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

COMPARATIVE PROPERTIES OF NONPARAMETRIC STATISTICS FOR THE ANALYSIS OF THE 2 x c LAYOUT FOR ORDINAL CATEGORICAL DATA

by

MARGARET A. POSCH

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 S. Sawilowsky 9/26/96
Advisor Date

Jo Anne Halbert 9/26/96

9/26/96

DEDICATION

This paper is dedicated to the memory of my mother. Her own determination and appreciation for knowledge were qualities which I admired and her support of my own pursuit of education was appreciated. She was a loving mother and a good friend.

ACKNOWLEDGMENTS

It is with deep gratitude that I mention the names of the following people. Their contribution of time and effort, as well as their support of this work, warrants particular notice:

Dr. Shlomo Sawilowsky, my major advisor, who provided me with direction and gave me constant support throughout the writing process of this document. His personal attention during the final stages of my doctoral program made this degree a reality.

Dr. Joseph Posch, Jr., my husband, whose suggestion to pursue this degree and whose encouragement, persistence, and support when discouragement prevailed gave me the strength and determination to finish this project.

Dr. Alan Hoffman, Dr. JoAnne Holbert and Dr. Larry Miller, who found time to serve on my committee, when their schedules were already laden with commitments to their own students.

Dr. Donald Marcotte, whose classes were engrossing, inspiring and comprehensible, kindling within me a desire to pursue this field, in particular.

All those at Wayne State University who enabled me to overcome obstacles, both great and small.

Mr. James Kollar, my father, who taught me the challenge to finish what I start, a lesson for which I will always be grateful; it is the underpinning of this degree.

My beautiful children, Joseph, III, David, Jean, Michael, and Christina, whose patience and undying love enabled me to complete this goal without any major domestic calamities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

Introduction

Behavioral research is concerned with making appropriate statistical inferences about a given population through evidence provided by samples. For example, consider two independent samples, where the problem becomes determining whether both samples were drawn from the same population. There is an on-going debate among applied researchers, however, regarding the statistical approach to data analysis for educational and clinical samples. Most real-world studies involve data from nonnormally distributed populations (Bradley, 1977; Micceri, 1989); and yet classical parametric statistics with the assumption of population normality are applied to these situations. The reliance on distribution and other population characteristics marks these procedures as parametric (Kerlinger, 1973).

Nonparametric techniques, however, do not require assumptions about the underlying population shapes from which the data are obtained. These are called nonparametric tests, and they are strong competitors to their parametric counterparts in certain situations, as noted by Hollander and Wolfe (1973):

> More often than not, the nonparametric procedures are only slightly less efficient than their normal theory competitors when the underlying populations are normal (the home court of normal theory methods), and they can be mildly and wildly more efficient than these competitors when the underlying populations are not normal. (p.1)

The use of nonparametric statistics increased in the 1950s and 1960s, but their popularity declined thereafter. The trend toward distribution-free statistics awakened the suspicions of some practitioners that valuable information was lost when data were converted to ranks. Anderson (1961), Blair and Higgins (1980), and Sawilowsky (1990) summarized the reasons for the post-1960s decline of the nonparametric approach to

analysis of social science data. First, parametric tests were considered extremely robust by researchers relative to underlying population nonnormality (Boneau, 1960; Glass, Peckham, & Sanders, 1972), obviating the need for nonparametric tests. Second, alternative choices were not necessary; there was an assumption that parametric tests were more powerful than their nonparametric counterparts (Kerlinger, 1964, 1973; Nunnally, 1975) seemingly without regard for the underlying distribution of the samples. Third, there were few nonparametric alternatives for complicated research designs (Bradley, 1968). Eventually, researchers recognized that the best statistical approach for data analysis may change with the shape of underlying distributions. Consequently, over the last two decades, nonparametric techniques have been explored, developed and applied in certain research situations. Ongoing studies using Monte Carlo methods to compare the power and efficiency of parametric and nonparametric statistical tests help to resolve data analysis issues in research designs chosen by practitioners in educational and psychological research. In addition, the advancement of computer software has contributed to the changed paradigm of researchers regarding their statistical approach to data analysis.

