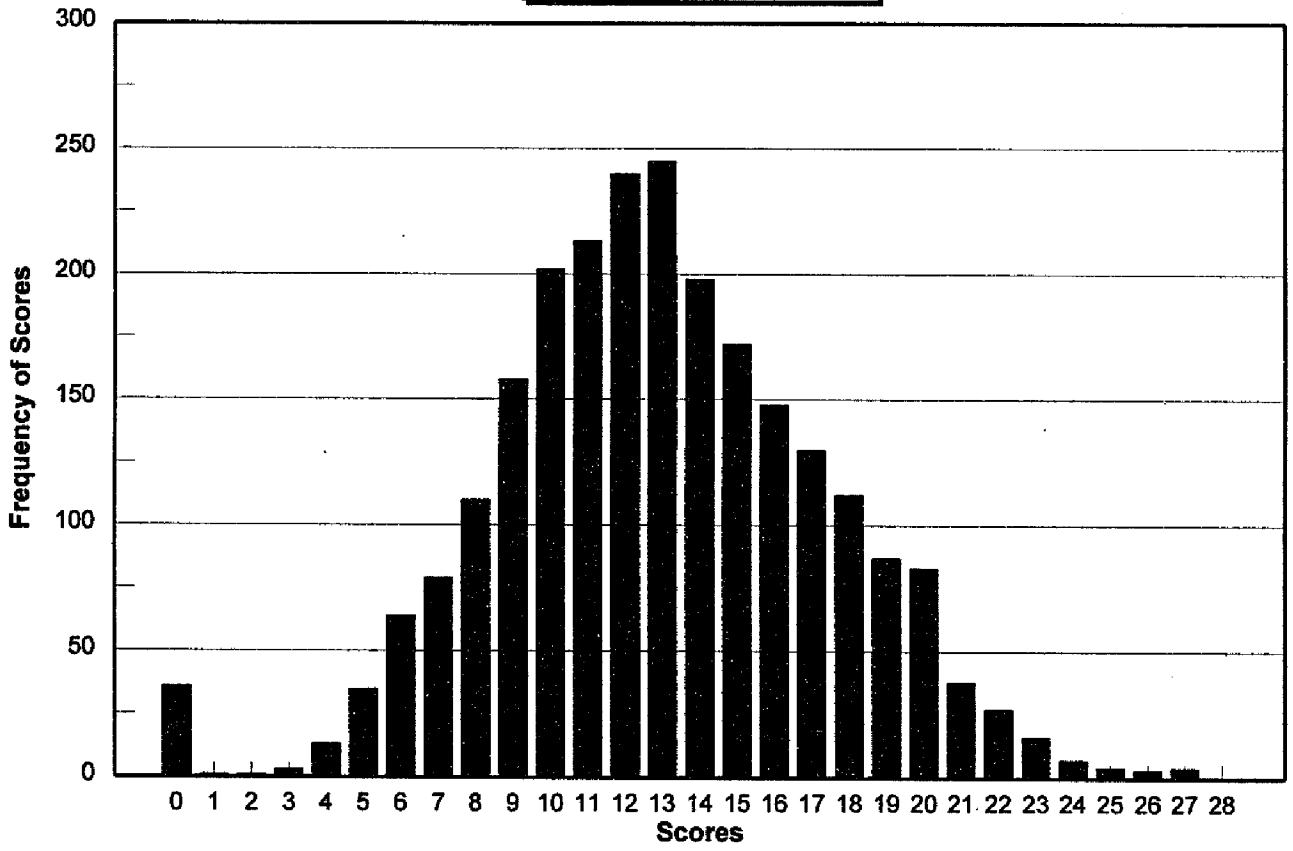
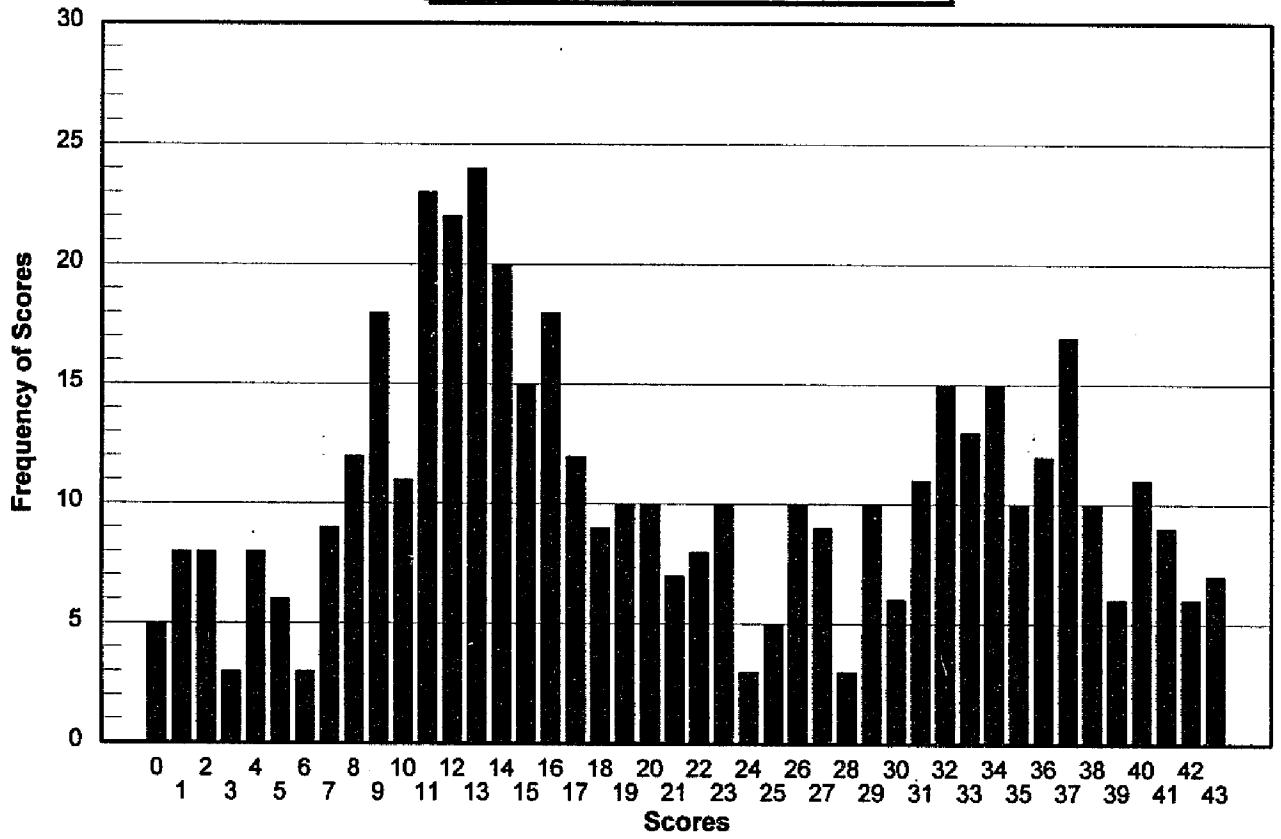


Figure 4
Multimodal and Lumpiness, Achievement



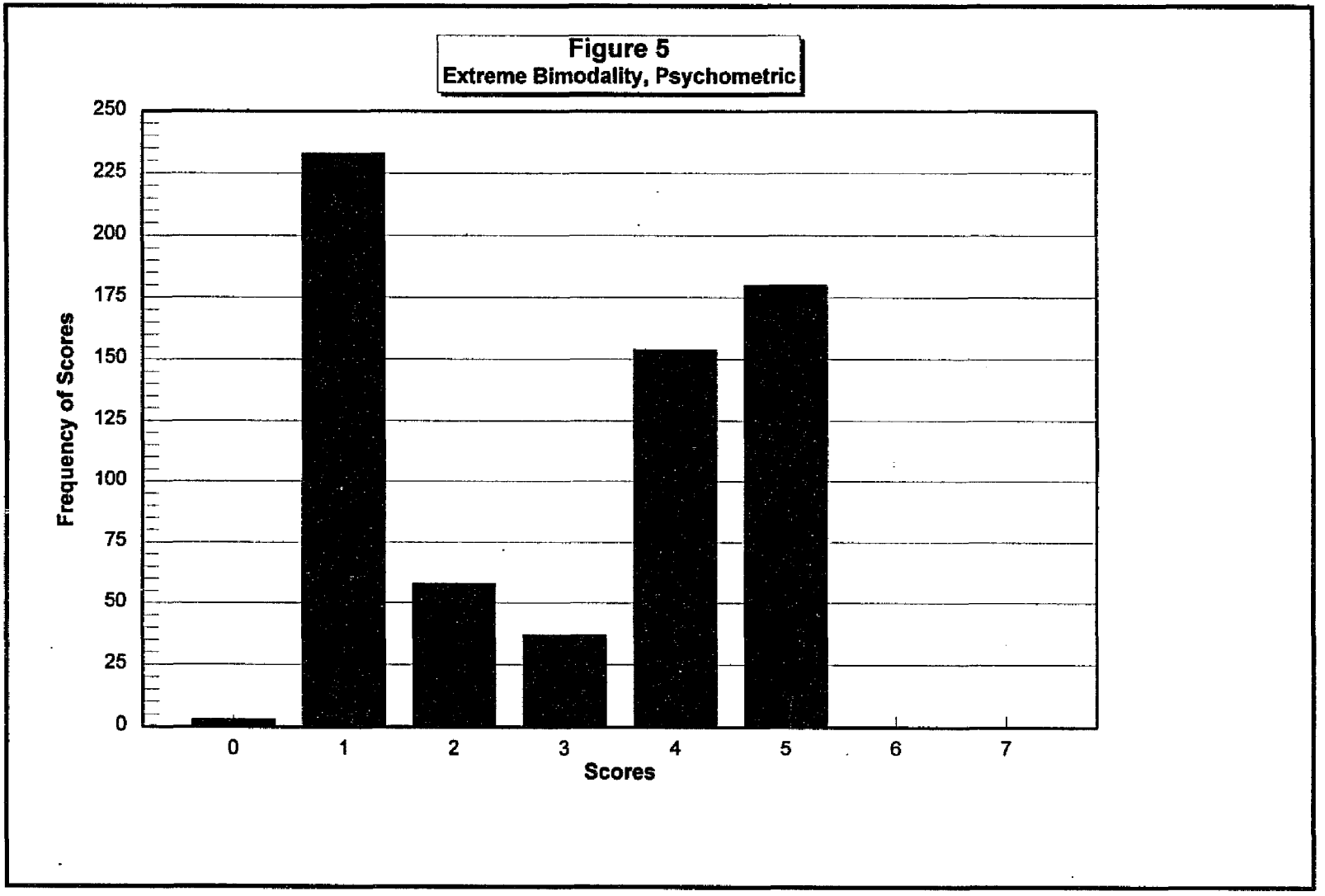


Figure 5
Extreme Bimodality, Psychometric
Frequency of Scores
250
225
200
175
150
125
100
75
50
25
0
0
1
2
3
4
5
6
7
Scores

Figure 6
Extreme Asymmetry, Psychometric

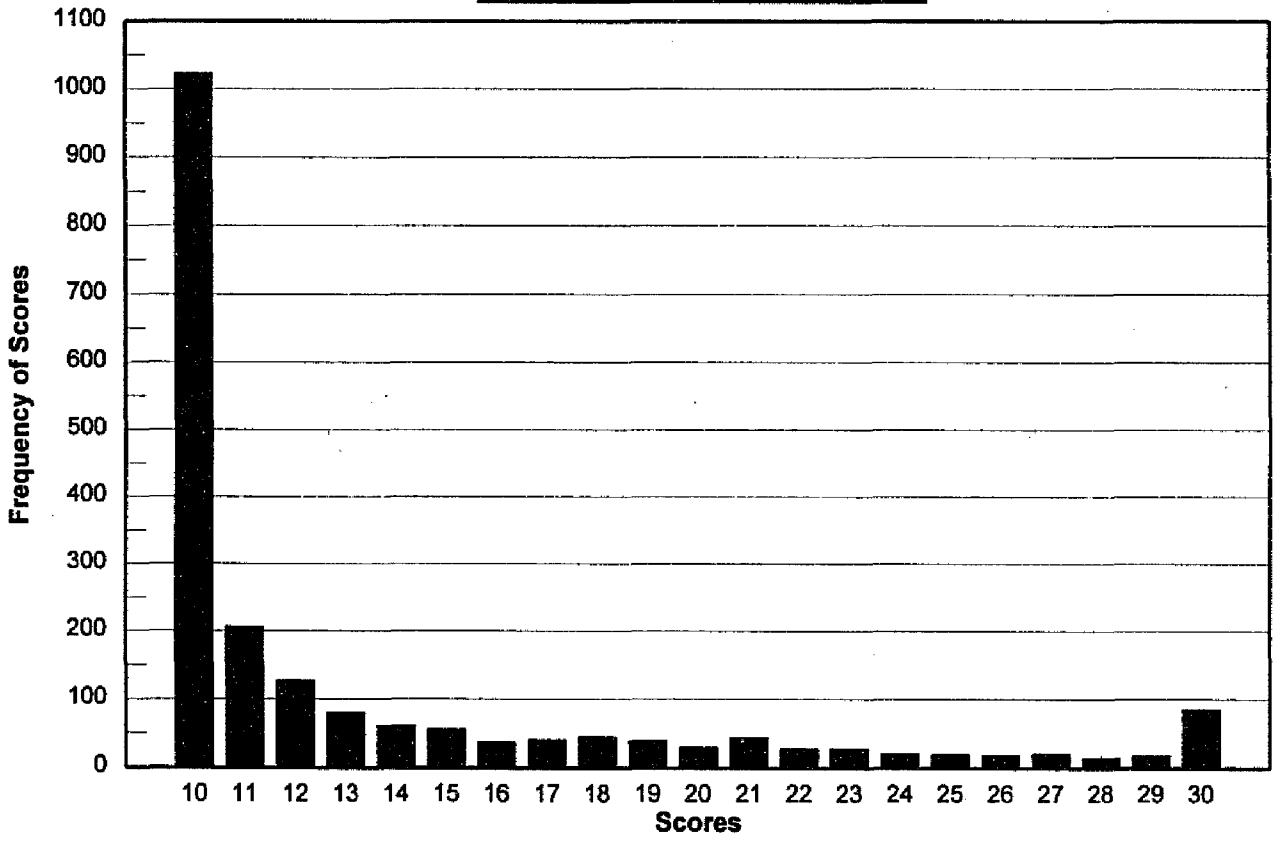
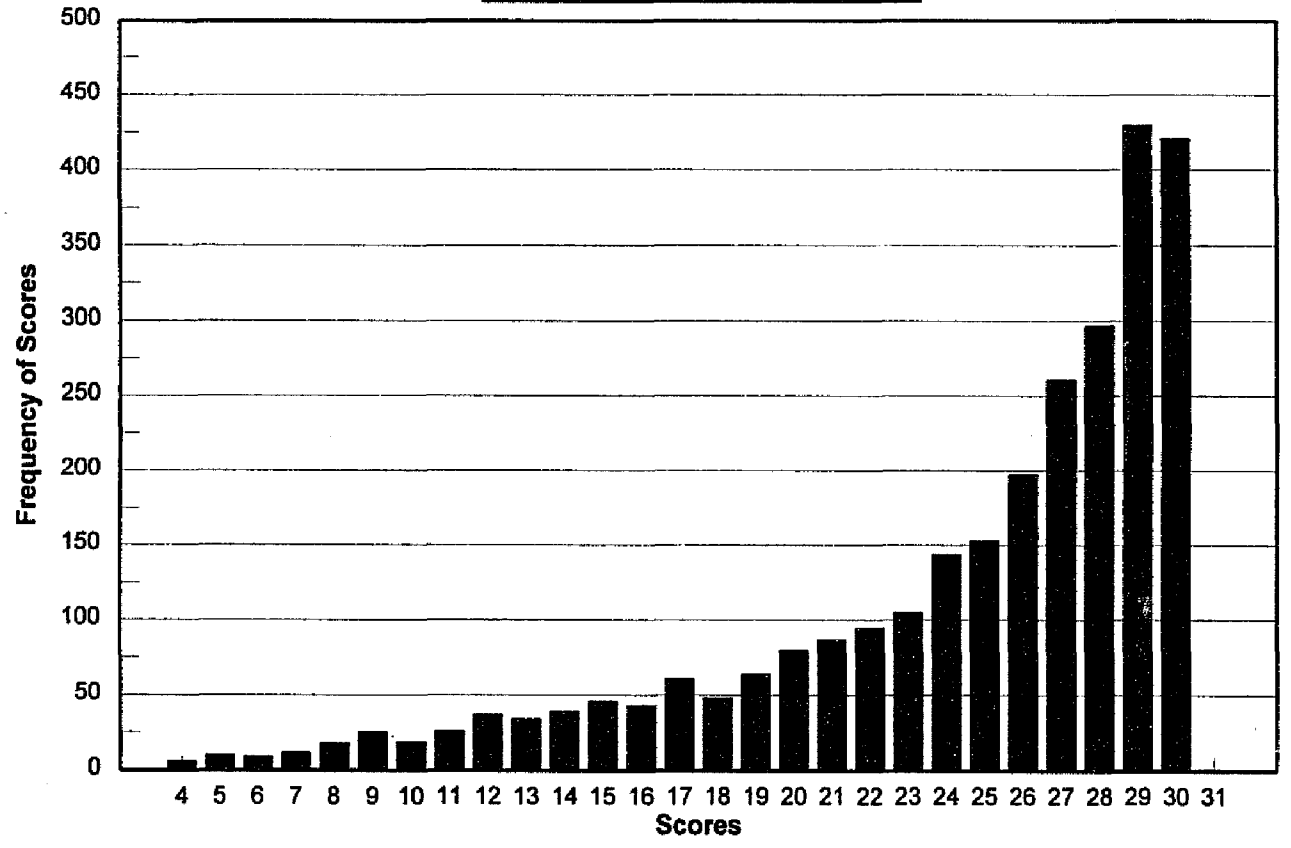


Figure 7
Extreme Asymmetry, Achievement



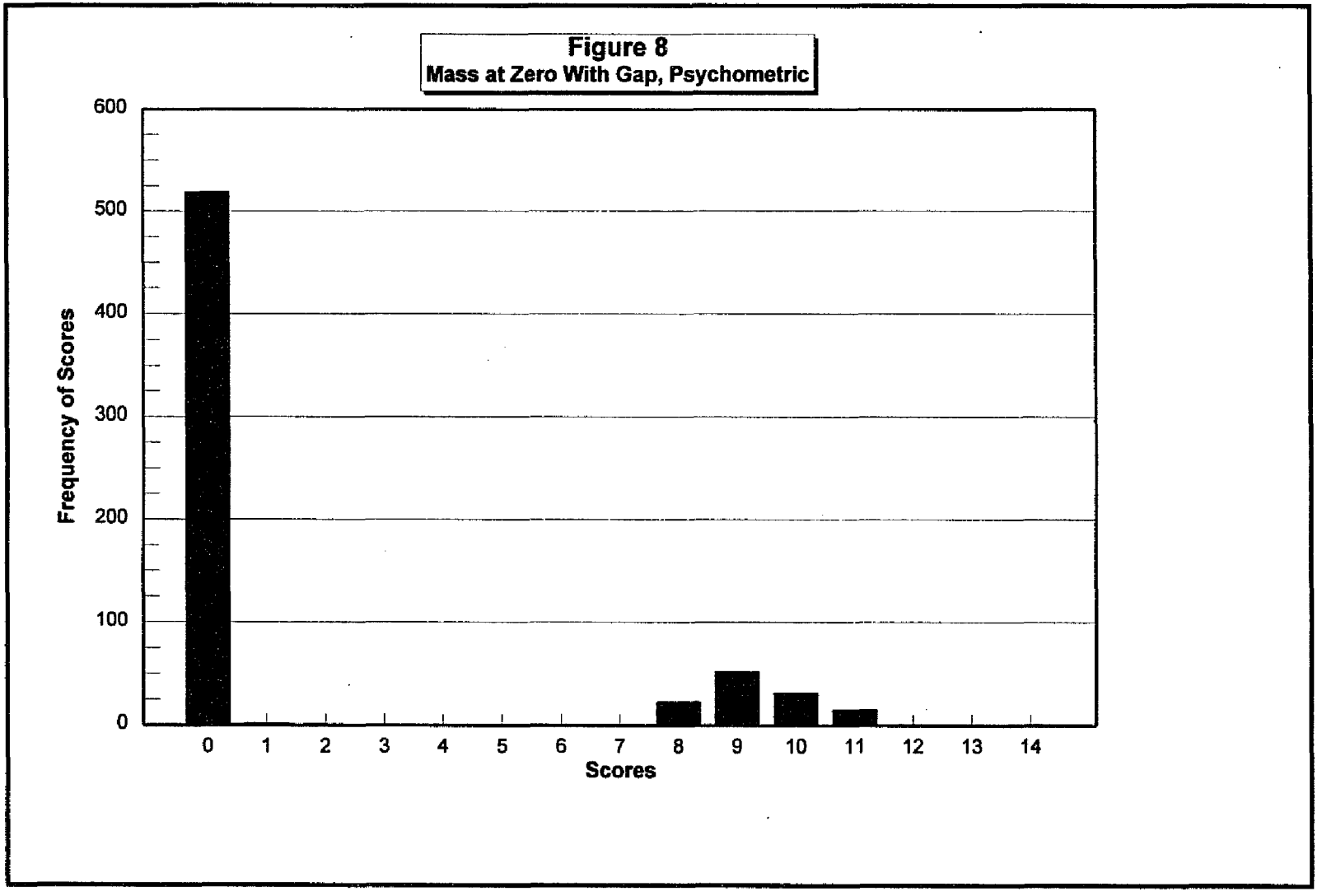


Figure 8
Mass at Zero With Gap, Psychometric
Frequency of Scores
600
500
400
300
200
100
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Scores

CHAPTER FOUR

RESULTS

Monte Carlo simulations were conducted to assess the comparative power properties of the independent samples t-test and Wilcoxon Rank Sum test. The results are summarized into 32 figures in this section. Tables with the actual results are referenced in Appendices A-H.

As previously discussed, there are eight distributions, four sample sizes of (n1,n2)= (10,10), (5,15), (30,30), and (15,45), and eight treatment effects. Each figure compares the power of the independent samples t-test and Wilcoxon Rank Sum test for a specified distribution, sample size, and treatment effect. The Y axis represents power (1-ß), and ranges from 0 to 1.00, with a power level of 1.00 theoretically demonstrating maximum power. The X axis represents the associated effect size with a range of .25σ to 2.00σ, where the standard deviation refers to the respective distribution. The effect size is a function of the distribution's standard deviation multiplied by a constant of .25σ. Power was obtained by comparing the actual alpha for each test statistic under each condition, to departures from nominal alpha. The obtained values were recorded and are represented as bar graphs. Although various alpha levels were used (.10, .05, .01) in the simulation, only alpha= .05 is presented in the results. The results for .10 and .01 are similar, and therefor are not presented here.