Computers and Statistical Analysis

The mini-computers with advanced operating systems manufactured by Data General and Wang in the 1960s and 1970s supported intermediary languages like FORTRAN, PASCAL and Business BASIC. Consequently, sophisticated statistical programs were developed and had a revolutionary impact on behavioral and educational research. Kerlinger (1973) expressed the importance of computers to behavioral research:

> The days of statistical and mathematical computational drudgery are over: the computer now does in minutes and seconds statistical and other operations that took days, weeks, and even months of clerical and desk calculator work. Research projects that would not have been attempted ten years ago because of the sheer bulk of

> necessary calculations to analyze the data of the projects are now readily approachable with the computer and computer auxiliary equipment (p.705).

The availability of personal computers with increased power in the late 1980s provided convenience to the researcher and statistical packages were produced to facilitate computations of statistical tests which heretofore were considered too complex to compute manually with ease and accuracy. For example, permutation tests (and randomization tests), could be executed with high-speed and accuracy and easy ability to repeat operations thousands of times. Bradley (1968) called permutation tests "stunningly efficient", and Blair, Higgins and Smitley (1980) showed that the permutation test is indeed a powerful and efficient test for applied research situations.

Permutation Statistics

Sawilowsky (1990) noted that for sample sizes in applied research, all possible permutations are difficult to calculate. He remarked, however, that a rank order of observations, which "results in rank tests that maintain the properties of the parent permutation test in being nonparametric exact tests, ...are often easy to compute" (p.94). For example, the Wilcoxon Rank Sum (WRS) test, also known as the Mann-Whitney *U* test for independent samples, a member of the family of permutation tests, is simple to compute and is accessible to researchers through sophisticated software packages.

The Mann-Whitney *U* test was introduced in 1947 (the WRS was developed two years earlier) to test the null hypothesis that *X* and *Y* scores come from the same population, H_0: $f(x) = g(x)$. This is a rank sum test and computes the positive differences for all possible pairs of *X* and *Y* scores (Mann & Whitney, 1947). Researchers in education and psychology who employ real-world data prefer to use the Mann-Whitney *U* test to the Chi Square Test because it takes into account the magnitude of the outcomes. Superiority of outcomes is important for making recommendations to clients or patients.

An article by Grissom (1994) recommended the use of the Mann-Whitney *U* test over the Chi Square test for analysis of therapy outcomes in a 2 x c layout using ordinal categorical data for the above cited reason. Grissom noted, "(the Mann-Whitney *U* test) addresses the clinically useful question of whether one or the other therapy produces a superior result on the graded scale of clinically meaningfully categorized outcomes" (p.282). Grissom's suggestion to hand-calculate the test, however, was a tedious approach, subject to error, and obsolete. Advancements in computer technology have made the Mann-Whitney *U* test and many other nonparametric statistics, including exact permutations tests, feasible alternatives for analyzing data in behavioral and social science research quickly and efficiently.

Research Problem

Given the increased use of computers in statistical research, investigators should consider alternative computerized statistical techniques that may have been overlooked or avoided in the past due to their complex nature. Grissom's recommendation to use the Mann-Whitney *U* test for testing the superiority of therapy outcomes brings the researcher up to 1947. This study investigates some alternatives available for use with the computer which will bring the researcher up to 1996.

This study examines the comparative properties of four nonparametric tests for various sample sizes of ordinal categorical data found in the context of applied psychological and educational settings. Specifically, the investigation will compare several competitors on independent samples with a variety of ordinal outcome levels, or *c*, encountered in ordinal categorical data representative of real-world settings. The competing tests include the Mann-Whitney *U* Test (WRS), the Random Normal Scores Test, the Savage Scores Test, and the Permutation Test. A Chi Square test will be performed to replicate Grissom's approach. This research will answer the following basic

question: Which test is most applicable and provides the most powerful test statistic for each situation?

Significance of the Research

Problem-identification and problem-solving studies assess situations or help one make decisions to solve any number of problems or to create opportunities affecting the dynamics of the discipline in a given situation. Real-world data sets are used to give researchers in the many fields of behavioral research, including education and psychology, something concrete to which they can relate when making decisions regarding appropriateness of statistical tests. In contrast to earlier times, when hand-calculation was necessary and burdensome, this research will enable investigators to employ accessible, sophisticated statistical computer software offering these computations, in a quick, simple and accurate manner.