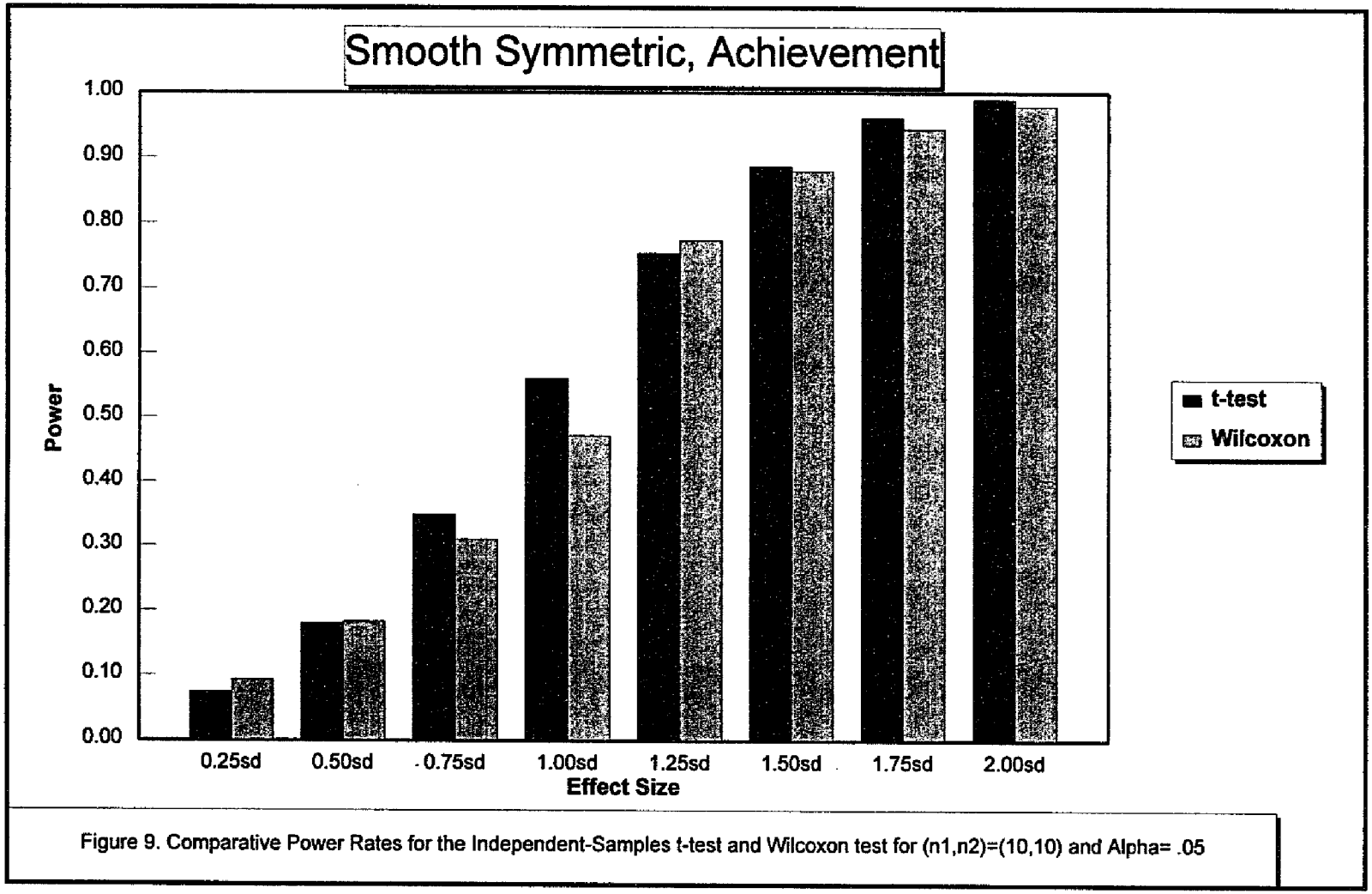


Figure 9. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(10,10) and Alpha= .05

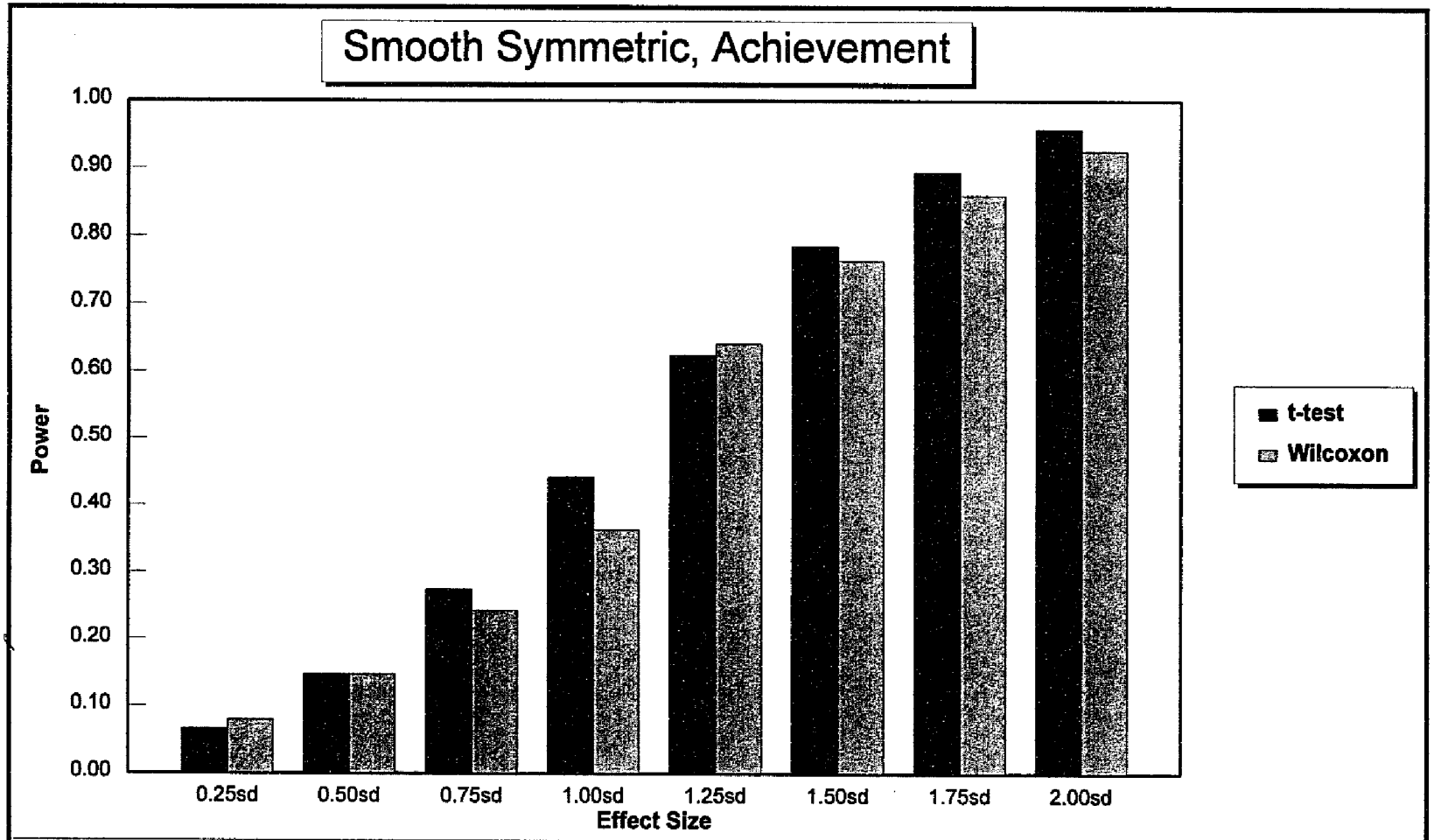


Figure 10. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(5,15) and Alpha= .05

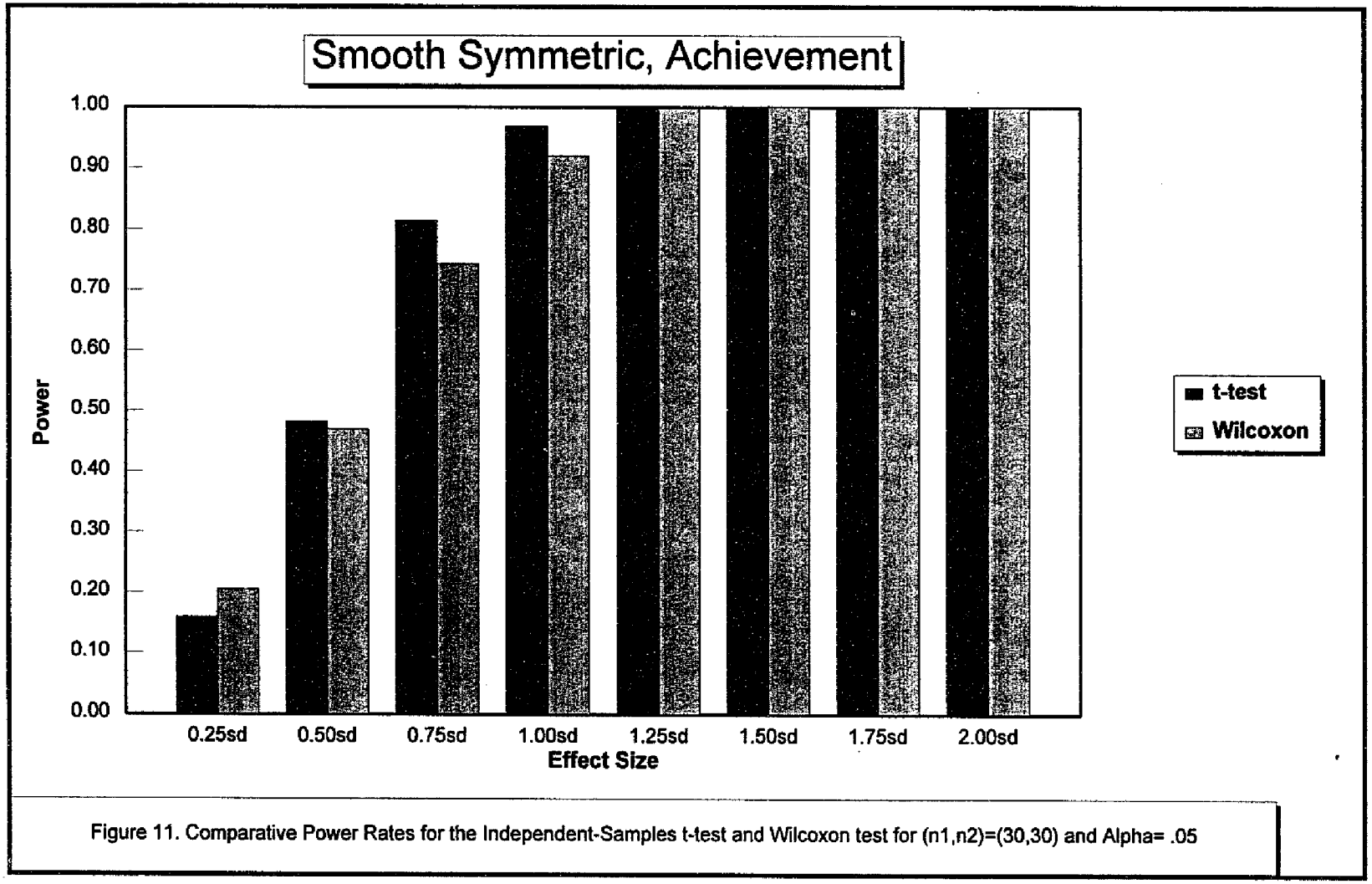


Figure 11. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(30,30) and Alpha= .05

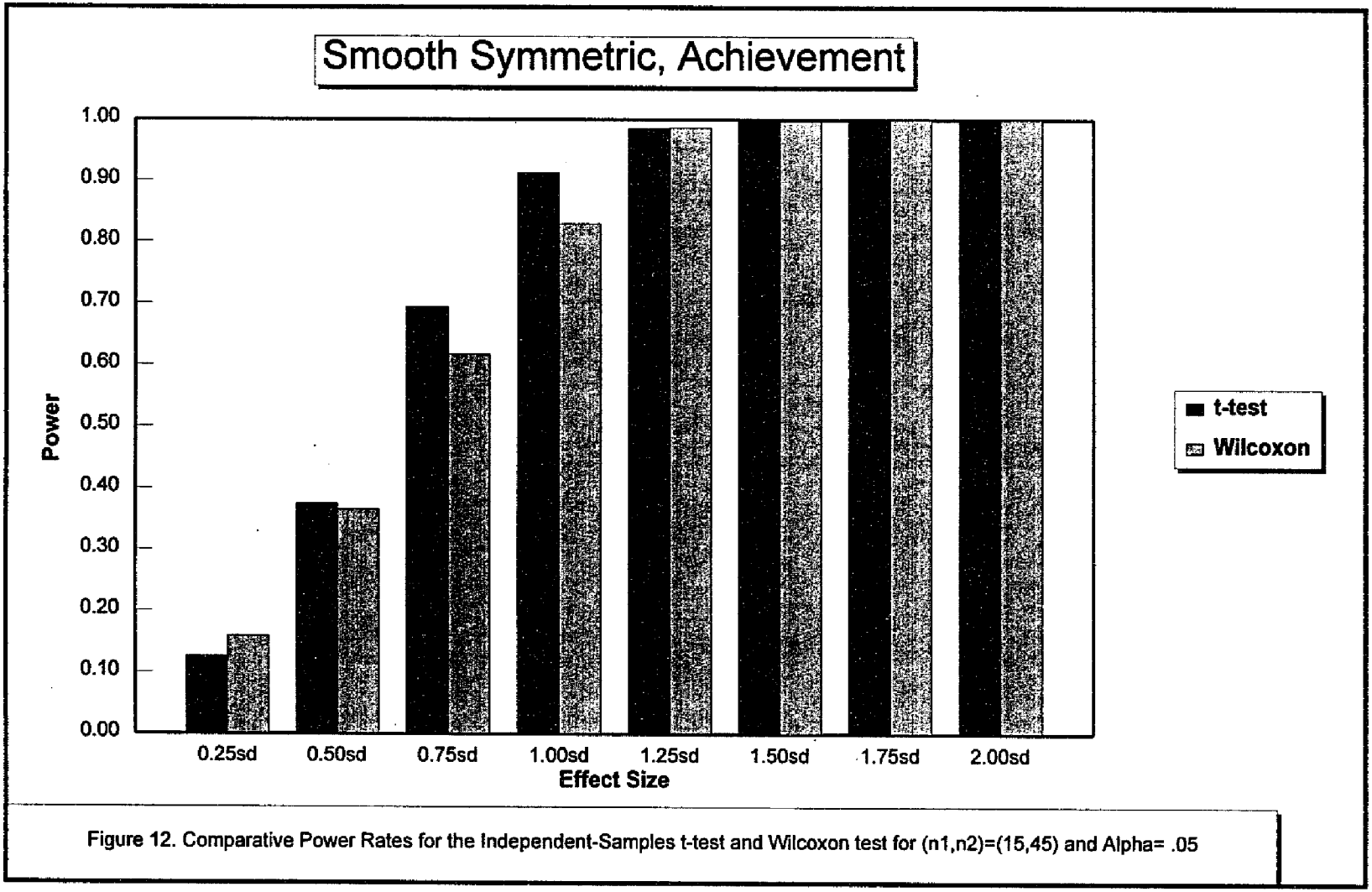


Figure 12. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(15,45) and Alpha= .05

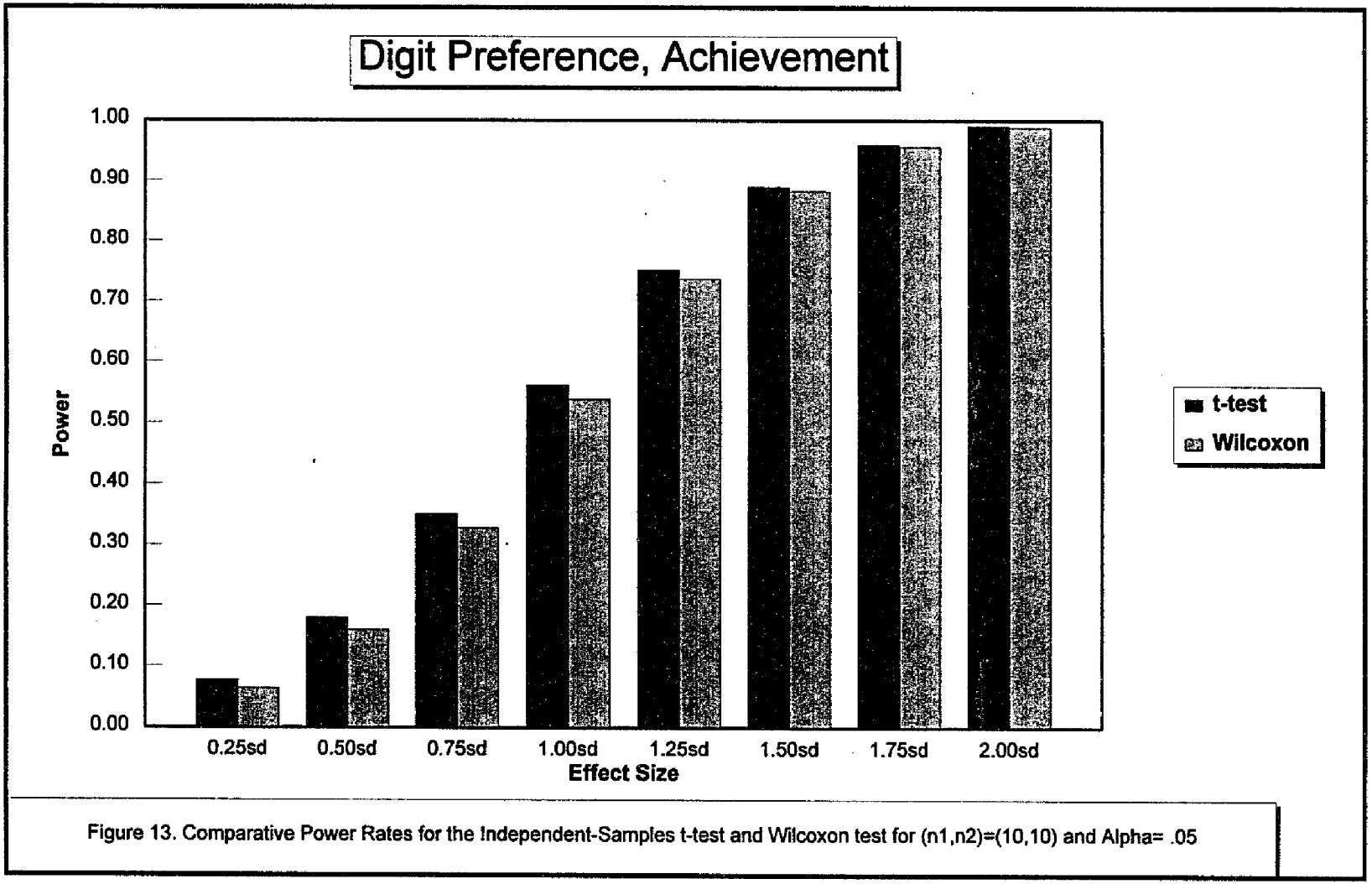


Figure 13. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(10,10) and Alpha= .05

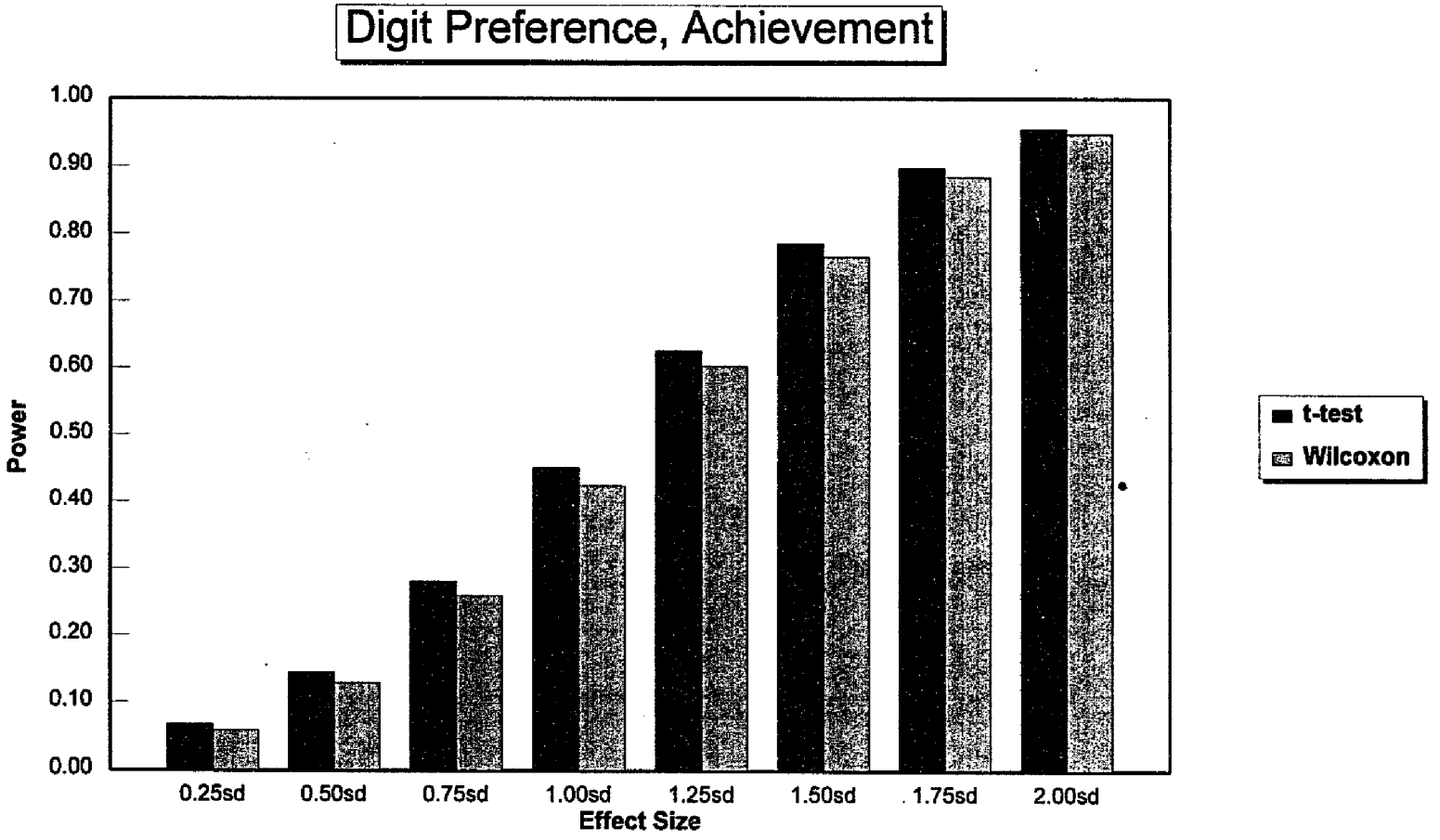


Figure 14. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(5,15) and Alpha= .05

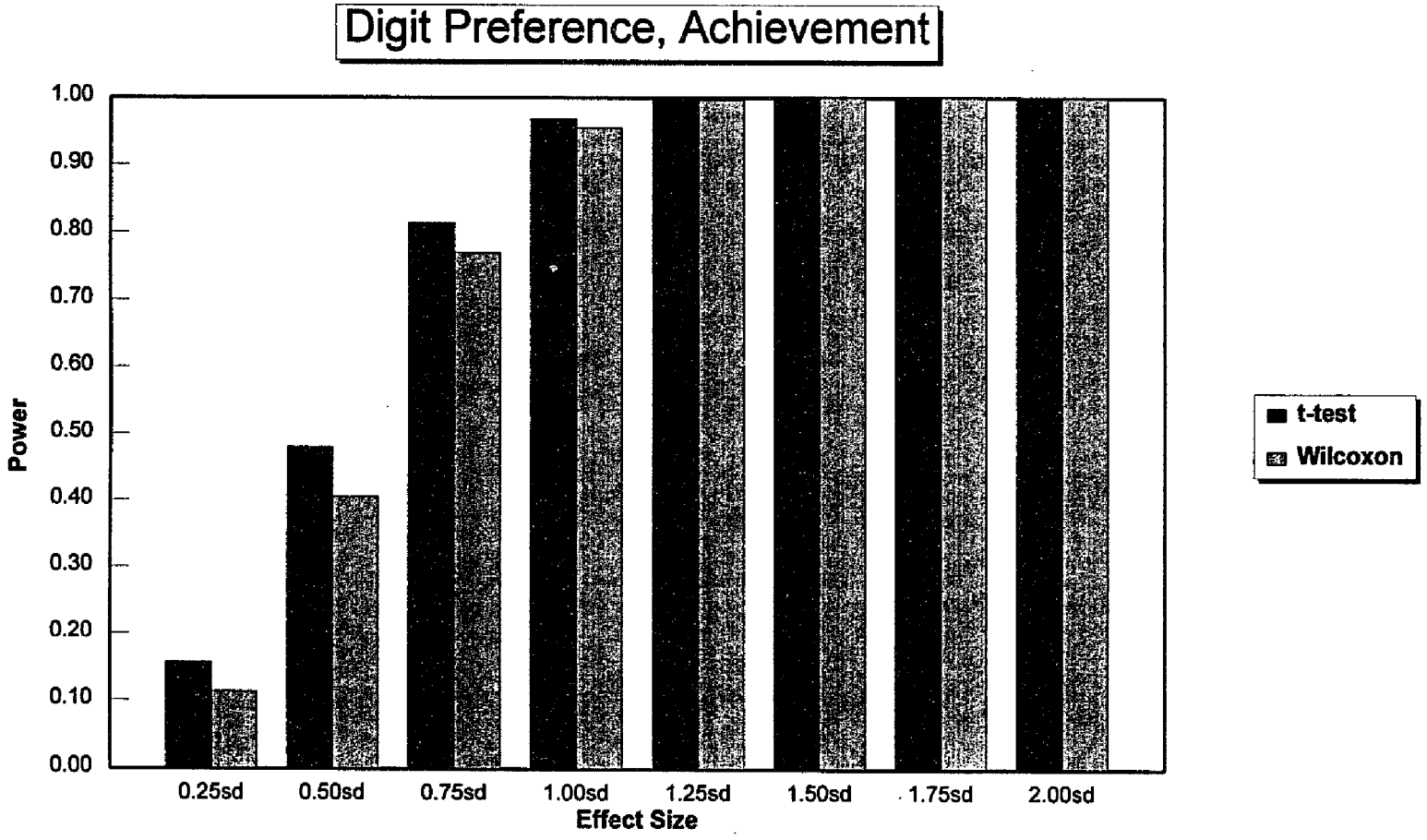


Figure 15. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(30,30) and Alpha= .05

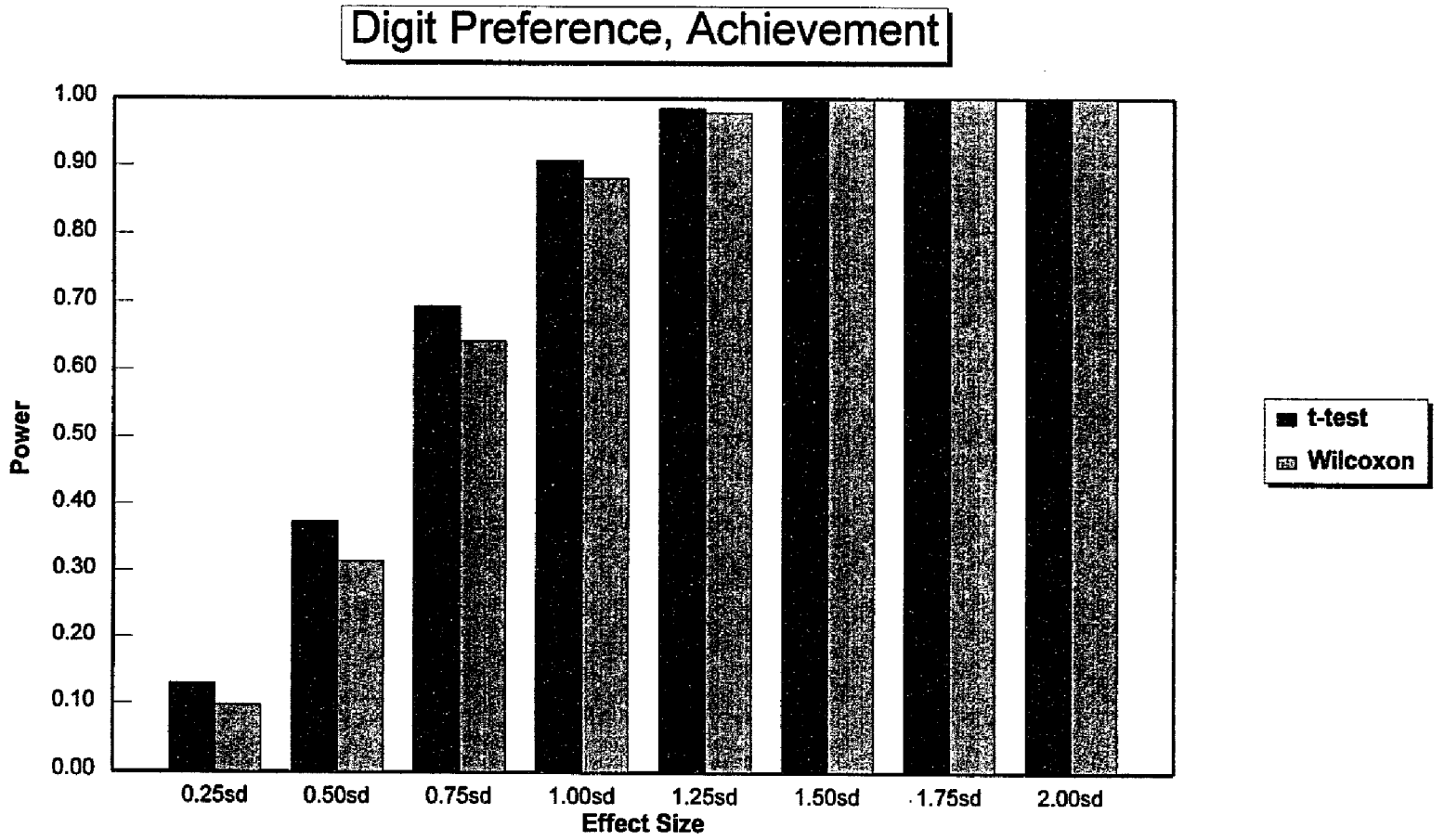


Figure 16. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(15,45) and Alpha= .05

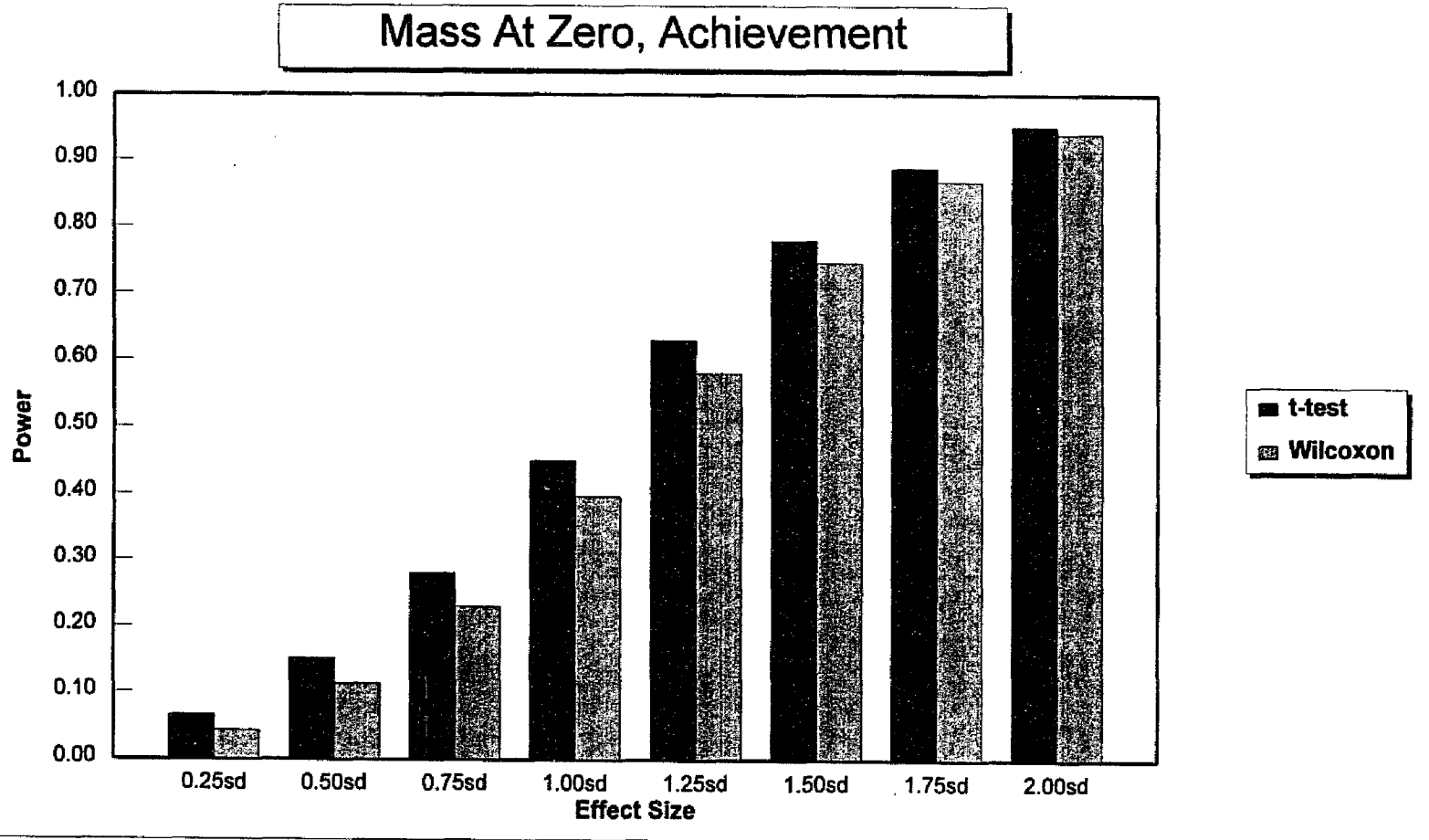


Figure 17. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(10,10) and Alpha= .05

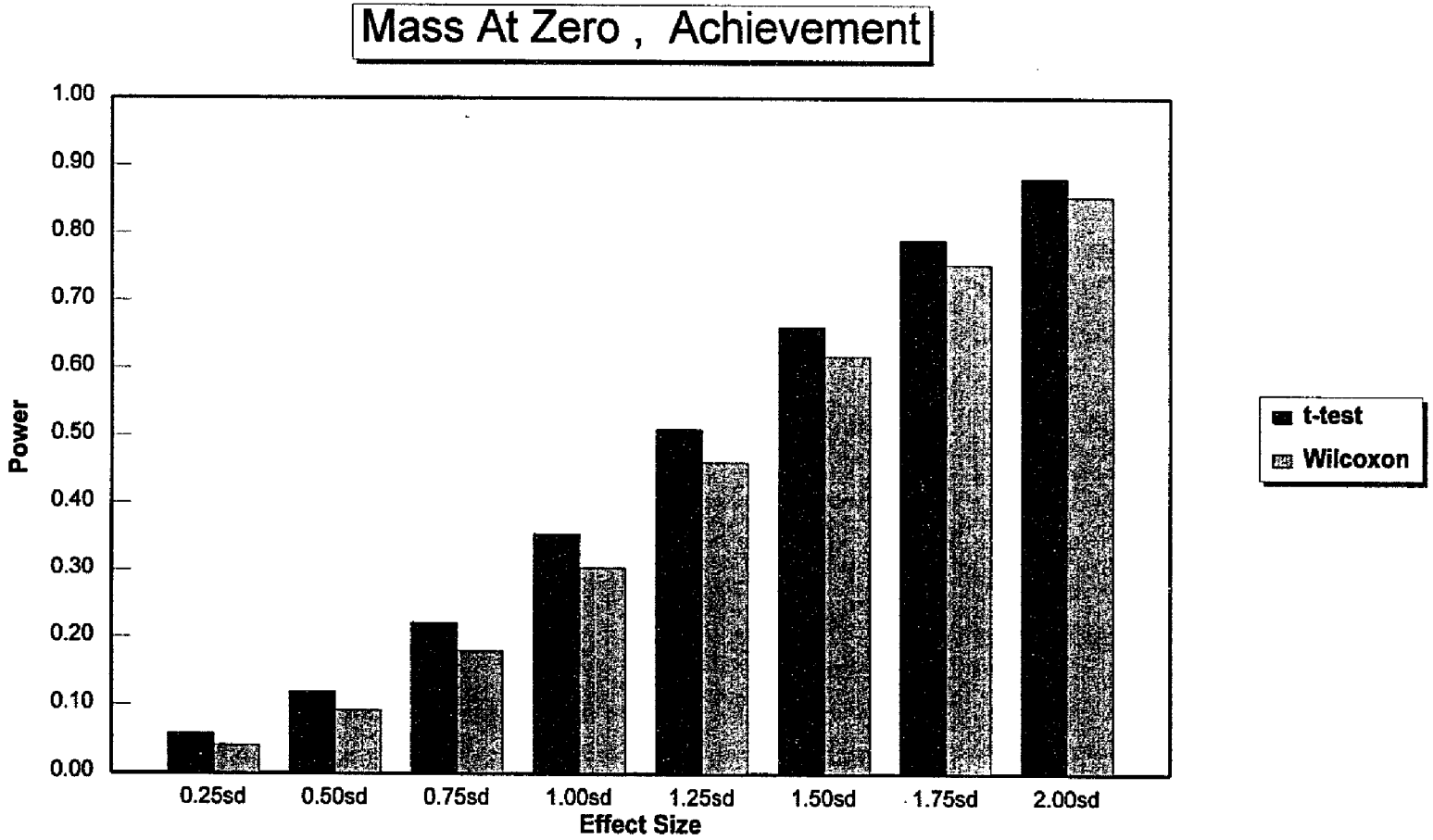


Figure 18. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(5,15) and Alpha= .05

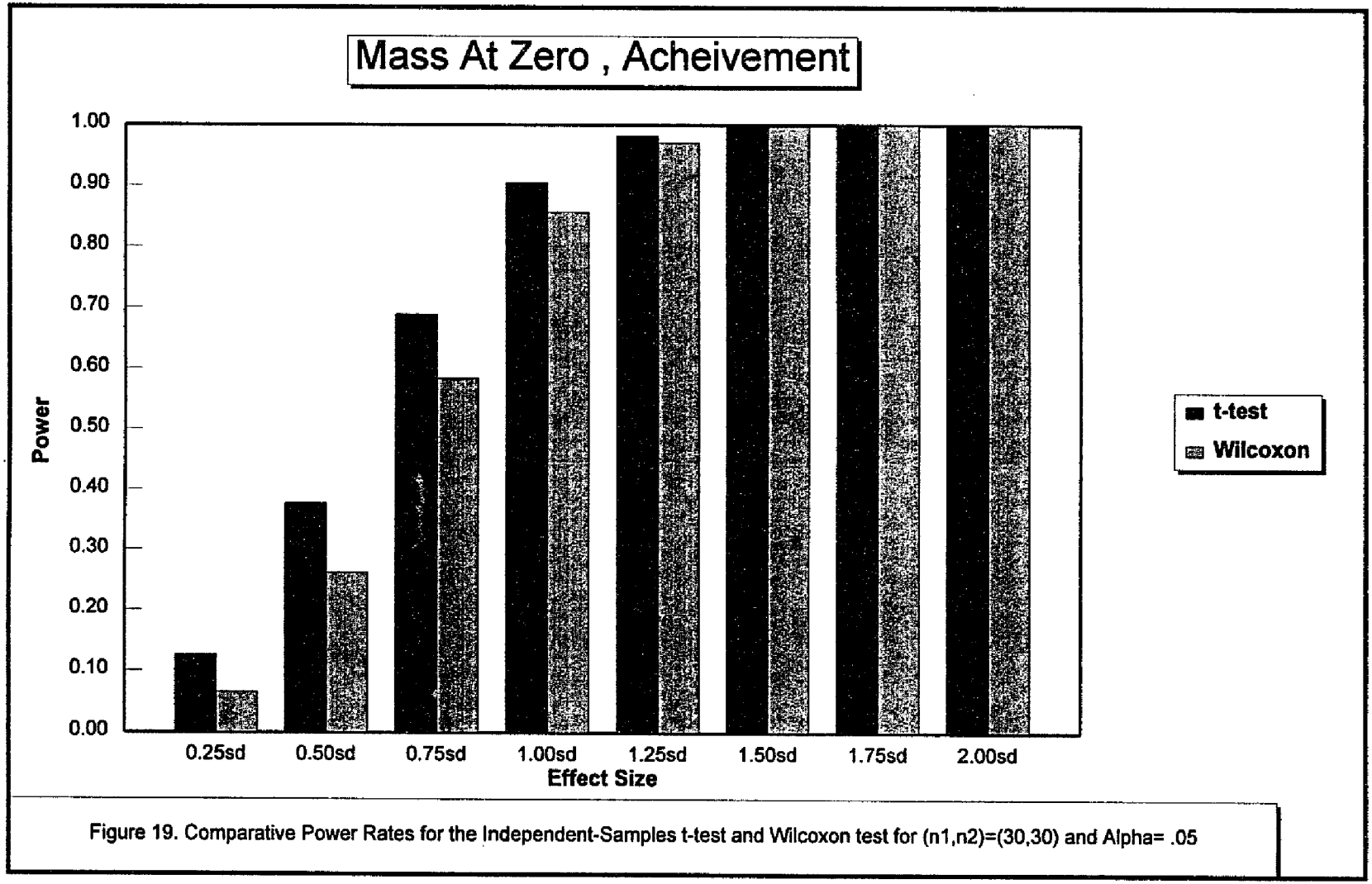


Figure 19. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(30,30) and Alpha= .05

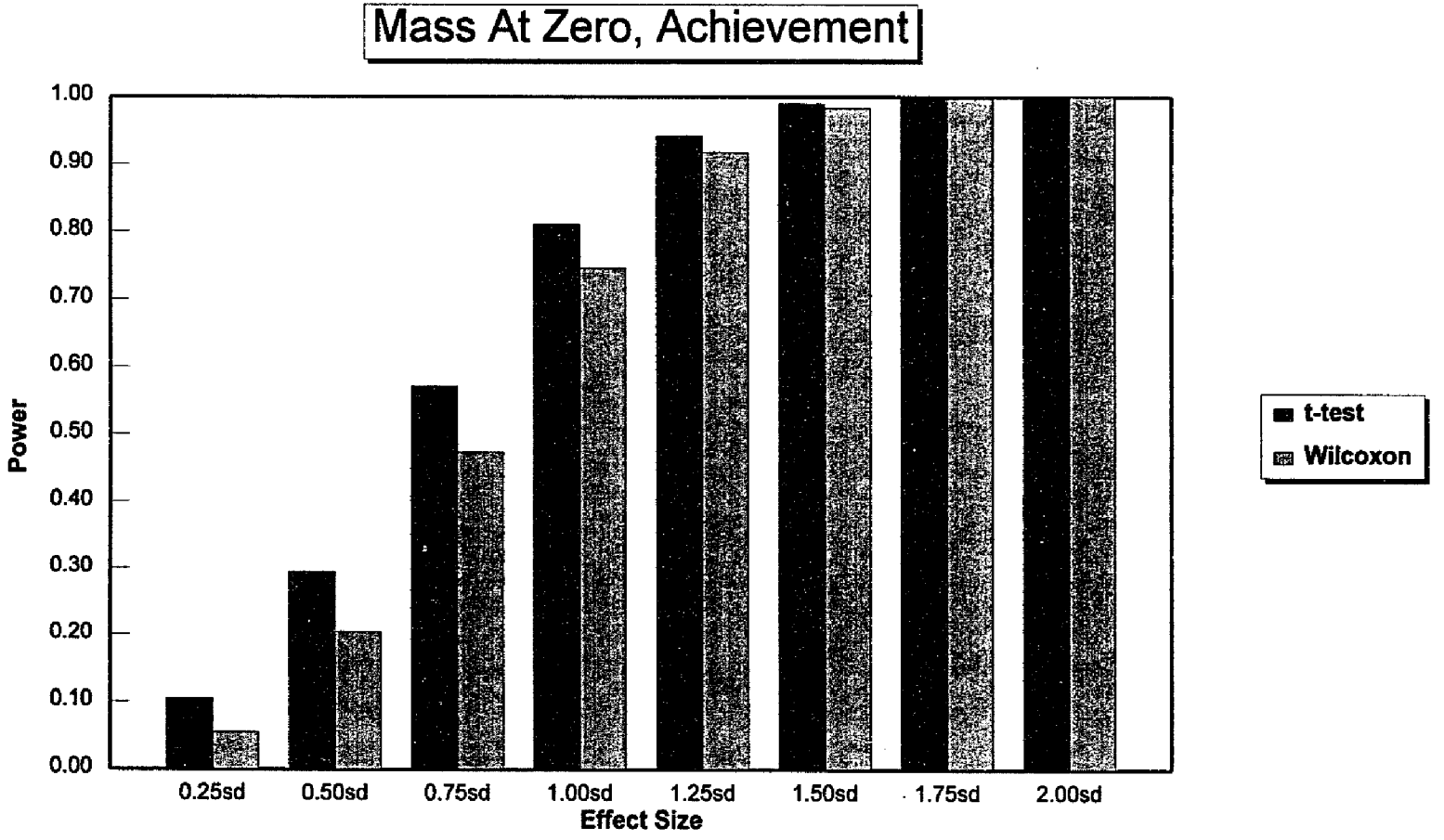


Figure 20. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(15,45) and Alpha= .05

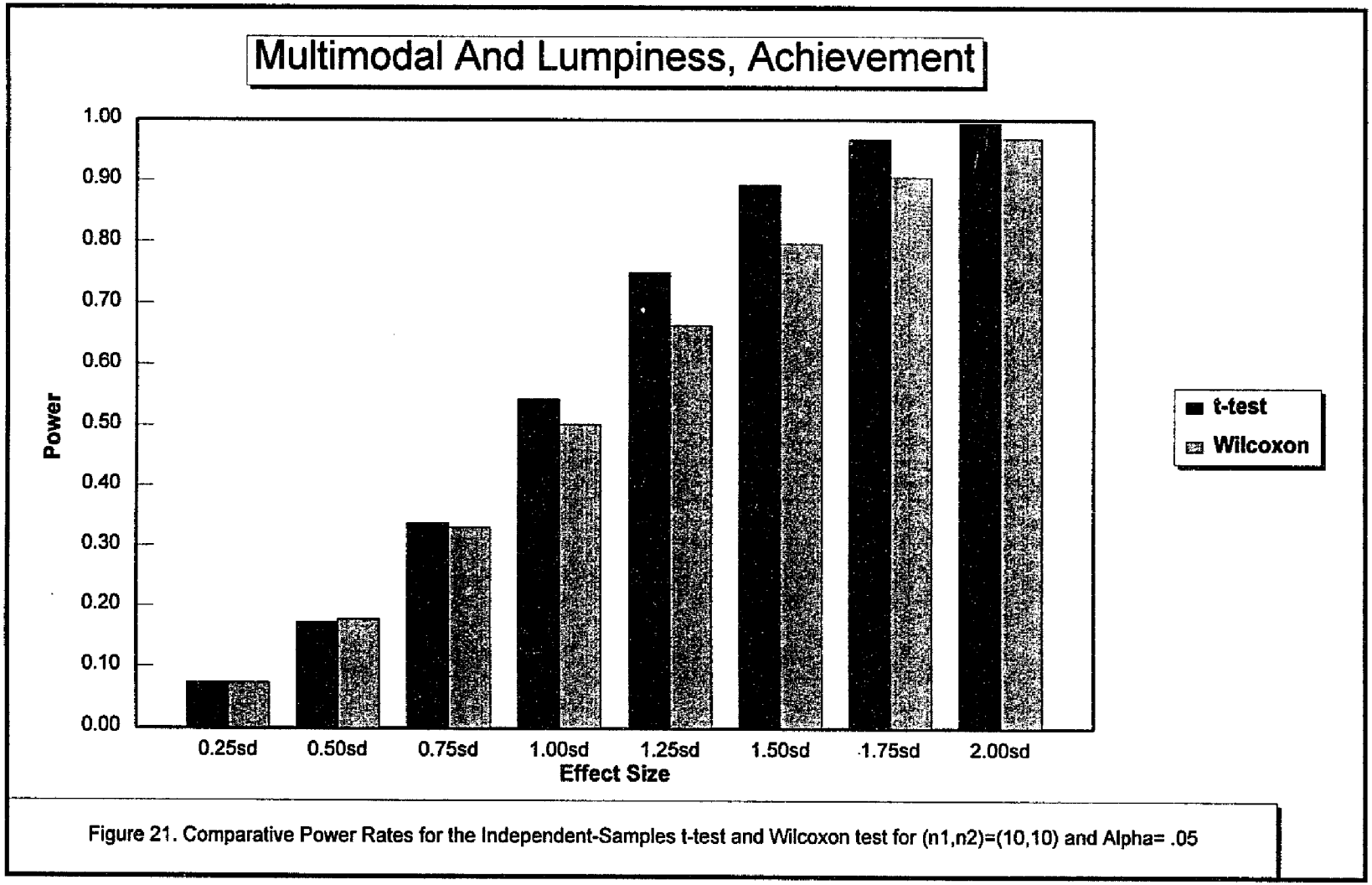


Figure 21. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(10,10) and Alpha= .05