Definitions

Ordinal Categorical Data

Many behavioral and social science research methodologies use groups which comprise two or more presumed ordinal levels of outcomes. For example, Grissom (1994) suggested that in using two types (categories) of marital therapies, three levels of outcomes may occur: divorced, no change, or improved. When an outcome includes categories, the data obtained is identified as categorical. When the outcome categories include inherent levels of magnitude, the data is identified as ordinal categorical.

Permutation Test

"A permutation of the integers (1,2, ... , *n*) is a rearrangement of these integers" (Hollander & Wolfe, 1973). A permutation test means "the permuting of data for the repeated-measures randomization trend test with *k* levels of a treatment and *n* subjects rearranges the order of the measurements over the *k* treatment levels for each of the *n* subjects" (Edgington & Khuller, 1992). The reference distribution, or permutation distribution, "is derived from all possible arrangements of the data" (May & Hunter, 1990).

Power

Power is defined as "1 - β, ... the probability that we will reject a false null hypothesis" (Daniel & Terrell, 1995), where β is the probability of a Type II error.

Distribution

Distribution refers to the population of scores. In the course of collecting data, the assumption of population normality in parametric statistics is often violated. This means that the normal distributional characteristics of scores does not occur.

Robustness

Robustness refers to the insensitivity of statistical tests to violations of assumptions. Bradley (1968) commented:

> Practically any violation of a parametric test's assumptions alters the distribution of the test statistic and changes the probabilities of Type I and Type II errors. The test is said to be robust against violation of a certain assumption if its probabilities of Type I or Type II errors (usually the former) are not appreciably affected by the violation" (p.25).

Robustness is a highly debated topic (Bradley, 1978; Blair, 1981; Boneau, 1960).There is no standard definition of what constitutes a robust statistic (Bradley, 1968). A range of p-values stated by the researcher for a given alpha level quantifies the level of robustness. Generally speaking, Monte Carlo studies take a liberal stand on the issue, employing alpha levels of 0.05 or 0.01, thereby permitting a broad range of p-values and plenty of opportunity for researchers to reject the null hypothesis. Other studies, such as those in the pure sciences and medicine, usually require a more conservative alpha level of, perhaps, 0.001 or less, restricting the range for rejection considerably. Thus, to define robustness adequately, a quantitative range of *p*-values must be assumed, and a complete quantitative statement about the degree of violation of the assumption with complete specifications regarding the sampling procedure and the conditions under which the test is performed must be provided (Bradley, 1968).

CHAPTER TWO

Review of Literature

The Debate

Behavioral studies begin with a research problem. For example, a null hypothesis of no difference between two independent means is stated, and an a priori level of significance for rejecting the null hypothesis is determined. The researcher performs the statistical test and makes inferences to the population based on results from the sample. This procedure of statistical analysis is based on the classical approach which assumes data are at least approximately normally distributed. It became the paradigm adopted by researchers during the late nineteenth and early twentieth centuries in disciplines of mathematics, science, and human behavior. Parametric tests developed by Fisher, Pearson, and "Student" were promoted in research manuals, although sometimes they neglected explanations of the mathematical logic underpinning this approach. Consequently, applied investigators often acquired statistical skills, but lacked understanding (Bradley, 1968).

The validity of hypothesis testing in social and behavioral sciences, such as psychological research, triggered a debate among investigators. Researchers in psychology took a more pragmatic, rather than theoretical, approach to data analysis. The simplicity of a "yes-no" interpretation of results using the scientific method was, indeed, very attractive to researchers, but the realization became apparent that decision-making is not a clear-cut phenomenon in human behavior studies. The accepted paradigm that a better understanding of psychological problems could be obtained by studying behavior through a laboratory linear process model was challenged because many aspects of psychology must be studied in context.

Petrinovich (1979) argued for the inclusion of ecological variables in the behavioral equation to maintain focus on the essence of human existence. He called for a

scientific revolution to find a better paradigm than the accepted one, permitting generalization to representative situations through a systematic framework for data collection. These sentiments were echoed by others (Bradley, 1968, 1977, 1978, 1982; Cohen, 1990; Dar, Serlin & Omer, 1994; Omer & Dar, 1992; Rossi, 1990) who claimed the adequacy of an inference is "a function of how the data were produced, not how they are analyzed" (Cohen & Cohen, 1975, p.5). These authors noted that well-formulated research questions lead to data sets which meet the general assumptions of a particular statistical test. The demands of clinical populations and individualized treatments, for example, did not coincide with the theoretical assumptions of random assignment and homogeneous treatment administration (Omer & Dar, 1992). A review of research articles in the *Journal of Counseling Psychology* over a two-year period (Thoreson, 1969) showed that 75% of analyses were descriptive/correlational and none employed controlled environments and individual behavioral changes. Seventy-five percent of the studies used Posttest-only designs; yet only twenty percent employed random selection of subjects, a clear violation of a fundamental assumption underlying the Posttest-only design.

Parametric versus Nonparametric Tests

The controversy over the correct approach to statistical analysis of applied data sets encouraged some investigators to pursue strategies nonparametric in nature to avoid making false assumptions about the data. Blair, Higgins and Smitley (1980) conducted Monte Carlo studies which endorsed Hollander and Wolfe's (1973) perspective that two sample rank tests outperform their parametric counterparts for a variety of nonnormal distributions, such as those found in clinical studies. Hollander and Wolfe referred to Kendall's (1948) authorship of the first textbook devoted solely to ranking methods. Kendall recognized that ranking methods had wide applicability to the fields of psychology, education, industrial experimentation, and economics. Sawilowsky (1990)

noted that, in terms of asymptotic theory, the Wilcoxon rank test is a very powerful test, citing Noether (1984) and Wolfowitz (1949), who published a survey of nonparametric methods and believed that the only wasted information in nonparametric techniques is that which is not available. He also cited numerous small-sample Monte Carlo studies supporting this. Moreover, Sawilowsky (1990) further stated that "many variables encountered in education and psychology that are treated as interval in scale may be better justified as ordinal in scale" (p.95).

Historically, other issues which arose with applied researchers were the time and labor required to convert data to ranks and the difficult interpretation of results employed in nonparametric approaches (Boneau, 1960, p.312). The prospect of giving up the straight-forward, "cookbook" methods of statistical analysis was more than some wished to manage. Today, this is not an issue because modern computers quickly perform simplified ranks procedures provided by statistical software packages.

Ripstra (1974) admitted that often many readers do not perceive misrepresentations of results made by researchers. He acknowledged: "Misleading or false results can be costly, especially in the fields that affect human beings" (p.48). Mehrens (1978) pleaded for more sophistication in the interpretation of results in counseling research and cited Glass (1976) who shared that perspective.

An increase in the literature devoted to the appropriateness of statistics in certain situations has appeared over the last three decades. Parametric statistics, such as the *t* and *F*, have most often been treated in these studies, exhibiting the growing concern of researchers about their validity in certain situations. When the underlying assumptions have been violated, alternative robust and powerful methods of analysis must be found.

Monte Carlo studies have shown that under certain circumstances (e.g., equal sample size, alpha set at least to 0.05, or when the researcher is doing a one-tailed versus a two-tailed test), the *t* is acceptably robust and is justified (Sawilowsky & Blair, 1992). Of course, investigations have shown that nonparametric statistics are preferred, in terms

of departures from population normality, with respect to Type I error (Blair, 1981). The issue, then, becomes one of comparative power.

To restate, when all parametric assumptions are met, including normality, the appropriateness of the *t* test is never a question because it is, by definition, uniformly the most powerful, unbiased test. Applied researchers, who are concerned with nonnormal data sets, would also accept that the *t* is sufficiently robust with respect to Type I and Type II error. The question remains, nevertheless, whether or not there is another statistic, a non-parametric one, for example, that is more powerful for these situations.