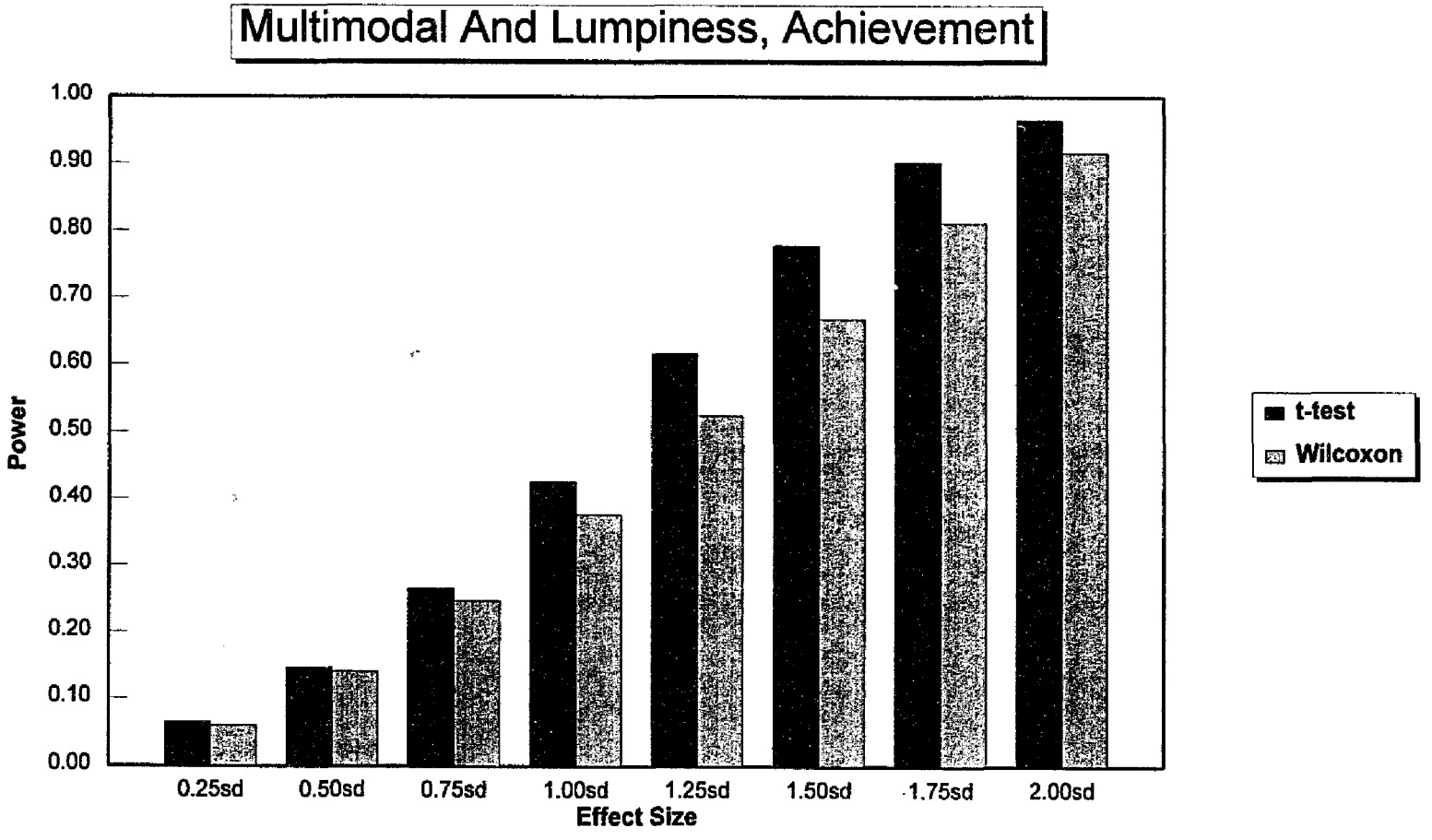


Figure 22. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(5,15) and Alpha= .05

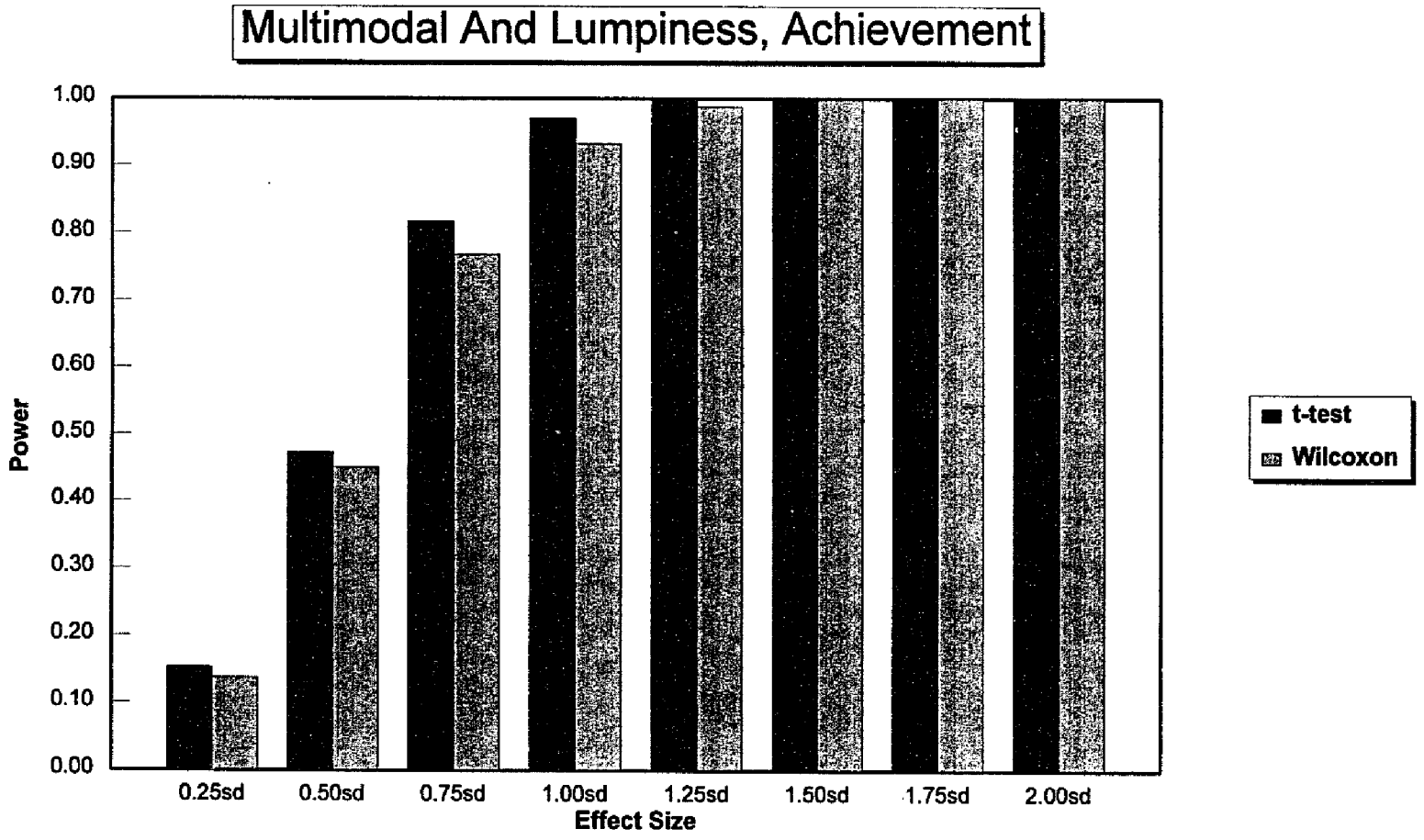


Figure 23. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(30,30) and Alpha= .05

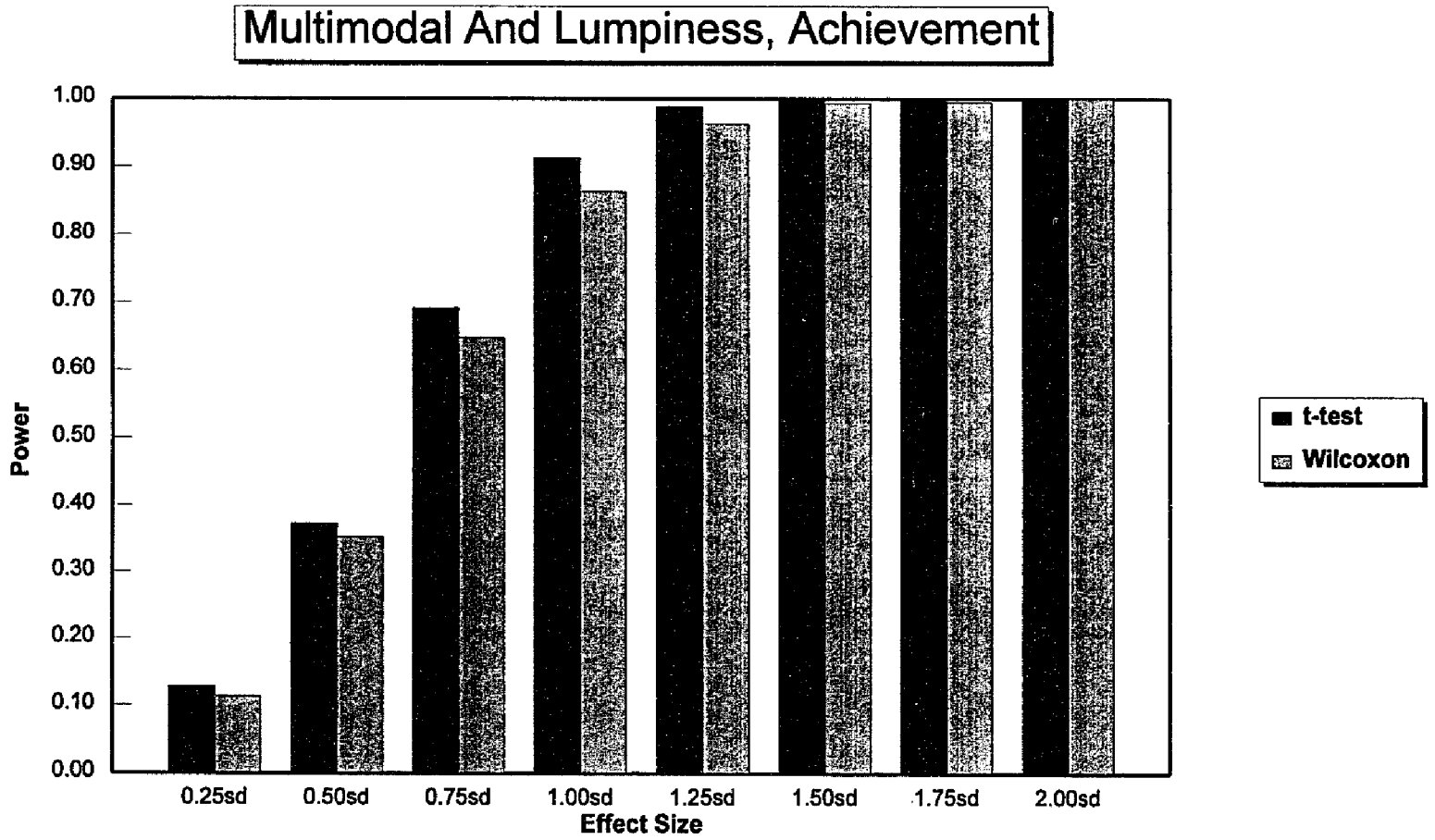


Figure 24. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(15,45) and Alpha= .05

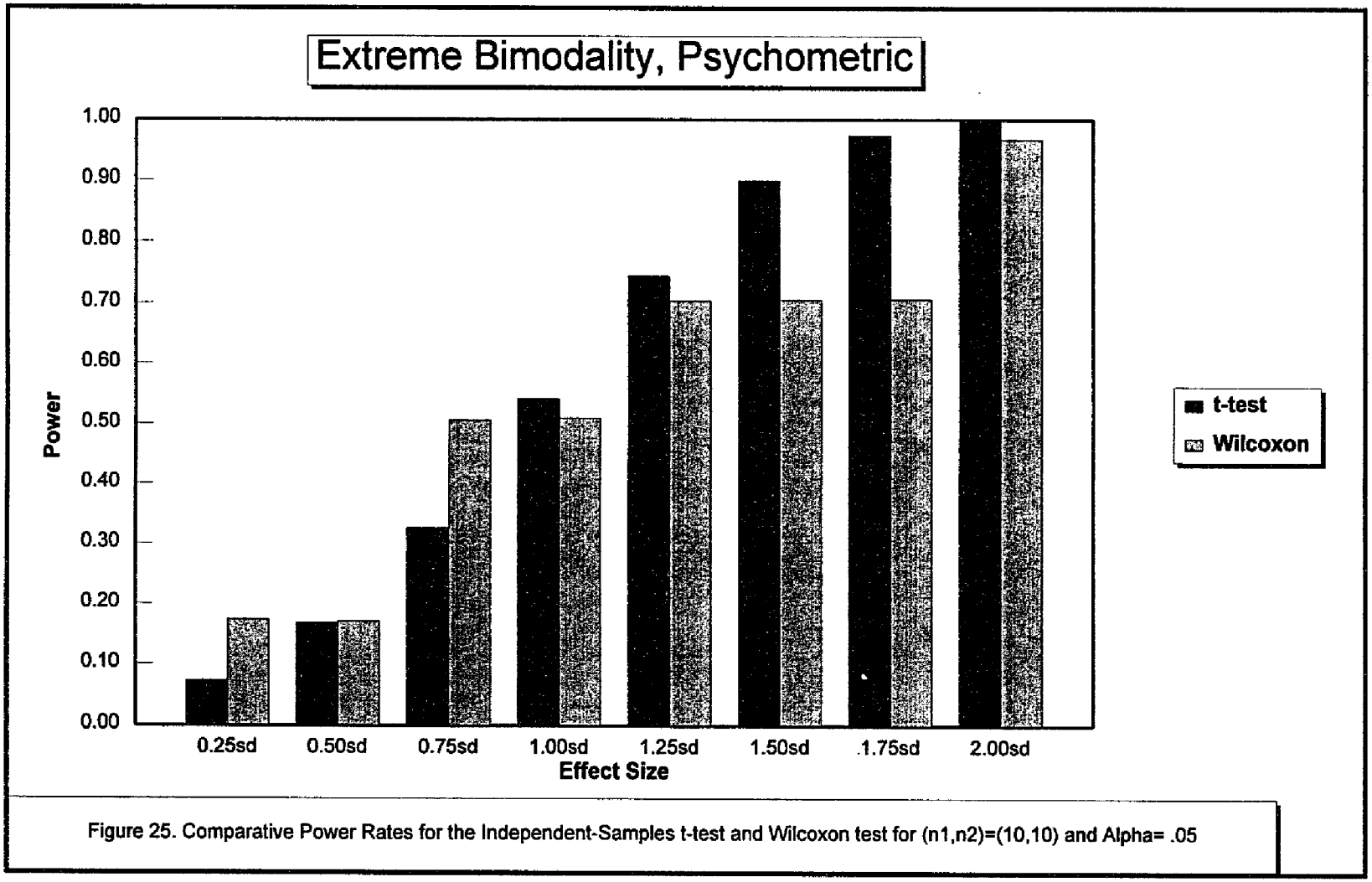


Figure 25. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(10,10) and Alpha= .05

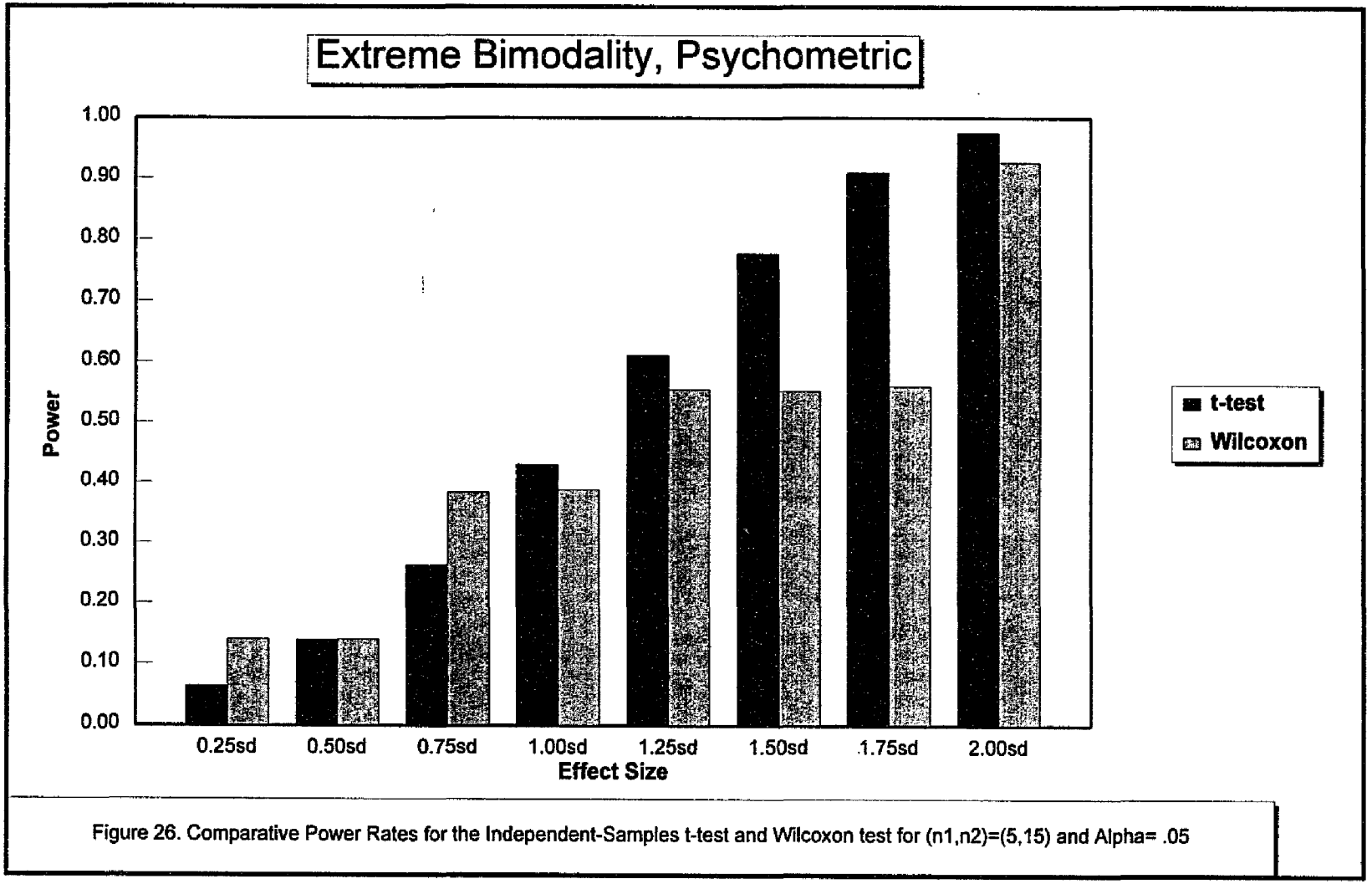


Figure 26. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(5,15) and Alpha= .05

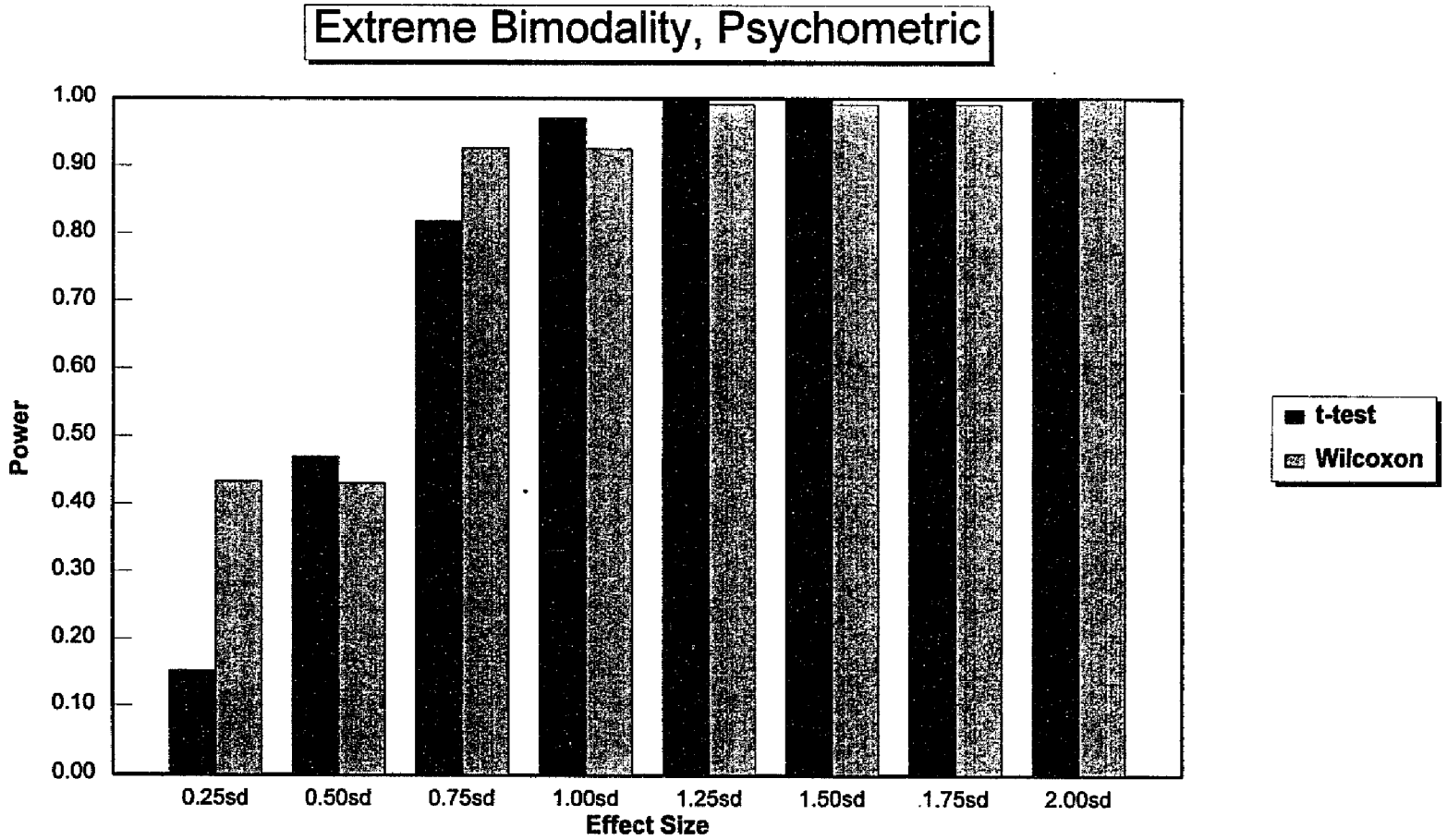


Figure 27. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(30,30) and Alpha= .05