The Data

A review of literature found that researchers are aware of the need to employ nonparametric approaches to data analysis in education and psychology. Jenkins, Fuqua and Froehle (1984) conducted an analysis of 1,698 empirical articles published in the *Journal of Counseling Psychology* and found that the trend continued to focus on parametric procedures, which assume normally distributed data, in psychological research. The distributional properties of the data, however, were rarely reported, rendering valid judgments about the appropriateness of the statistic used nearly impossible. Efficiency of the statistical test is of optimal concern in any research, but it is hindered by the nature of the data generated within the behavioral setting.

Micceri (1989) conducted an investigation of 440 large-sample distributions of psychometric and achievement measures. At the $\alpha = 0.01$ significance level, all but 3% of the data sets represented radical departures from normality. "None of the examined distributions passed all tests of normality, and very few seemed to be an even reasonably close approximation to the Gaussian" (p.161). Micceri's findings raised doubts about robustness studies, such as Boneau's (1960), which proclaimed the robustness of the *t* and *F* statistics under convenient theoretical distributions, but did not represent real life situations. Rather, Micceri's investigation identified population characteristics which

produced several nonnormal distributions relevant to behavioral research, theory development and decision-making.

Monte Carlo studies compared nonparametric alternatives to investigate the properties of these statistics under certain nonnormal distributions. However, data acquired by pseudo-random number generators do not accurately model distributions encountered in real world situations. Micceri (1989) and Sawilowsky and Hillman (1992) attested to the need for employing real data sets to determine the most appropriate, powerful test in nonnormal situations. Consequently, the researcher is not answering a question about theoretical data, but answering a theoretical question about real data.

The Current Study. The current study is concerned with real data which is ordinal and categorical, under a variety of sample sizes, forming a 2 x c design. The data are categorized into two treatment groups, for example, Curriculum I and Curriculum II, or Therapy I and Therapy II. Then, the groups are categorized again into x number of ordered levels, such as low, medium, and high; or not improved, somewhat improved, much improved, and cured. Categorizing outcome data into three or more levels provides a more informative result for educators and psychologists using real world data than "yes-no" scales.

P-values

Gajjar, Mehta, Patel, and Senchaudhuri (1992) noted the importance of p-values and the accuracy with which they are calculated. Statistical inference relies on probability; and, subsequently, the reliability of the interpretation of data and correct decision-making are dependent on this information. When researchers in education and psychology employ real data sets in their analyses, they frequently find that the distribution of their samples is more extreme than the null distribution, and the p-value obtained produces evidence against the null hypothesis.

In the past, mathematical statistics have shown, via asymptotic theory (i.e., ARE, BRE), the promise of more power for nonparametric statistics over their parametric counterparts under a variety of nonnormal distributions. This has led to small samples research, and ultimately to the adoption of nonparametric statistics in many research situations. The *p*-values often obtained represented asymptotic values, as opposed to exact *p*-values, due to the lack of computer ability to derive them.

Gajjar, et al. (1992) explained the importance of using exact *p*-values versus the asymptotic *p*-value with the following example: a Chi Square (X^2) test for row and column interaction in a 3 x 9 sparse contingency table, for example, produced an observed test statistic of $X^2 = 22.29$. The asymptotic *p*-value is the tail area to the right of 22.29 for a chi-squared distribution with 16 degrees of freedom. The observed *p*-value of 0.1342 implies there is no row and column interaction. However, by computing the exact *p*-value of the Pearson chi-square statistic for the tail area to the right of 22.29, an exact *p*-value of 0.0013 was obtained, implying that there was, in fact, a significant row and column interaction.

StatXact-Turbo, a statistical software package for exact nonparametrical inference, computes exact permutational *p*-values. Gajjar, et al. (1992) explain that the process is straight-forward:

> One must construct a reference set of all possible outcomes in which the exact null probability of each outcome is known. The exact *p*-value is then the sum of exact probabilities of those outcomes in the reference set that are at least as extreme as the one observed (pp.1-4).

Example from the Literature

Grissom (1994) examined a study of patients in marital therapy (Snyder, Wills, & Grady-Fletcher, 1991) and another study of headache therapy (Holroyd, Nash, Pingel, Cordingley, & Jerome, 1991). A 2 x 3 table produced in the former study revealed three

ordered levels of outcomes for two types of therapies (ordinal categorical data), where each sample was larger than ten. Grissom questioned the use of Chi Square, a technique introduced in the nineteenth century, to compare expected and actual occurrences, because that statistic does not provide information about superiority of the outcomes; it only suggests there is a difference. Information about superiority is important for decision-making in both education and psychology. Therefore, Grissom suggested the use of the Wilcoxon Mann-Whitney *U* test, which does provide information about superiority of the outcome categories.

The question remains, however, why the analysis of the data employed in this study was not applied to the original scores, producing a more accurate analysis to answer the hypothesis of which treatment is superior. Rather, reference was made to clinical outcome scales to which arbitrary cut-off points produced ordinal outcome levels. The appearance of gaining information here may be deceiving when the assumptions and decision rules for defining categories are examined. The reliability of instruments providing measurements like "headache intensity" described by Barlow, Hayes, and Nelson (1984) and cited in Grissom (1994), for example, are in question. How does one determine intensity of a headache? Labels of identification may be very subjective in nature, producing excessive inaccuracy of decision rules. The ability of the researcher to show how clients or subjects move from one category to another, however, has always been important in the fields of education and psychology. The above questions regarding instrument reliability may be topics for further research; this research examines other theoretical issues.

The contingency table in Grissom's study provided a count of clients which produced a particular outcome level for the treatment given. Grissom's suggestion to use the Wilcoxon-Mann-Whitney *U* test will indeed answer the hypothesis about movement from one category to another and will simultaneously respond to the hypothesis of which treatment is more effective. However, that author has only progressed the approach

toward clinical research to 1947 when the Mann-Whitney *U* test was introduced (or 1945 in terms of the WRS test). The advent of computers has presented new opportunities for researchers to select from a variety of alternative procedures for certain situations.

Grissom referred to Emerson and Moses (1985) in suggesting that it is not necessary to compute exact *p*-values for data sets which contain sample sizes over ten where many ties are present, for the approximate *p*-value obtained usually comes within fifty percent of the exact *p*-value. The current study will address this issue. Gajjar, et al. (1992) showed that one can compute, just as easily, an exact *p*-value for samples large or small, whether using ranked or original data, providing the researcher with more accurate information than approximations. The approach by Gajjar, et al. brought education and psychology research into the 1990s, incorporated the employment of user-friendly computer software, and eliminated the guesswork of determining significance.

The Concept of Error

Researchers in education and psychology often compare statistics sampled from two or more groups. The *p*-value and confidence intervals are generally accepted as "the two most useful quantitative measures for determining whether, and by how much, the populations differ" (Gajjar, et al., 1992). *P*-values and confidence intervals can be interpreted precisely. "The *p*-value ...represents the probability of a Type I error (i.e., α)" (Runyon & Haber, 1991). In other words, the *p*-value represents the probability of rejecting a null hypothesis which is, in fact, true. A region for rejection of the null hypothesis (*Ho*) is determined at the onset of the experiment. This region, located at the extreme end(s) of the distribution, is described by α (also referred to as the alpha level, or level of significance). For example, if alpha (α) is set at 0.05, the probability of the outcome occurring due to chance alone is twenty to one. That is, five percent of the time the researcher will mistakenly reject the null hypothesis (making a Type I error). The 0.05 level of significance is most often used in social and behavioral science research unless

there is a specific reason for being more conservative about making a Type I error (Runyon & Haber, 1991).

A Type II error (Type β error) is defined by Runyon and Haber (1991) as "the probability of failing to reject (or accepting) *Ho* when it is actually false." thus, β is the probability of making a Type II error. Type II errors are common in behavioral research for two reasons: 1) the sample sizes are so small, the statistics lack power; and 2) sometimes the researcher sets the alpha level too conservatively for the hypothesis being tested. For example, if $\alpha = 0.01$ and the results observed in the experiment occur by chance just two percent of the time, the researcher fails to reject a null hypothesis (the *p*-value does not fall within the rejection region). However, the experimental effect may, in fact, be significant. If α was set at 0.05, the researcher would have found significance, indicating the treatment had an effect. The lower the researcher sets the level of rejection, the less likely, then, he is of making a Type I error, and the more likely he is of making a Type II error, and visa versa. The key is to find a balance between the probabilities of a Type I and a Type II error (Cohen, 1983).