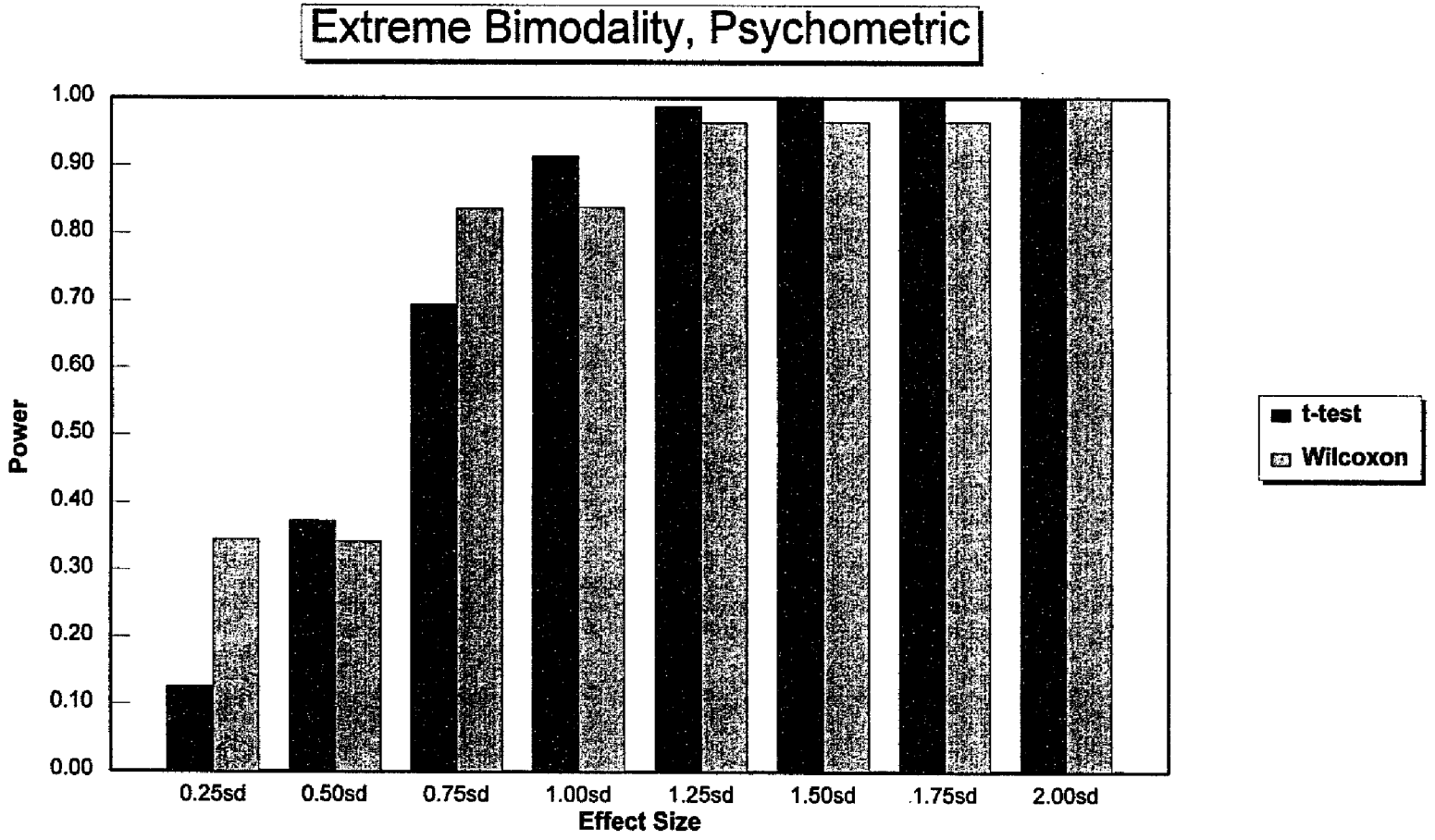


Figure 28. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(15,45) and Alpha= .05

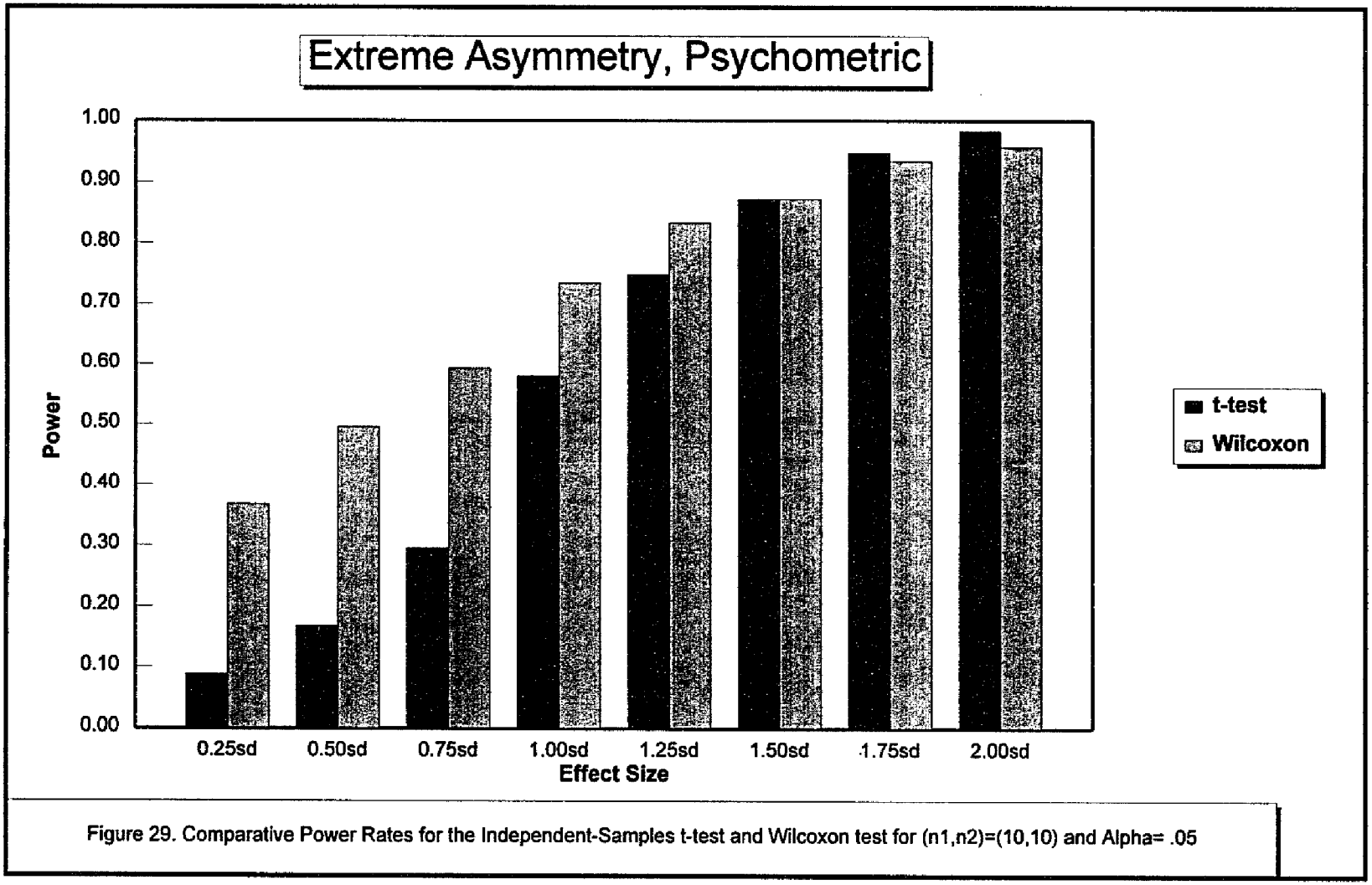


Figure 29. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(10,10) and Alpha= .05

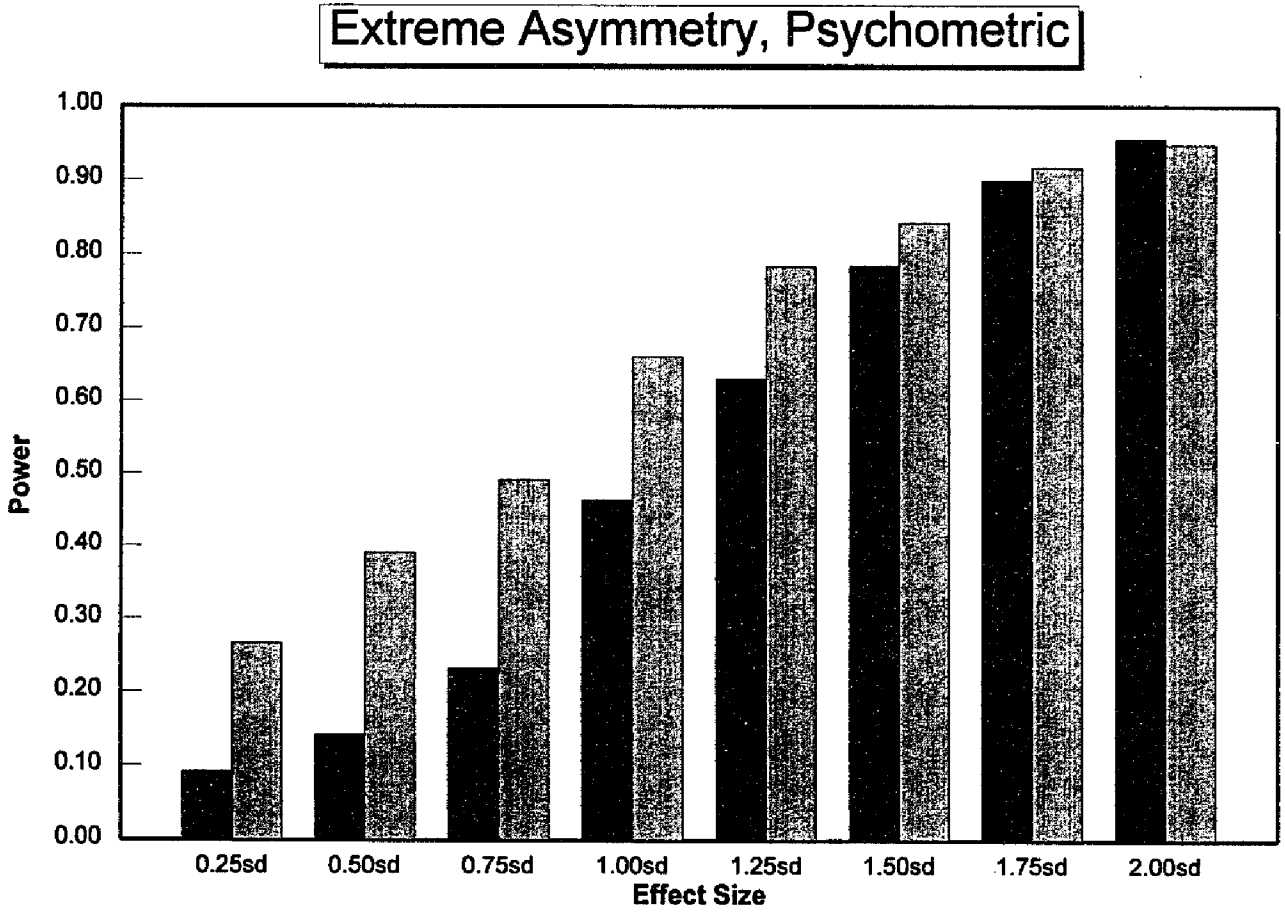


Figure 30. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(5,15) and Alpha= .05

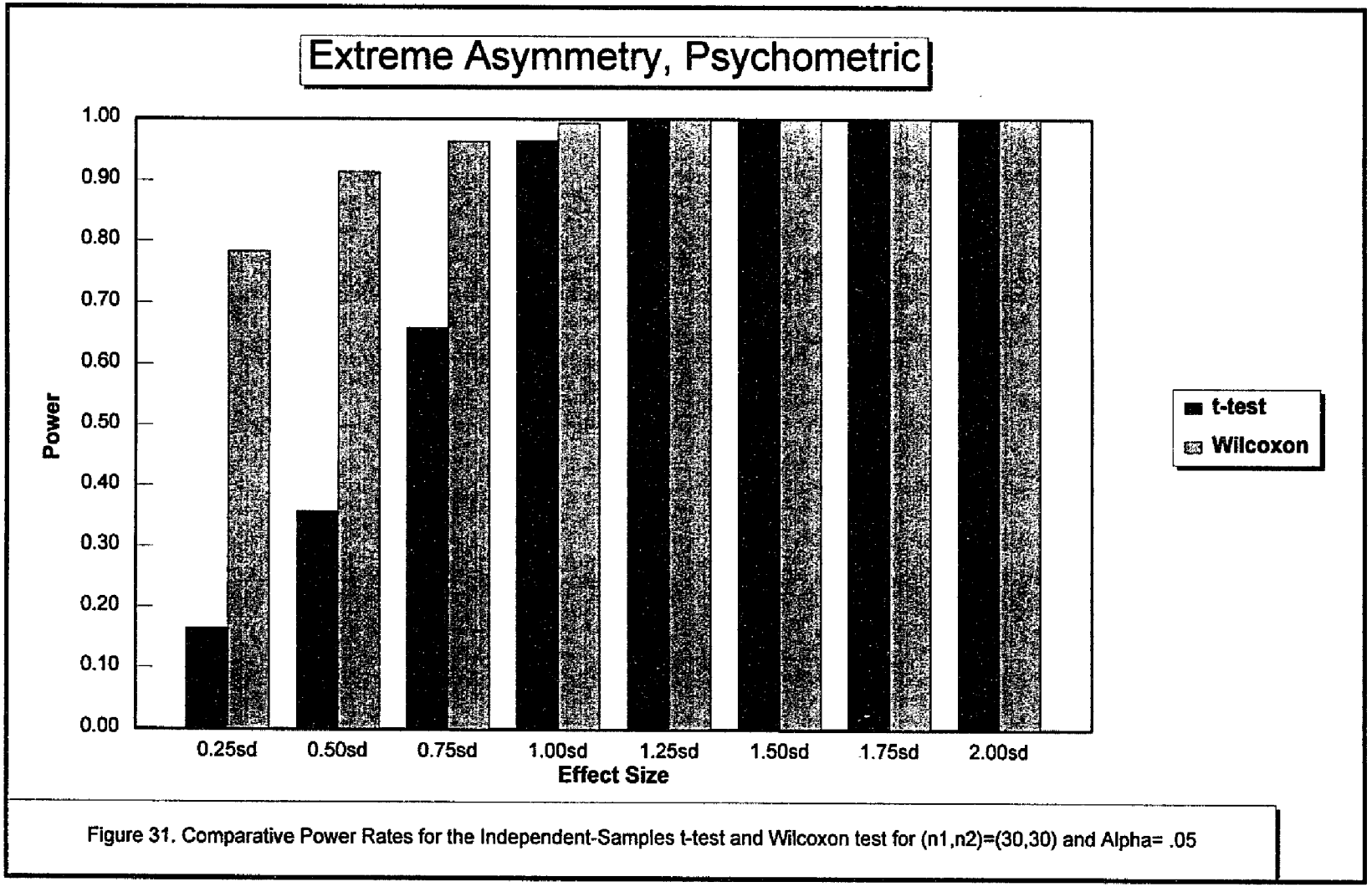


Figure 31. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(30,30) and Alpha= .05

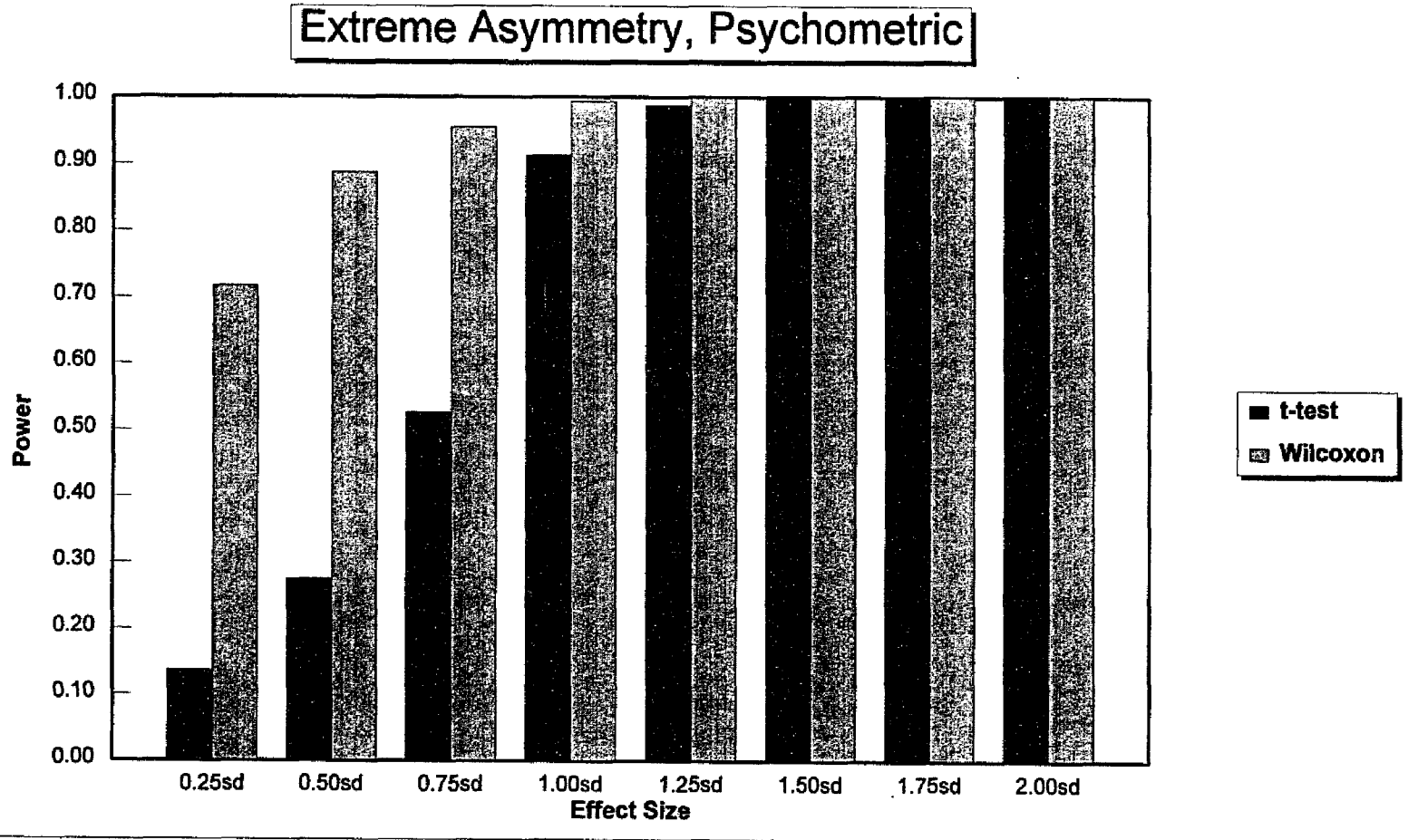


Figure 32. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(15,45) and Alpha= .05

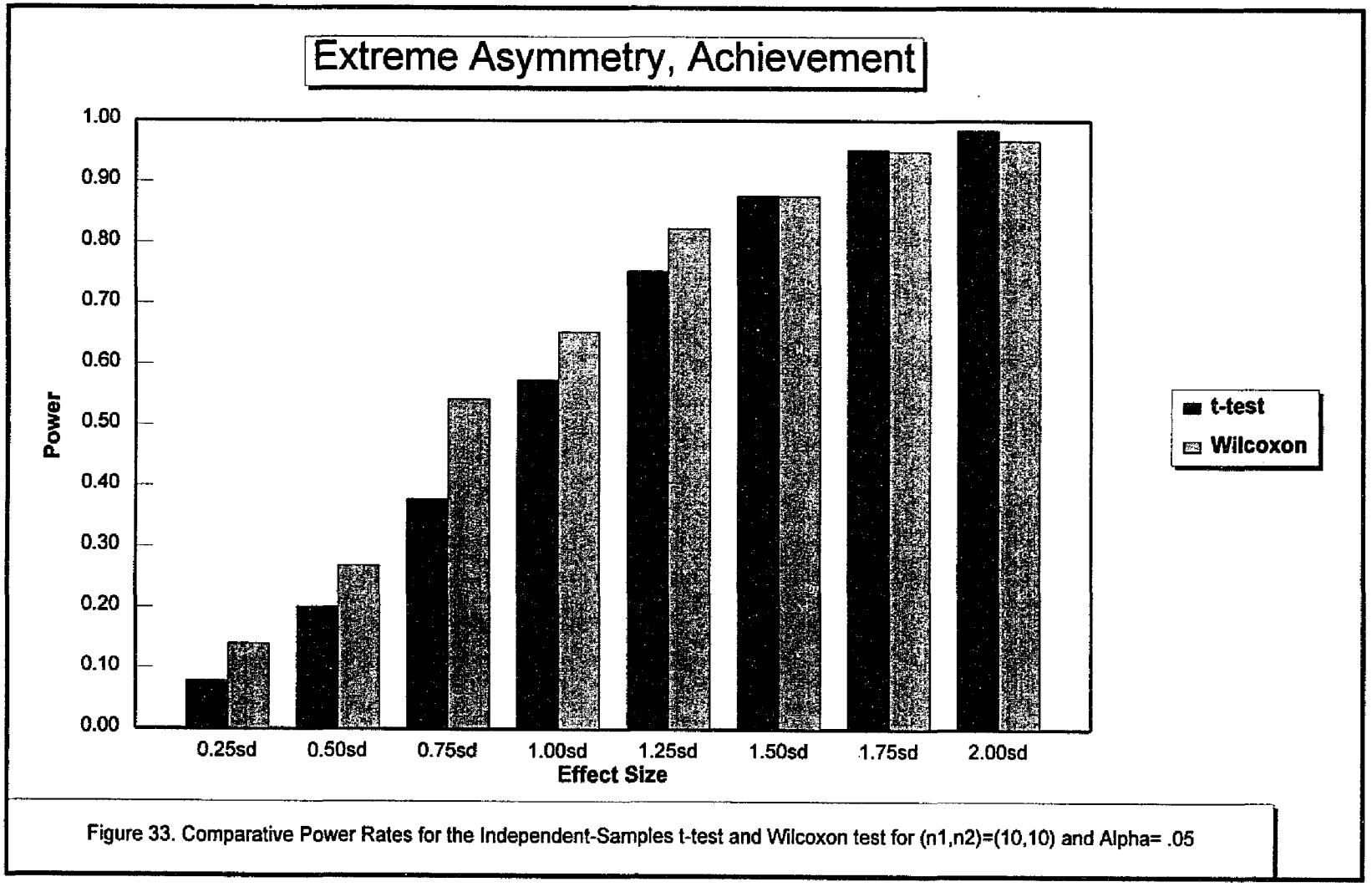


Figure 33. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(10,10) and Alpha= .05

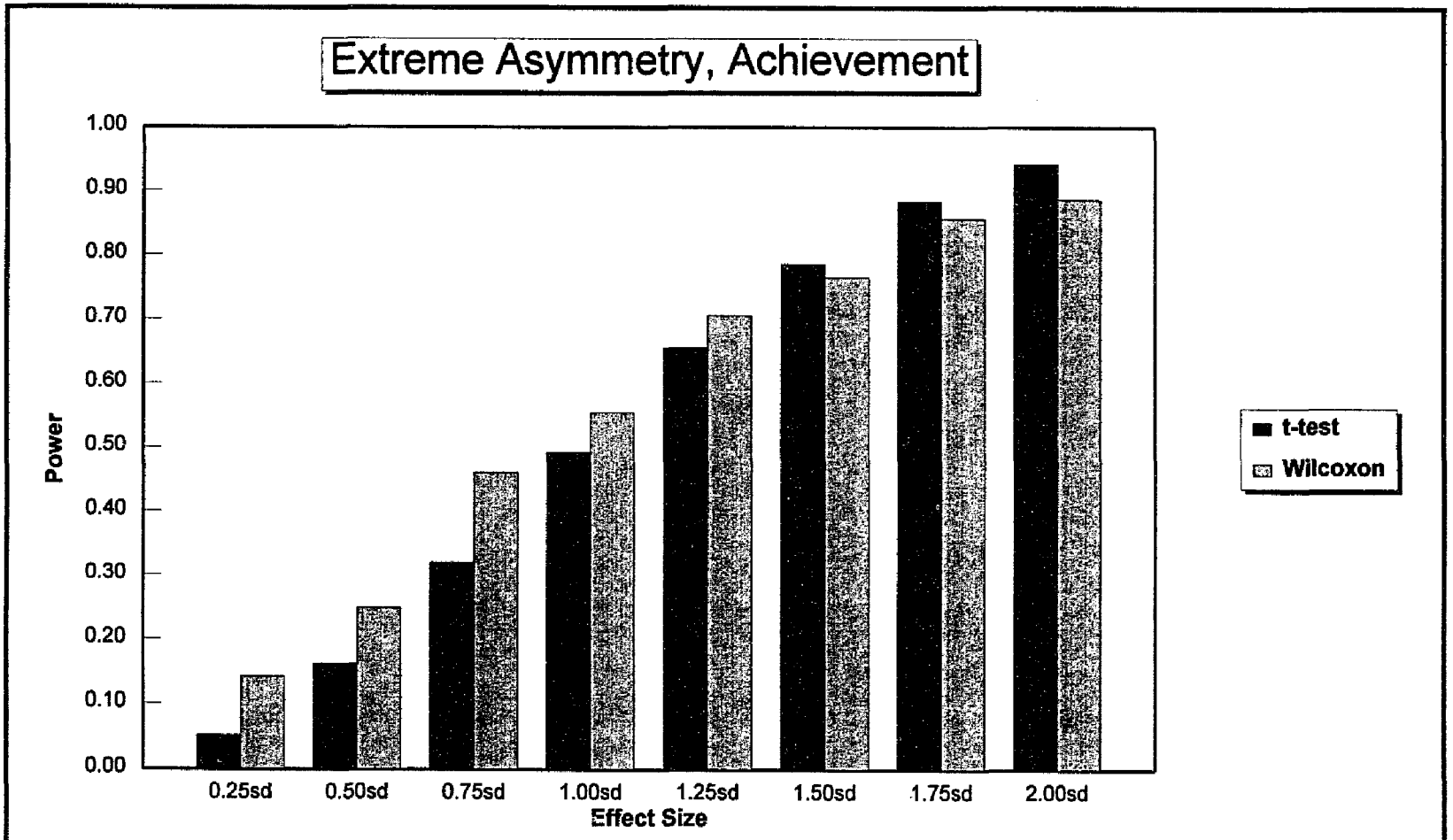


Figure 34. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(5,15) and Alpha= .05

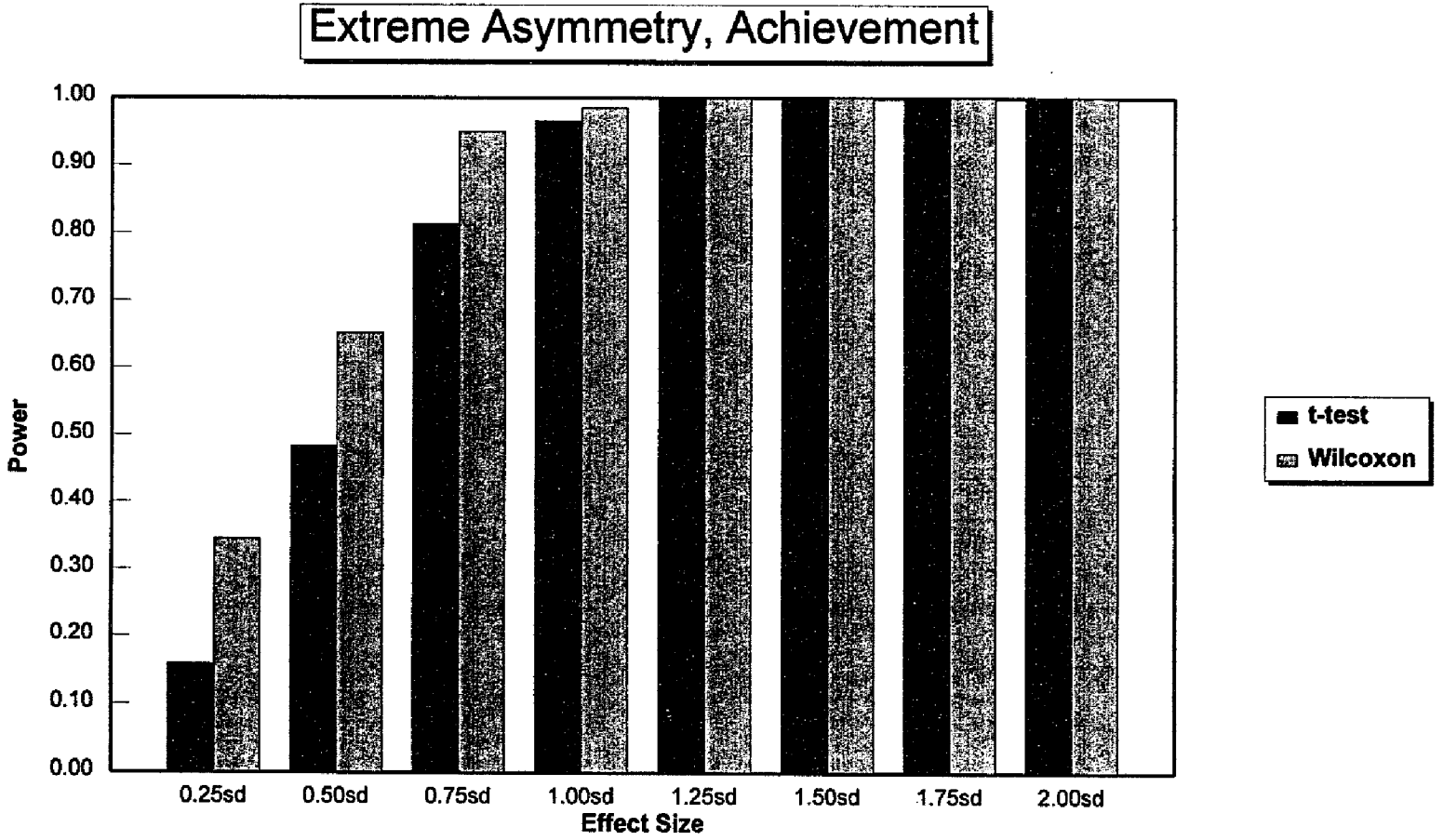


Figure 35. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(30,30) and Alpha= .05

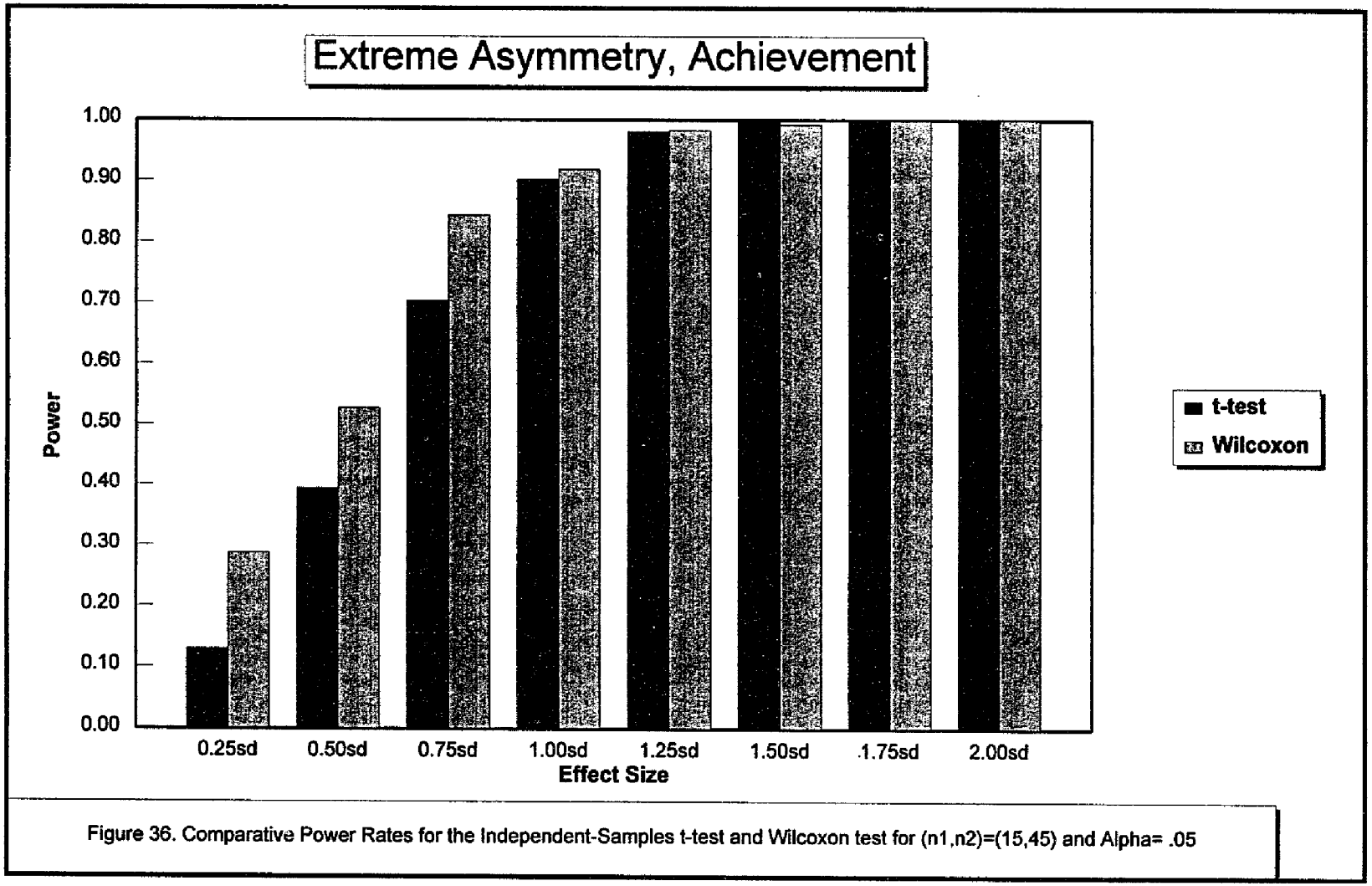


Figure 36. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(15,45) and Alpha= .05

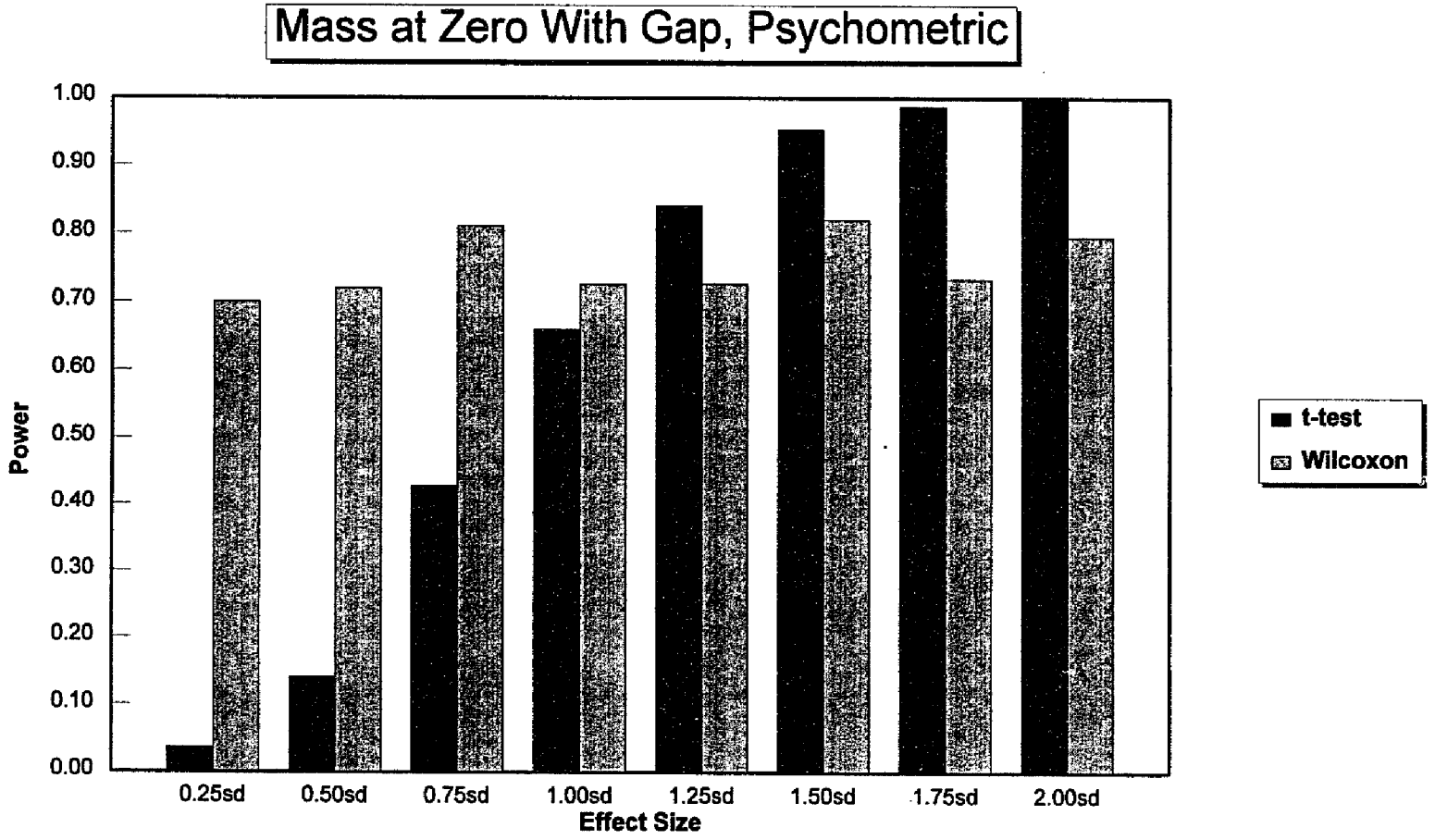


Figure 37. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(10,10) and Alpha= .05

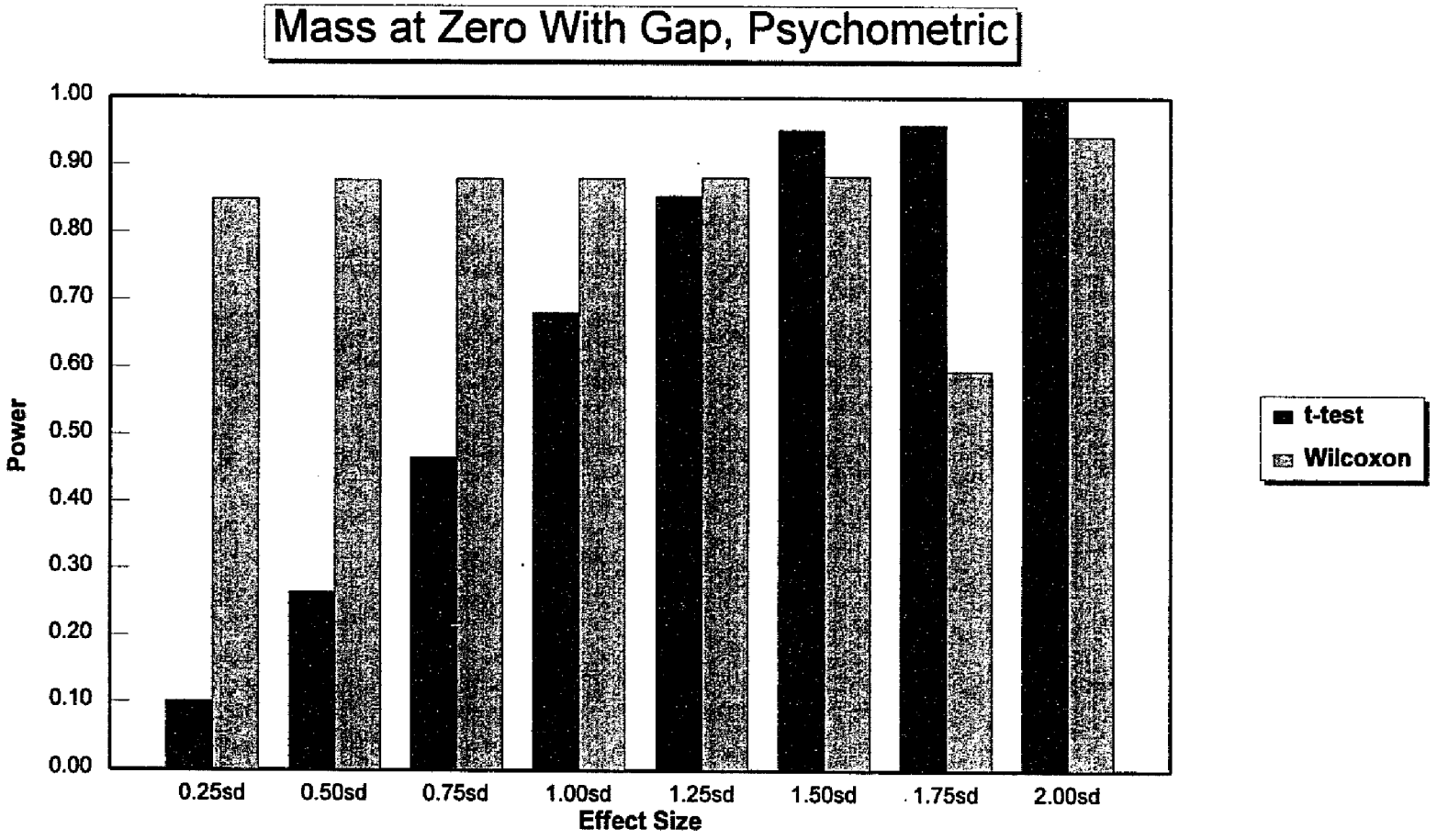


Figure 38. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(5,15) and Alpha= .05

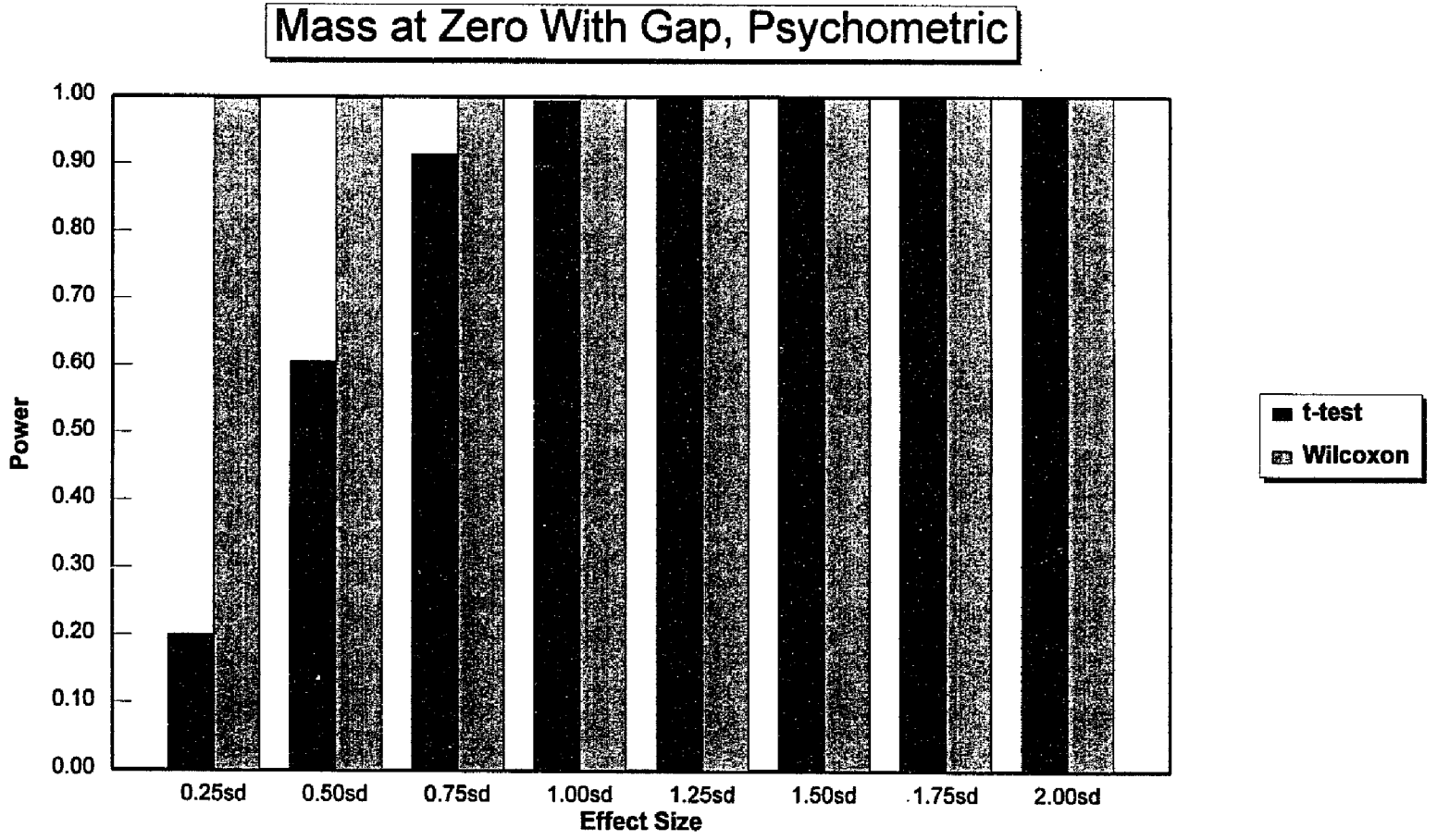


Figure 39. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(30,30) and Alpha= .05

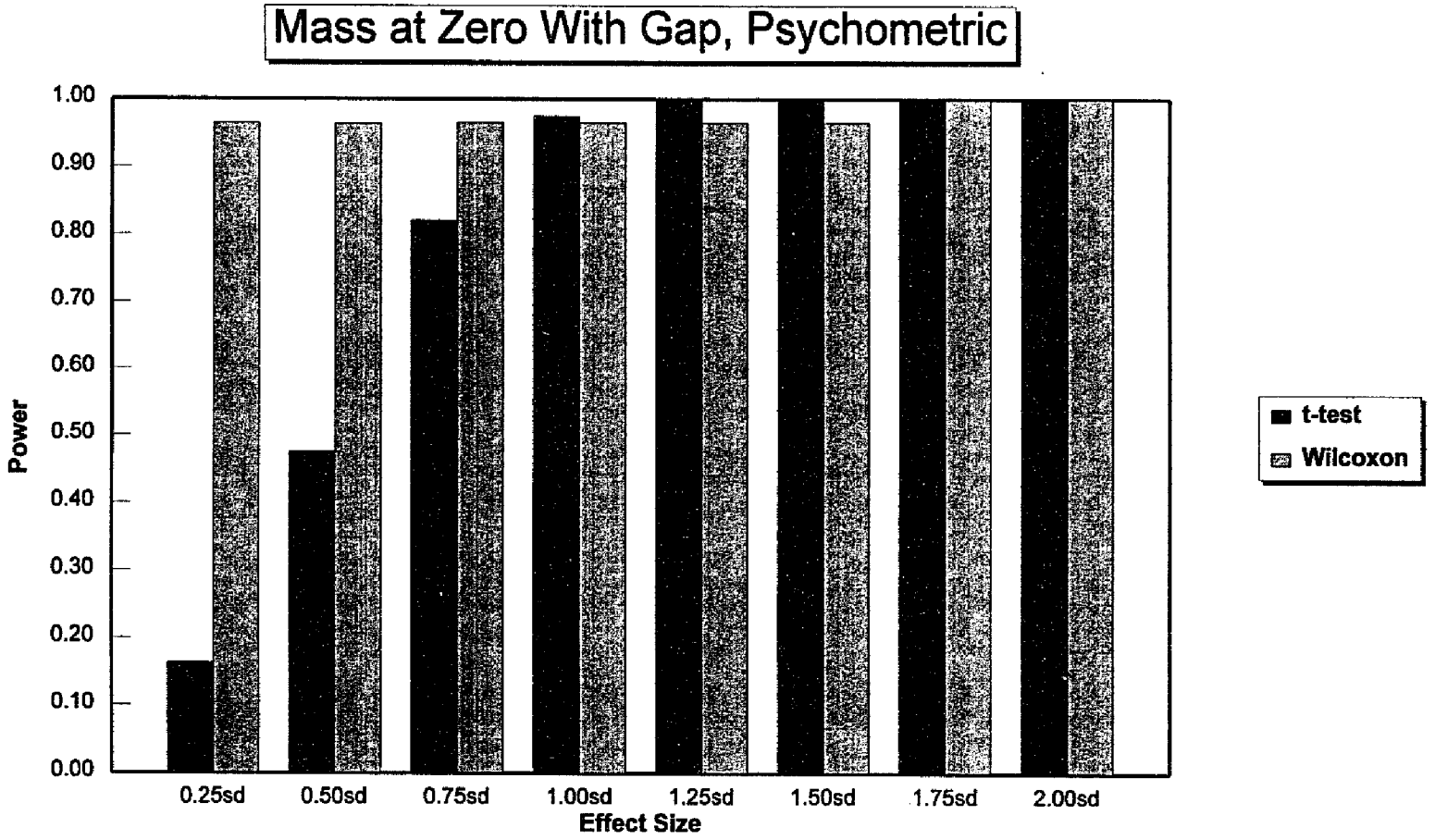


Figure 40. Comparative Power Rates for the Independent-Samples t-test and Wilcoxon test for (n1,n2)=(15,45) and Alpha= .05

CHAPTER FIVE

DISCUSSION AND CONCLUSION

Study Overview

When conducting research, the primary concern for any statistician or researcher should be the validity of the design and the reliability of instruments. However, if the wrong statistics are utilized, are used incorrectly, or ineffecient statistics are used, this will void even the best designed studies. Many view the application of statistical tests as a rather simplified clear cut process. Yet, as previously discussed the appropriate application of statistical tests may be unclear and in some situations controversial.

The robustness and power of the independent samples t-test and Wilcoxon Rank Sum test to violations of normality has been in the forefront of this controversy. In the past, the widespread belief that applied data sets are normal, the influence of Monte Carlo studies on known mathematical distributions demonstrating robustness, and the assumption that nonparametric tests are less powerful, regardless of the shape of the population, has generated support for the use of parametric tests.

In education and psychology, the belief in normality is prevelent. In turn, the t-test being considered the Uniformally Most Powerful Unbiased test (UMPU) under normal

curve theory, becomes the primary tool in education and psychology to measure shift in location parameter. According to Micceri (1989), researcher should question the frequency of normal distributions in real world settings; yet, educators and psychometricians have been slow in recognizing this trend. Bradley (1977, 1978, 1980a, 1980b, 1980c, 1982), Micceri (1989), and other researchers have identified the prominence of non-normal distributions in real world settings in education and psychology.

A study by Sawilowsky and Blair (1992) demonstrated the robustness of the independent samples t-test (Type I & Type II error) to violations of normality in various real world data sets identified by Micceri (1989). However, they cautioned researchers and statisticians using the independent samples t-test, that a nonparametric test may be more powerful. Today, the focus has shifted from robustness to the comparative power of statistical tests to violations of the underlying assumptions. For the purpose of this study the comparative power of the independent samples t-test and its nonparametric counterpart the Wilcoxon Rank Sum test to violations from normality on real world data sets will be reviewed.

Restatement of the Purpose of the Study

The purpose of the study is to educate researchers and statisticians in education, psychology, and other disciplines in understanding the comparative power properties of the independent samples t-test and Wilcoxon test to violations from normality. Using Monte Carlo techniques of repeated

sampling with replacement from eight real data sets found in education and psychology will assist statisticians and researchers in determining the utility of past studies, but more important, heighten the awareness in the prevalence of nonnormal data in these fields, and ensure appropriate test application for future studies.

The following analysis will review the comparative power of a one tailed independent samples t-test and Wilcoxon Rank Sum test under eight real world distributions, four sample sizes (n), and eight treatment effects (c). These advantages will be defined by the rate of rejection of the test and the difference in the rejection rates between the tests. In addition, robustness to assess Type I error rates and the adequacy of the algorithms used in the Fortran program will also be reviewed. Both of Bradley's (1978) liberal and stringent definitions of robustness will be used.

Smooth Symmetric, Achievement

Figures 9-12 represent the comparative power rates of the independent samples t-test and Wilcoxon Rank Sum test (WRS) under the smooth symmetric achievement distribution when (n1,n2)= (10,10), (5,15), (30,30), and (15,45) respectively. A one tailed independent samples t-test and Wilcoxon Rank Sum test produced Type 1 error rates very close to nominal alpha for all sample sizes. Both tests met Bradley's liberal and stringent definitions of robustness (see Appendix A).

The smooth symmetric achievement distribution is similar to the normal distribution with kurtosis = 2.66 and a skew =

0.01. When assessing the comparative power of these two tests the WRS held modest power advantages in 25% (8 of 32) of the comparisons, with power differences ranging from .003 to .046. Four of these advantages occurred under all four distributions when c=.25σ. The largest power advantage occurred when (n1,n2)= (30,30) and c=.25σ. At this level the rate of rejection for the WRS was .205 as compared to .159 for the t-test.

The t-test held power advantages in 53% (17 of 32) of the comparisons. These power differences were also very modest, ranging from .001 to .088. The largest power advantages occurred when c= 1.00σ and (n1,n2)= (10,10), (5,15), and (15,45). At this level the rate of rejection was .56, .442, .912 for the t-test and .472, .363, and .829 for the WRS, respectively. In the remaining 22% (7 of 32) of the distributions neither test held an advantage.

Digit Preference, Achievement

The comparative power of the t-test and WRS under the digit preference achievement distribution is presented in figures 13-16. This distribution can also be considered relatively symmetric, with a kurtosis = 2.76 and a skew = -0.07. Both tests demonstrated robust results to Type I error under all four sample sizes, meeting the liberal and stringent definitions of robustness (see Appendix B).

The independent samples t-test demonstrated slight power advantages in 81% (26 of 32) of the comparisons, with power differences ranging from .001 to .075. The maximum power

advantage of .075 was obtained when (n1,n2)= (30,30) and c=.50σ. Under these conditions the t-test rejected at a rate of.48 and the WRS rejected at a rate of .40. In the remaining 19% (6 of 32) of the distributions the t-test or WRS held power advantages.

Mass At Zero, Achievement

Figures 17-20 represent the comparative power of the WRS and t-test under the mass at zero achievement distribution. This distribution is relatively symmetric as demonstrated by a kurtosis of 3.31 and a skew = -0.03. Again, both tests produced Type I error rates near nominal alpha under all four sample sizes, using Bradley's (1978) liberal and stringent definition of roubustness (see Appendix C).

Similar comparative power results occurring with digit preference, also occurred with mass at zero achievement. In 91% (29 of 32) of the comparisons the t-test demonstrated slight advantages over the WRS, with a power difference ranging from .001 to .116. Although modest, the largest advantage of .116 occurred when (n1,n2)= (30,30), and c=.50σ. At this level the t-test rejected at a rate of .377 as compared to .261 for the WRS. In the remaining 9% (3 of 32) of the distributions the t-test or WRS demonstrated power advantages.

Multimodal And Lumpiness, Achievement

Figures 21-24 represent the comparative power results of the t-test and WRS under a multimodal and lumpiness achievement distribution. As previously discussed, this

distribution has not been considered in past robustness or comparative power studies, until Sawilowsky and Blair (1992). This distribution demonstrated a kurtosis = 1.80 and skew = .19. The t-test and WRS demonstrated robustness when Type I error rates approached nominal alpha levels in both the liberal and stringent definitions of robustness (see Appendix D).

This distribution produced comparative power results similar to digit preference and mass at zero. The t-test held slight power advantages in 84% (27 of 32) of the comparisons, with a power difference ranging from .002 to .11. The largest advantage of .11 occurring when (n1,n2) = 5,15 and $c=1.50\sigma$. At this level the rate of rejection for the t-test was .778 as compared to .668 for the WRS.

The WRS held a power advantage in only 3% (1 of 32) of the comparisons, although the power difference was only .006. In the remaining 13% (4 of 32) of the distributions neither test held an advantage.

Extreme Bimodality, Psychometric

Figures 25-28 represent the comparative power of the t-test and WRS under an extreme bimodal psychometric distribution. This distribution demonstrated a skew = -0.08 and kurtosis = 1.30. Under all four sample sizes the t-test and WRS produced Type I error rates near nominal alpha (see Appendix E).

Unlike the previous distributions, extreme bimodality has produced the most extreme results thus far. The WRS showed

power advantages in 31% (10 of 32) of the comparisons, with power differences ranging from .001 to .28. Four of these power advantages occurred at all four sample sizes when $c=.25\sigma$. For example, when (n1,n2)= (30,30), the rate of rejection for the WRS was .434, as compared to .153 for the t-test, a power difference of .281. When (n1,n2)= (15,45), the power for both tests decreased, but the WRS still held moderate power advantages over the t-test, with a rate of rejection of .345, and for the t-test .126.

When (n1,n2)= (10,10), (5,15) and $c=.25\sigma$, $.50\sigma$, $.75\sigma$ the WRS held modest power advantages. The largest advantages occurred when (n1,n2)= (10,10) and $c=.75\sigma$. The rate of rejection for the WRS was .506 as compared to .326 for the t-test. With the uneven samples of n=20 the rate of rejection for the t-test was .262 as compared to .383 for the WRS. The remaining power advantages for the WRS were extremely modest.

In contrast, the t-test held power advantages in 63% (20 of 32) of the distributions, with power differences ranging from .007 to .349. The advantages for the t-test were modest except when (n1,n2)= (10,10), (5,15), and $c=1.50\sigma$ and 1.75σ. When (n1,n2)= (10,10), the rate of rejection at $c=1.50\sigma$ was .899 for the t-test and .706 for the WRS, a power difference of .193. At the same sample size and $c=1.75\sigma$ the rate of rejection for the t-test was .973 as compared to .707 for the WRS. For unequal sample sizes of (5,15) and $c=1.50\sigma$ and 1.75σ the power of both tests decreased, although the magnitude of the difference increased in favor of the t-test. The rate of

rejection for the t-test when c=1.50σ, and 1.75σ was .777 and .909, while the WRS showed a rate of rejection of .552 and .56, demonstrating power differences of .225 and .349 respectively.

Extreme Asymmetry, Psychometric

Figures 29-32 show the comparative power of the t-test and WRS under an extreme asymmetry psychometric distribution. This distribution has a kurtosis= 4.52 and skew = 1.64.

The robustness of the t-test to Type 1 error under this distribution, demonstrated rather liberal and conservative results (see Appendix F). When (n1,n2)= (10,10) the one tailed t-test produced slightly conservative results with actual alpha=.021, thus failing to meet the stringent definition of robustness. As the even samples increased to (30,30) the t-test produced rates near nominal alpha. With an uneven sample size = (5,15) the t-test failed to meet the liberal definition of robustness, with actual alpha = .038. As the sample size was increased to (15,45) actual alpha decreased to .031, thus meeting the liberal definition of robustness, but exceeding the stringent criterion.

The WRS test produced Type I error rates near nominal alpha except when (n1,n2)= (5,15). For this sample size an actual alpha =.029 exceeded the stringent definition of robustness (see Appendix F). Similar deviations demonstrating liberal and conservative variations and the potential causes for these variations will be discussed in the proceeding sections.

The extreme asymmetry psychometric distribution is the first distribution in this study to overwhelming favor the WRS. The WRS demonstrated power advantages in 69% (22 of 32) of the comparisons, and in many instances moderate to extreme advantages ranging from .002 to .618. When n=20, the WRS dominated the comparisons in small to medium effect sizes (.25σ to 1.25σ). For example when c=.25σ the rate of rejection for the WRS was .368 for the even samples and .266 for the uneven samples. In contrast the t-test rejected at a rate of .088 for even and .092 for the uneven samples, a power difference of .28 and .174 respectively. Under the same sample sizes, when c=.50σ the power differences for the WRS increased to .329 for the even samples and .249 for the uneven. As the effect size increased to c=.75σ and 1.00σ the WRS maintained rather liberal power advantages.

As the sample size was increased to n=60, the magnitude of the WRS power advantage increased. When (n1,n2) = 30,30 and c=.25σ,.50σ, and .75σ, the rate of rejection for the WRS was .783, .914, and .963. In contrast, the t-test rejected at a rate of .165, .358, and .659, a power difference of .618, .556 and .304 respectively. With (n1,n2) = (15,45) and c=.25σ, .50σ, and .75σ the power of the two tests decreased, but the WRS still held large power advantages with power differences of .58, .612, and .429. The remaining power advantages for the WRS very moderate.

Extreme Asymmetry, Achievement

Figures 33-36 represents the comparative power of the WRS

and t-test under the extreme asymmetry, achievement distribution. This distribution has a skew = 1.64 and kurtosis = 4.11.

The robustness of the WRS under this distribution produced Type I error rates near nominal alpha (see Appendix G). In retrospect, the t-produced results that were rather conservative when samples were uneven. When (n1,n2)= (5,15), actual alpha was equal to .008, clearly violating Bradley's stringent and liberal definition of robustness. Although when (n1,n2)= (15,45), actual alpha increased to .018. Again, this violated the stringent definition, but met the liberal criterion. The even samples produced results that were near nominal values.

When assessing the comparative power, the WRS demonstrated similar results as the extreme asymmetry psychometric distribution, although the power advantages under this distribution were moderate. The WRS held power advantages in 63% (20 of 32) of the comparisons, with power differences ranging from .002 to .187. Power advantages for the WRS were maintained in all samples when c=.25σ, .50σ, .75σ, 1.00σ and 1.25σ. The largest power advantage occurring when (n1,n2) = 30,30 and c=.25σ. At this level the rejection rate for the WRS was .345 as compared to .158 for the t-test.

The t-test held slight power advantages in 25% (8 of 32) of the distributions, with power differences ranging from .001 to .056. Six of these power advantages occurred when n=20 and c=1.50σ, 1.75σ and 2.00σ. A maximum power advantage was

obtained when (n1,n2) = (5,15) and c=2.00σ. The rate of rejection for the t-test was .942 and WRS .886, a difference of .056. The remaining power advantages for the t-test occurred when (n1,n2) = (15,45) and c=1.50σ and 1.75σ, although the advantages were minimal. The remaining 12% (4 of 32) of the power advantages neither the t-test or WRS held power advantages.

Mass at Zero With Gap, Psychometric

Figures 37-40 represent the comparative power of the WRS and t-test under a mass at zero with gap psychometric distribution. This distribution can be considered the most extreme with 80% of the scores accumulating at zero. This data set has a skew = 1.65 and kurtosis = 3.98.

The robustness of the t-test and WRS to Type I error under this distribution produced both liberal and conservative results (see Appendix H). When (n1,n2)= (10,10), the t-test and WRS failed to achieve the stringent or liberal definitions of robustness, with actual alpha = .006. As the equal sample sizes were increased to n=60, both tests obtained Type 1 error rates near nominal alpha.

The unequal sample sizes produced liberal results. When (n1,n2)= (5,15) and (15,45) actual alpha for the WRS was .031 and .032, and for the t-test .035 and .032 respectively. As shown, these results met the liberal definition of robustness, but failed to achieve the stringent criterion.

As demonstrated in the extreme asymmetry psychometric distribution and again in this distribution, the t-test

produced Type I error rates failing to meet the liberal and stringent definitions of robustness identified by Bradley (1978). As shown by Sawilowsky and Blair (1992), and replicated in this study, the t-test has been found to be nonrobust with extreme skews. In addition, the WRS being considered a distribution free test also produced three nonrobust results in these distributions. Probable causes for this discrepancy will be reviewed in the conclusion.

The WRS produced power advantages in 50% (16 of 32) of the comparisons and in many instances extreme power advantages in all four sample sizes when c=.25σ, .50σ, and .75σ. The magnitude of these advantages is as follows. When (n1,n2)= (10,10), (15,15), (30,30), (15,45) and c=.25σ, the WRS rejected at a rate of .70, .849, .996, and .964 respectively. In contrast, the t-test rejected at a rate of .036, .101, .20, and .161, clearly showing the most extreme power advantages in this study. As the effect size increased to .50, the WRS rejected at a rate of .72, .877, .997 and .963, while the t-test rejected at a rate of .14, .264, .605, and .475. When c=.75σ the WRS maintained the advantages, although the magnitude of the advantages decreased.

The t-test showed power advantages in all four sample sizes when the effect size was moderate to large (1.25σ to 2.00σ). The most extreme advantages occurring when (n1,n2)= (10, 10). The rejection rates for the t-test in accordance to the above effect sizes were .841, .953, .987, and .998. In contrast, the WRS rejected at a rate of .727, .82 .733 and

.795. When (n1,n2)= (5,15) and c=1.75σ the t-test detected its largest power advantage. Under these conditions the rejection rate for the t-test was .96 as compared to .595 for the WRS, a power difference of .365. The remaining advantages for the t-test were extremely moderate.

The WRS produced obscure robustness results under this distribution. In addition, the power of the WRS under the discrete mass at zero distribution, also demonstrated unusual fluxuations. For example, when (n1,n2)= (10,10) and c=.75σ the rejection rate of the WRS was .811. As the effect size increased to 1.00σ, the power rate decreased to .726. This type of decrease also occurred under the same sample size when c=1.75σ and again when (n1,n2)= (5,15) and c= 1.75σ. In theory for a distribution free test, as the effect size increases the power of the test should also increase. As previously discussed the t-test encountered its largest advantage under these unusual circumstances. Further discussion of these results will be review in the following section.

Conclusion

In the fields of education and psychology we must begin to understand the impact of statistical tests on research outcomes. The prevalence of nonnormality in these fields has been well documented, yet we are just beginning to understand the types of distributions these fields encompass and the impact of test application.

This study demonstrated the comparative power of the t-test and WRS under real world distributions. The results

indicate that the t-test was more powerful under distributions that were relatively symmetric, although the magnitude of these differences were minimal. In contrast, the WRS held more power advantages in distributions with extreme skews or heavy tails. The magnitude of these advantages were extremely large.

In those distributions (smooth symmetric, digit preference, mass at zero, and multimodal and lumpiness) considered relatively symmetric, the t-test maintained its reputation as being the Uniformly Most Powerful Unbiased Test under normal theory. Under these distributions the t-test held power advantages in 77% (99 of 128) of the comparisons, while the WRS held power advantages in only 7% (9 of 128). As shown in other studies and replicated in this study, the power advantages of the t-test in near symmetric distributions is extremely modest and in many instances near equivalent to the WRS. In fact, this study demonstrated that in only 1.56% (2 of 128) of the comparisons did the t-test exceed a power difference greater than .10 in relatively smooth distributions.

In the remaining data sets (extreme bimodal, extreme asymmetry achievement, extreme asymmetry psychometric and mass at zero with gap) the WRS held power advantages in 53% (68 of 128) of the comparisons, as compared to 37% (47 of 128) for the t-test. The impact of these results appear to be minimal, although the magnitude of the power differences overwhelmingly favors the WRS. For example, of the 68 power comparisons favoring the WRS under these four data sets, 65% (44) of the

power differences were greater than .10 as compared to 17% (8 of 47) for the t-test. In addition, 14 of the 44 differences in favor of the WRS were in the range of $.20 < x \leq .40$, while the t-test had only 5. In the more extreme comparisons the WRS had an additional 7 of the 44 ranging from $.50 < x \geq .60$, and 3 greater than .70. Overall, the WRS held 81% (44 of 54) of the power differences greater than .10.

Based on these results and other similar results conducted on mathematical distributions, it is recommended when the characteristics of a population are known to be relatively symmetric the t-test should be applied. When distributions consist of heavier tails or skews the WRS should be the test of choice. In turn, when population characteristics are unknown, the WRS is recommended because of the magnitude of the power differences in extreme skews, the modest variations in symmetric distributions, and the comparative power and robustness of the WRS to Type I and Type II errors in small, medium, and large effect sizes.

In addition to the comparative power differences, there were also robust discrepancies identified for both tests. According to the stringent definition of robustness identified by Bradley (1978), the t-test failed to produce Type I error rates near nominal alpha in 25% (8 of 32) of the distributions and sample sizes. Using the liberal definition of robustness the t-test failed in 9% (3 of 32) of the comparisons.

Sawilowsky and Blair (1992) demonstrated the nonrobustness of the t-test under distributions with extreme skew, therefore

the robustness results in this study were not surprising. The more complex issue is the nonrobust results generated by the WRS under the mass at zero with gap distribution.

As discussed in the methodology section, the WRS in this study was conducted by applying the independent samples t-test on the ranks of the original scores. Sawilowsky and Brown (1991), found that employing the t-test on ranks does not take into account a correction for ties or continuity, and in many situations reduces the power of the test. The mass at zero with gap distribution has 80% (519 of 648) of the scores accumulating at zero. It is believed this distinction not only influenced the results of the Type I error properties, but is a direct cause for the obscure variations in power under the mass at zero with gap data set.

In distributions with extreme ties, the Wilcoxon Rank Sum using the t-test on the ranks did not perform adequately. Therefore, it is recommended to use the Wilcoxon Rank Sum test on the original scores when extreme ties are present or distribution characteristics are unknown.

Additional Study Areas

This study assessed the comparative power of the t-test and WRS to violations from normality in eight real world distributions. Since the WRS was conducted by applying the independent samples t-test on the ranks of the original scores, further research could be undertaken by applying the original Wilcoxon Rank Sum test, under the same conditions. Further research can be conducted on the dependent samples

test to violations from normality, as well as violations to the assumption of hetrogenious variances for both the independent and dependent samples.

In addition, as identified by Micceri (1989), there is a prevalence of nonnormality under real world conditions. As more research is conducted similar to Micceri, there will be opportunities to compare the robustness and comparative power of these tests under new distributions. In fact, this study demonstrates the need to understand the wide array of distributions in other areas of education, psychology and other professions.

APPENDIX A

Smooth Symmetric, Achievement

Comparative Power Rates and Robustness for a One-tailed Independent-Samples t-test and Wilcoxon test for Alpha =.05

Statistic/Sample Size	Robustness	.025sd	.50sd	.75sd	1.00sd	1.25sd	1.50sd	1.75sd	2.00sd
10,10									
t-test	0.025	0.074	0.18	0.35	0.56	0.754	0.887	0.961	0.99
WRS	0.026	0.093	0.183	0.31	0.472	0.774	0.88	0.944	0.979
Power Difference	NA	-0.019	-0.003	0.04	0.088	-0.02	0.007	0.017	0.011
5,15									
t-test	0.024	0.066	0.147	0.274	0.442	0.624	0.785	0.894	0.957
WRS	0.024	0.079	0.147	0.242	0.363	0.641	0.764	0.86	0.925
Power Difference	NA	-0.013	0	0.032	0.079	-0.017	0.021	0.034	0.032
30,30									
t-test	0.024	0.159	0.482	0.814	0.969	0.998	1	1	1
WRS	0.025	0.205	0.47	0.743	0.92	0.998	1	1	1
Power Difference	NA	-0.046	0.012	0.071	0.049	0	0	0	0
15,45									
t-test	0.027	0.126	0.375	0.694	0.912	0.985	0.999	1	1
WRS	0.026	0.159	0.366	0.617	0.829	0.987	0.998	1	1
Power Difference	NA	-0.033	0.009	0.077	0.083	-0.002	0.001	0	0

Note: a negative power difference demonstrates a power advantage for the WRS

APPENDIX B

Digit Preference, Achievement

Comparative Power Rates and Robustness for a One-tailed Independent-Samples t-test and Wilcoxon test for Alpha =.05

Statistic/Sample Size	Robustness	.025sd	.50sd	.75sd	1.00sd	1.25sd	1.50sd	1.75sd	2.00sd
10,10									
t-test	0.025	0.077	0.18	0.351	0.562	0.753	0.889	0.959	0.99
WRS	0.025	0.064	0.16	0.328	0.539	0.738	0.882	0.956	0.988
Power Difference	NA	0.013	0.02	0.023	0.023	0.015	0.007	0.003	0.002
5,15									
t-test	0.025	0.067	0.145	0.28	0.451	0.626	0.785	0.897	0.955
WRS	0.025	0.058	0.129	0.259	0.424	0.603	0.766	0.884	0.948
Power Difference	NA	0.009	0.016	0.021	0.027	0.023	0.019	0.013	0.007
30,30									
t-test	0.025	0.157	0.48	0.814	0.968	0.997	1	1	1
WRS	0.025	0.114	0.405	0.77	0.954	0.996	1	1	1
Power Difference	NA	0.043	0.075	0.044	0.014	0.001	0	0	0
15,45									
t-test	0.024	0.13	0.374	0.693	0.908	0.985	0.998	1	1
WRS	0.025	0.098	0.314	0.643	0.881	0.978	0.998	1	1
Power Difference	NA	0.032	0.06	0.05	0.027	0.007	0	0	0

Note: a negative power difference demonstrates a power advantage for the WRS

APPENDIX C

Mass At Zero, Achievement

Comparative Power Rates and Robustness for a One-tailed Independent-Samples t-test and Wilcoxon test for Alpha =.05

Statistic/Sample Size	Robustness	.025sd	.50sd	.75sd	1.00sd	1.25sd	1.50sd	1.75sd	2.00sd
10,10									
t-test	0.025	0.065	0.151	0.279	0.449	0.629	0.779	0.888	0.951
WRS	0.027	0.043	0.113	0.229	0.396	0.58	0.746	0.868	0.94
Power Difference	NA	0.022	0.038	0.05	0.053	0.049	0.033	0.02	0.011
5,15									
t-test	0.026	0.058	0.12	0.221	0.354	0.51	0.661	0.79	0.881
WRS	0.025	0.041	0.092	0.18	0.304	0.461	0.618	0.754	0.854
Power Difference	NA	0.017	0.028	0.041	0.05	0.049	0.043	0.036	0.027
30,30									
t-test	0.025	0.126	0.377	0.689	0.905	0.982	0.999	1	1
WRS	0.025	0.065	0.261	0.584	0.857	0.97	0.997	1	1
Power Difference	NA	0.061	0.116	0.105	0.048	0.012	0.002	0	0
15,45									
t-test	0.025	0.105	0.294	0.572	0.811	0.942	0.989	0.998	1
WRS	0.024	0.055	0.204	0.474	0.746	0.918	0.983	0.997	1
Power Difference	NA	0.05	0.09	0.098	0.065	0.024	0.006	0.001	0

Note: a negative power difference demonstrates a power advantage for the WRS

APPENDIX D

Multimodal and Lumpiness, Achievement

Comparative Power Rates and Robustness for a One-tailed Independent-Samples t-test and Wilcoxon test for Alpha =.05

Statistic/Sample Size	Robustness	.025sd	.50sd	.75sd	1.00sd	1.25sd	1.50sd	1.75sd	2.00sd
10,10									
t-test	0.025	0.073	0.173	0.338	0.543	0.75	0.893	0.968	0.994
WRS	0.025	0.073	0.179	0.331	0.502	0.664	0.797	0.905	0.969
Power Difference	NA	0	-0.006	0.007	0.041	0.086	0.096	0.063	0.025
5,15									
t-test	0.025	0.065	0.145	0.265	0.426	0.616	0.778	0.902	0.965
WRS	0.023	0.06	0.14	0.246	0.377	0.525	0.668	0.812	0.916
Power Difference	NA	0.005	0.005	0.019	0.049	0.091	0.11	0.09	0.049
30,30									
t-test	0.025	0.153	0.473	0.817	0.971	0.999	1	1	1
WRS	0.026	0.137	0.45	0.768	0.933	0.987	0.998	1	1
Power Difference	NA	0.016	0.023	0.049	0.038	0.012	0.002	0	0
15,45									
t-test	0.025	0.128	0.372	0.691	0.912	0.988	0.999	1	1
WRS	0.024	0.114	0.352	0.648	0.863	0.963	0.993	0.994	1
Power Difference	NA	0.014	0.02	0.043	0.049	0.025	0.006	0.006	0

Note: a negative power difference demonstrates a power advantage for the WRS

APPENDIX E

Extreme Bimodality, Psychometric

Comparative Power Rates and Robustness for a One-tailed Independent-Samples t-test and Wilcoxon test for Alpha =.05

Statistic/Sample Size	Robustness	.025sd	.50sd	.75sd	1.00sd	1.25sd	1.50sd	1.75sd	2.00sd
10,10									
t-test	0.025	0.073	0.168	0.326	0.542	0.744	0.899	0.973	0.996
WRS	0.025	0.174	0.171	0.506	0.51	0.704	0.706	0.707	0.967
Power Difference	NA	-0.101	-0.003	-0.18	0.032	0.04	0.193	0.266	0.029
5,15									
t-test	0.023	0.064	0.139	0.262	0.43	0.611	0.777	0.909	0.975
WRS	0.024	0.14	0.14	0.383	0.388	0.555	0.552	0.56	0.926
Power Difference	NA	-0.076	-0.001	-0.121	0.042	0.056	0.225	0.349	0.049
30,30									
t-test	0.026	0.153	0.47	0.818	0.971	0.998	0.999	1	1
WRS	0.027	0.434	0.432	0.926	0.925	0.991	0.991	0.991	1
Power Difference	NA	-0.281	0.038	-0.108	0.046	0.007	0.008	0.009	0
15,45									
t-test	0.024	0.126	0.372	0.694	0.914	0.987	0.999	1	1
WRS	0.025	0.345	0.342	0.836	0.838	0.963	0.964	0.964	1
Power Difference	NA	-0.219	0.03	-0.142	0.076	0.024	0.035	0.036	0

Note: a negative power difference demonstrates a power advantage for the WRS

APPENDIX F

Extreme Asymmetry, Psychometric

Comparative Power Rates and Robustness for a One-tailed Independent-Samples t-test and Wilcoxon test for Alpha =.05

Statistic/Sample Size	Robustness	.025sd	.50sd	.75sd	1.00sd	1.25sd	1.50sd	1.75sd	2.00sd
10,10									
t-test	0.021	0.088	0.167	0.295	0.58	0.749	0.872	0.948	0.983
WRS	0.025	0.368	0.496	0.593	0.735	0.833	0.872	0.934	0.957
Power Difference	NA	-0.28	-0.329	-0.298	-0.155	-0.084	0	0.014	0.026
5,15									
t-test	0.038	0.092	0.142	0.232	0.463	0.63	0.784	0.899	0.955
WRS	0.029	0.266	0.391	0.491	0.661	0.783	0.842	0.917	0.948
Power Difference	NA	-0.174	-0.249	-0.259	-0.198	-0.153	-0.058	-0.018	0.007
30,30									
t-test	0.025	0.165	0.358	0.659	0.964	0.997	0.999	1	1
WRS	0.025	0.783	0.914	0.963	0.993	0.999	0.999	1	1
Power Difference	NA	-0.618	-0.556	-0.304	-0.029	-0.002	0	0	0
15,45									
t-test	0.031	0.136	0.275	0.526	0.913	0.987	0.999	1	1
WRS	0.026	0.716	0.887	0.955	0.993	0.999	0.999	1	1
Power Difference	NA	-0.58	-0.612	-0.429	-0.08	-0.012	0	0	0

Note: a negative power difference demonstrates a power advantage for the WRS

APPENDIX G

Extreme Asymmetry, Achievement

Comparative Power Rates and Robustness for a One-tailed Independent-Samples t-test and Wilcoxon test for Alpha =.05

Statistic/Sample Size	Robustness	.025sd	.50sd	.75sd	1.00sd	1.25sd	1.50sd	1.75sd	2.00sd
10,10									
t-test	0.023	0.079	0.201	0.378	0.574	0.753	0.877	0.952	0.984
WRS	0.024	0.14	0.269	0.542	0.653	0.823	0.876	0.949	0.967
Power Difference	NA	-0.061	-0.068	-0.164	-0.079	-0.07	0.001	0.003	0.017
5,15									
t-test	0.008	0.052	0.162	0.32	0.492	0.656	0.786	0.883	0.942
WRS	0.024	0.142	0.249	0.461	0.554	0.706	0.765	0.857	0.886
Power Difference	NA	-0.09	-0.087	-0.141	-0.062	-0.05	0.021	0.026	0.056
30,30									
t-test	0.024	0.158	0.484	0.814	0.965	0.997	1	1	1
WRS	0.025	0.345	0.652	0.949	0.985	0.999	1	1	1
Power Difference	NA	-0.187	-0.168	-0.135	-0.02	-0.002	0	0	0
15,45									
t-test	0.018	0.129	0.394	0.704	0.902	0.98	0.997	1	1
WRS	0.025	0.287	0.527	0.843	0.918	0.982	0.992	0.999	1
Power Difference	NA	-0.158	-0.133	-0.139	-0.016	-0.002	0.005	0.001	0

Note: a negative power difference demonstrates a power advantage for the WRS

APPENDIX H

Mass At Zero With Gap, Psychometric

Comparative Power Rates and Robustness for a One-tailed Independent-Samples t-test and Wilcoxon test for Alpha =.05

Statistic/Sample Size	Robustness	.025sd	.50sd	.75sd	1.00sd	1.25sd	1.50sd	1.75sd	2.00sd
10,10									
t-test	0.006	0.036	0.14	0.428	0.66	0.841	0.953	0.987	0.998
WRS	0.006	0.7	0.72	0.811	0.726	0.727	0.82	0.733	0.795
Power Difference	NA	-0.664	-0.58	-0.383	-0.066	0.114	0.133	0.254	0.203
5,15									
t-test	0.035	0.101	0.264	0.467	0.681	0.854	0.952	0.96	0.999
WRS	0.031	0.849	0.877	0.879	0.88	0.881	0.883	0.595	0.943
Power Difference	NA	-0.748	-0.613	-0.412	-0.199	-0.027	0.069	0.365	0.056
30,30									
t-test	0.024	0.2	0.605	0.914	0.993	0.999	1	1	1
WRS	0.024	0.996	0.997	0.997	0.997	0.997	0.997	0.998	0.999
Power Difference	NA	-0.796	-0.392	-0.083	-0.004	0.002	0.003	0.002	0.001
15,45									
t-test	0.032	0.161	0.475	0.82	0.974	0.999	0.999	1	1
WRS	0.032	0.964	0.963	0.965	0.964	0.964	0.964	0.999	0.966
Power Difference	NA	-0.803	-0.488	-0.145	0.01	0.035	0.035	0.001	0.034

Note: a negative power difference demonstrates a power advantage for the WRS

REFERENCES

Akritas, M.G. (1991). Limitations of the rank transform procedure: A study of repeated measures designs, part 1. Journal of the American Statistical Association, 86, 457-460.

Anderson, N.H. (1961). Scales and statistics: Parametric and nonparametric. Psychological Bulletin, 58, 305-316.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., & Tukey, J.W. (1972). Robust estimates of location survey and advances. Princeton, NJ: Princeton University Press.

Baker, B.O., Hardych, C.D., & Petrinovich, L.F. (1966). Weak measurement versus strong statistics: An empirical critique of S.S Stevens' proscriptions on statistics. Educational and Psychological Measurement, 26, 219-309.

Blair, R.C. (1981). A reaction to "consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 51(4), 499-507.

Blair, R.C., & Higgins, J.J. (1980). A comparison of the power of the Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions. Journal of Educational Statistics, 5(4), 309-335.

Blair, R.C., & Higgins, J.J. (1981). A note on the asymptotic relative efficiency of the Wilcoxon rank-sum test relative to the independent means t test under mixtures of two normal distributions. British Journal of Mathematical and Statistical Psychology, 31, 124-128.

Blair, R.C., & Higgins, J.J. (1985). Comparison of the power of the paired samples t-test to that of Wilcoxon's signed-ranks tests under various populating shapes. Psychological Bulletin, 97, 119-128.

Blair, R.C., Higgins, J.J., & Smitley W.D.S. (1980). On the relative power of the U and t-tests. British Journal of Mathematical and Statistical Psychology, 33, 114-120.

Blair, R.C., Sawilowsky, S.S., & Higgins, J.J. (1987). Limitations of the rank transform in tests for interaction. Communications in Statistics: Computation and Simulation, B16, 1133-1145.

Boneau, C.A. (1960). The effects of violations of assumptions underlying the t-test. Psychological Bulletin, 57, 49-64.

Boneau, C.A. (1961). A note on measurement scales and statistical tests. American Psychologist, 16, 160-261.

Boneau, C.A. (1962). A comparison of the power of the U and t-tests. Psychological Review, 69, 246-256.

Bradley, J.V. (1968). Distribution-free statistical tests. Englewood Cliffs, NJ:Prentice Hall.

Bradley, J.V. (1977). A common situation conducive to bizarre distribution shapes. American Statistician, 31, 147-150.

Bradley, J.V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Bradley, J.V. (1980a). Nonrobustness in classical tests on means and variances: A large-scale sampling study. Bulletin of the Psychonomics Society, 15, 275-278.

Bradley, J.V. (1980b). Nonrobustness in one-sample Z and t tests: A large-scale sampling study. Bulletin of the Psychonomics Society, 15, 29-32.

Bradley, J.V. (1980c). Nonrobustness in Z, t, and F tests at large sample sizes. Bulletin of the Psychonomics Society, 16, 333-336.

Bradley, J.V. (1982). The insidious L-shaped distribution. Bulletin of the Psychometrics Society, 20(2), 85-88.

Brewer, J.K. (1972). On the power of statistical tests in the American Educational Research Journal. American Educational Research Journal, 9, 391-401.

Chernoff, H., & Savage, I.R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. Annals of Mathematical Statistics, 29, 972-999.

Cochran, W.G. (1947). Some consequences when the assumption for the analysis of variances are not satisfied. Biometric, 3, 27-38.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. Journal of Abnormal and Social Psychology, 65, 145-153.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112 (1), 155-159.

Conover, W.J., & Iman, R.L. (1981). Rank transformations as a

bridge between parametric and nonparametric statistics. American Statistician, 35, 124-129.

Dixon, W.J. (1954). Power under normality of several nonparametric tests. Annals of Mathematical Statistics, 25, 610-614.

Friedman, J.A., Chalmers, T.C., Smith, H., & Kuebler, R.R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. New England Journal of Medicine, 299, 690-694.

Gaito, J. (1986). Some issues in the measurement-statistics controversy. Canadian Psychology, 27, 63-68.

Geary, R.C. (1947). Testing for normality. Biometrika, 34, 209-242.

Gibbons, J.D. (1985). Nonparametric methods for quantitative analysis (2nd ed.). Columbus, OH: American Sciences.

Gibbons, J.D., & Chakraborti, S. (1991). Comparisons of the Mann-Whitney, Student's t, and Alternate t tests for means of normal distributions. Journal of Experimental Education, 59, 258-267.

Gibbons, J.D., & Chakraborti, S. (1992). Response to Zimmerman. Journal of Experimental Education, 60(4), 365-366.

Glass, G., Peckham, P., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 42, 237-288.

Hack, H.R.B. (1958). An empirical investigation into the distribution of the F-ratio in samples from two nonnormal populations. Biometrika, 45, 260-265.

Hajek, J., & Sidak, Z. (1967). Theory of rank tests. New York: Academic Press.

Harwell, M.R (1990). A general approach to hypothesis testing for nonparametric tests. Journal of Experimental Education, 58(2), 143-156.

Harwell, M.R. (1990). Summarizing Monte Carlo results in methodological research. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Hill, M., & Dixon, W.J. (1982). Robustness in real life: A

study of clinical laboratory data. Biometrics, 38, 377-396.

Hoaglin, D.C., Mosteller, F., & Tukey, J.W. (1983). Understanding robust and exploratory data analysis. New York: Wiley.

Hodges, J.C., & Lehmann, E.L. (1956). The efficiency of some nonparametric competitors of the t-test. Annals of Mathematical Statistics, 27, 324-335.

Hoeffding, W. (1952). The large sample power of tests based on permutations of observations. Annals of Mathematical Statistics, 23, 169-172.

Hsu, T.C., & Feldt, L.S. (1969). The effect of limitations on the number of criterion score values on the significance level of the F-test. American Educational Research Journal, 6, 515-527.

Hunter, M.A., & May, R.B. (1993). Some myths concerning parametric and nonparmetric tests. Canadian Psychology, 34 (4) 384-389.

International Mathematical and Statistical Libraries. (1987). IMSL library reference manual (10th ed.) Houston, TX: Author.

Ito, P.K. (1980). Robustness of ANOVA and MANOVA test procedures. In P.R. Krishnaiah (Ed.) Handbook of Statistics 6, 199-236.

Kelly, D.L. (1994). The comparative power of several nonparametric alternatives to the analysis of variance test for interaction in a 2 x 2 x 2 layout. Unpublished doctoral dissertation, Wayne State University.

Kendall, M.G., & Buckland W.R. (1981). A Dictionary of Statistical Terms (4th ed.). Edinburgh: Oliver and Boyd.

Kerlinger, F.N.(1964). Foundations of behavioral research. New York: Holt, Reinehart, & Winston.

Kerlinger, F.N.(1973). Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart, & Winston.

Kraemer, H.C., & Thiemann, S. (1987). How many subjects?: statistical power analysis in research. Newbury Park: Sage Publications.

Lehmann, E.L. (1975). Nonparametrics. San Francisco: Holden-Day.

Lehmann, E.L., & Stein, C. (1959). Testing statistical

hypothesis. New York: John Wiley.

Linquist, E.F. (1953). Design and analysis of experiments in education and psychology. Boston: Houghton Mifflin.

Lord, F.M. (1953). On the statistical treatment of football numbers. American Psychologist, 8, 750-751.

Mansfield, E. (1986). Basic statistics with applications. New York: W.W. Norton & Company.

Meddis, R. (1984). Statistics using ranks: A unified approach. New York: Basil Blackwell Inc.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105(1), 156-166.

Neave, H.R., & Granger, C.W.J. (1968). A Monte Carlo study comparing various two-sample tests for differences in mean. Technometrics, 10, 509-522.

Nunnally, J. (1975). Introduction to statistics for psychology and education. New York: McGraw-Hill.

Nunnally, J. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Pearson, K. (1895). Contributions to the mathematical theory of evolution: II. Skew variation in homogeneous material. Philosophical Transactions of the Royal Society, Ser.A,186, 343-414.

Pearson, E.S., & Please, N.W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. Biometrika, 62(2), 223-241.

Penfield, D.A. (1994). Choosing a Two-Sample Location Test. Journal of Experimental Education, 62(4), 343-360.

Pratt, J. W., Gibbons J.D. (1981). Concepts of nonparametric theory. New York: Springer-Verlag.

Randles, R.H., & Wolfe, D.A. (1979). Introduction to the theory of nonparametric test. New York: John Wiley.

Rey, J.J. (1983). Introduction to robust and quasi-robust statistical methods. Berlin, Germany: Springer-Verlag.

Rider, P.R. (1929). On the distribution of the ratio of mean to standard deviation in small samples from nonnormal populations. Biometrika, 21, 124-143.

Runyon, R.P., & Haber, A. (1991). Fundamentals of behavioral

statistics (7th ed.). New York: McGraw-Hill.

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

Sawilowsky, S.S., & Brown, M.T. On using the t-test on ranks as an alternative to the Wilcoxon test. Perceptual and Motor Skills, 72, 860-862.

Sawilowsky, S.S. (1993). Comments on using alternatives to normal theory statistics in social and behavioural 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 to departures from population normality. Psychological Bulletin, 111(2), 352-360.

Sawilowsky, S.S., Blair, R.C., & Higgins, J.J. (1989). An investigation of the type 1 error and power properties of the rank transform procedure in factorial ANOVA. Journal of Educational Statistics, 1(3), 255-267.

Senders, V.L. (1958). Measurement and statistics. New York: Oxford University Press.

Scheffe', H. (1959). The analysis of variance. New York: Wiley.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill.

Siegel, S., & Castellan, J.N. (1988). Nonparametric statistics (2nd ed.). New York: McGraw-Hill.

Stevens, S.S. (1946). On the theory of scales of measurement. Science, 130, 677-680.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. Handbook of experimental psychology. New York: Wiley.

Stigler, S.M (1977). Do robust estimators work with real data? The Annals of Statistics, 5, 1055-1098.

Still, A. W., & White, A.P. (1981). The approximate randomization test as an alternative to the F test in analysis of variance. British Journal of Mathematical and Statistical Psychology, 34, 243-252.

Tan, W.Y. (1982). Sampling distributions and robustness of t, F, and variance-ratio in two samples and ANOVA models with respect to departures from normality. Communications in Statistics, A11, 2485-2511.

Tapia, R.A., & Thompson, J.R. (1978). Nonparametric probability density estimation. Baltimore,MD: John Hopkins University Press.

Thompson, G.L. (1991). A unified approach to rank tests for multivariate and repeated measures designs. Journal of the American Statistical Association, 86(414), 410-419.

van den Brink, W.P., & van den Brink S.G.J. (1989). A comparison of the power of the t-test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem. British Journal of Mathematical and Statistical Psychology, 42, 183-189.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics, 1, 80-83.

Zimmerman, D.W. (1987). Comparative power of Student T test and Mann-Whitney U test for unequal sample sizes and variances. Journal of Experimental Education, 55, 171-174.

Zimmerman, D.W., & Zumbo, B.D. (1990). Effects of outliers on the relative power of parametric and nonparametric statistical tests. Perceptual and Motor Skills, 71, 339-349.

Zimmerman, D.W. (1991). Failure of the Mann-Whitney Test: A note on the simulation study of Gibbons and Chakraborti. Journal of Experimental Education, 60, 359-364.

Zumbo, B.D., & Zimmerman, D.W. (1991). Levels of measurement and the relation between parametric and nonparametric statistical tests: a review of recent findings. Working paper 91-1, Edumetrics Research Group, University of Ottawa.

Zumbo, B.D., & Zimmerman, D.W. (1993a). Introduction to the Symposium. Canadian Psychology, 34(4), 381-383.

Zumbo, B.D., & Zimmerman, D.W. (1993b). Postscript: Some closing comments on "alternatives to classical statistics" (normal theory). Canadian Psychology, 34(4), 441-445.

ABSTRACT

THE COMPARATIVE POWER OF THE INDEPENDENT-SAMPLES T-TEST AND WILCOXON RANK SUM TEST IN NON NORMAL DISTRIBUTIONS OF REAL DATA SETS IN EDUCATION AND PSYCHOLOGY

by

PATRICK DAVID BRIDGE
May 1996

Advisor: Shlomo Sawilowsky

Major: Theoretical Evaluation and Research

Degree: Doctor of Philosophy

Historically, small samples Monte Carlo studies testing the robustness and comparative power properties of the independent-samples t-test and Wilcoxon Rank Sum test have been restricted to known mathematical distributions. Recently, the prevalence of nonnormally distributed data sets has been recognized in the fields of education and psychology. This, in turn, has generated a need in understanding appropriate test application under these conditions. Using Monte Carlo techniques, the purpose of this study was to assess the Type I error properties and comparative power of the Wilcoxon Rank Sum test (algebraic equivalent: t-test on the ranks of original scores) and the independent samples t-test to violations of normality. The sampling is from eight real distributions in eduction and psychology which were identified in a study by Micceri (1989). Sample sizes (n1,n2)= (10,10), (5,15), (30,30), and (15,45) were used, with nominal alpha set at .05. Eight treatment effects ranging from .25σ to 2.00σ were used for each distribution and sample size to measure

shift in location parameters.

Comparative power results demonstrated that in those distributions considered relatively symmetric, the t-test maintained its reputation as being the Uniformly Most Powerful Unbiased test under normal theory. Yet, the power advantages were extremely modest and in most instances near equivalent to the Wilcoxon Rank Sum test. When the distribution demonstrated extreme skews or heavy tails, the power advantages overwhelmingly favored the Wilcoxon Rank Sum test.

The t-test generated nonrobust results to Type I error in 25% (8 of 32) of the distributions and sample sizes studied, with most occurring in the distributions with extreme skews. In addition, the WRS also demonstrated nonrobust results and obscure power results in a distribution characterizing extreme ties.

When ever possible use the original WRS to ensure corrections for ties and continuity are applied. If the original WRS cannot be used and distribution characteristics known to have heavy tails or extreme skews, it is recommended to use the WRS applying the t-test on the ranks. In turn, when population characteristics are unknown, the WRS is recommended because of the magnitude of the power differences in extreme skews, the modest variations in symmetric distributions, and the comparative power and robustness of the WRS to Type I and Type II errors in small, medium, and large effect sizes. If distribution characteristics of a population are known to be relatively symmetric the t-test should be the test of choice.

AUTOBIOGRAPHICAL STATEMENT
Patrick D. Bridge

WORK EXPERIENCE

1995 to-present	**SelectCare Inc., Troy, Mi** Senior Research Administrator
1993-1995	**Sinai Hospital, Detroit, Mi** Quality Improvement Specialist/Biostatistician
1991-1992	**Burns International Security, Taylor, Mi** Branch Manager
1989-1990	**Kids "R" Us Distribution Center, Southgate, Mi** Security Manager

Education

1991-present	**Wayne State University, Detroit, Michigan** Doctoral Candidate (ABD) Major: Theoretical Evaluation and Research
1987-1990	**University of Detroit, Detroit, Michigan** **Masters of Science** Major: Security Administration
1981-1986	**Eastern Michigan University, Ypsilanti, Michigan** **Bachelor of Science** Major: Criminal Justice/Sociology

Selected Publications

Rosenberg, M.K., Bridge, P.D., Brown, M. 1994. Cost comparison: A desflurane versus a propofol based general anesthetic technique. Journal of Anesthesiology and Analgesia, Vol 79, pp. 852-855.

Rosenberg, M.K., Raymond, C., Bridge, P.D. 1995. Comparison of midazolam/ketamine with methohexital for sedation during peribulbur block. Journal of Anesthesiology and Analgesia, Vol 81, pp. 173-175.

Cohen, G.I., White, M., Sochowski, R.A., Bridge, P.D., Klein A.L., Stewart W.J., Chan K. 1995. Reference values for normal transesophageal echocardiographic measurements. Journal of the American Society of Echocardiologists, Vol 8, (3), pp. 221-230.