Power

Effecting the balance between Type I and Type II error underlies the concept of power. "The power of a test is defined simply as the probability of rejecting the null hypothesis when it is in fact false" (Runyon & Haber, 1991). Symbolically, power is defined as $1 - \beta$, where β = the probability of a Type II error. Note that power can only be estimated when the *Ho* is false. The goal of the researcher is to use a statistical test which yields high power, that is, a test that will reject a null hypothesis which is not true.

Cohen (1990) noted that researchers in behavioral studies, particularly in psychological research, devote little attention to power. He attributed this neglect to a lack of understanding of the term and an inaccessibility of a reference handbook, even though he wrote such a reference in 1969 and revised in 1977 and 1988. Still, in 1990,

Cohen found that little research reflected discussions on power analysis; and, when they did, 11% of the studies incorrectly interpreted the analysis.

"Power-efficiency refers to the power of a test relative to the sample, and permits one to compare the power of two different statistical tests" (McCall, 1980). If the difference in central tendency is being considered, for example, traditionally, researchers used the *t* test (a parametric test to compare means), rather than nonparametric tests. In real educational and psychological data, however, investigators have learned that distributions are often skewed, sparse, or sample sizes are small; and, subsequently, for these situations, nonparametric methods have led the way in efficiency of statistical power (Blair & Higgins, 1980; Blair, Higgins & Smitley, 1980; Sawilowsky, 1990). Power will be studied here in the context of four nonparametric approaches for real data collected in the educational and psychological arenas.

The Distributions

Educators and psychometricians have failed to realize how nonnormal their data really are, although their data are characterized by discrete data points hardly describing a normal distribution (Micceri, 1989). Distributions, in Micceri's review of psychometric and ability measures, were described in terms of symmetrical properties, tail weight, multimodality and digit preferences for varied sample sizes.

Symmetry. Distributions can take on any number of shapes in behavioral research; but there are a few standard types which are based on the properties of symmetry, namely skewness and kurtosis. These properties are noteworthy with regard to this study because similar distributional characteristics will likely be present in the data sets found in educational and psychological journals.

The foremost distribution from a theoretical perspective is the symmetrical bell-shaped distribution shown in Figure 2.1. When the distribution is symmetrical, the mean and median coincide. Other distributions may be described as bell-shaped, but are

asymmetrical or skewed. In these situations, the values are concentrated at one end of the distribution, and extreme values form a "tail." If the tail is on the left, the skew is negative (Figure 2.2) and some of the low values are not offset by corresponding high values. A tail on the right describes a positive skew (Figure 2.3) and some of the high values are not offset by low values.

Kurtosis refers to the degree to which a frequency distribution is peaked, and it is caused by the "relative concentration of scores in the center of the distribution rather than the tails" (Levine, Ramsey, & Barenson, 1995). Unlike symmetry, which compares the positions of the mean and median, the kurtosis of a distribution has no means of assessment. For example, three distributions of equal sample sizes may have the same median, but the concentration of scores may differ, resulting in different displays of the data. Figure 2.4 shows how three distributions with the same median appear with different kurtoses.

Tails. The extreme ends of the distribution are called the tails. Researchers often have an idea of whether the distribution has outliers (extreme scores) or not, and would know whether the sample tails of the distribution *f* are long/heavy or short/light. Long tails occur with very extreme scores; heavy tails occur with many extreme scores (see Figure 2.5). By standardizing the distribution parameters into percentiles, for example, one can compare densities and tails of samples from real data.

Bimodal and Multimodal Distributions. When distributions overlap but have different "maxima" (Hammond & Householder, 1963), they have "lumpy" appearances. These distributions are called bimodal, if there are two peaks, and multimodal, if there are more than two peaks.

Bimodality and multimodality of distributions requires some subjectivity in the interpretation. Therefore, Micceri (1989) applied two methods to his data:

1. Histograms identified distribution with more than a single mode.

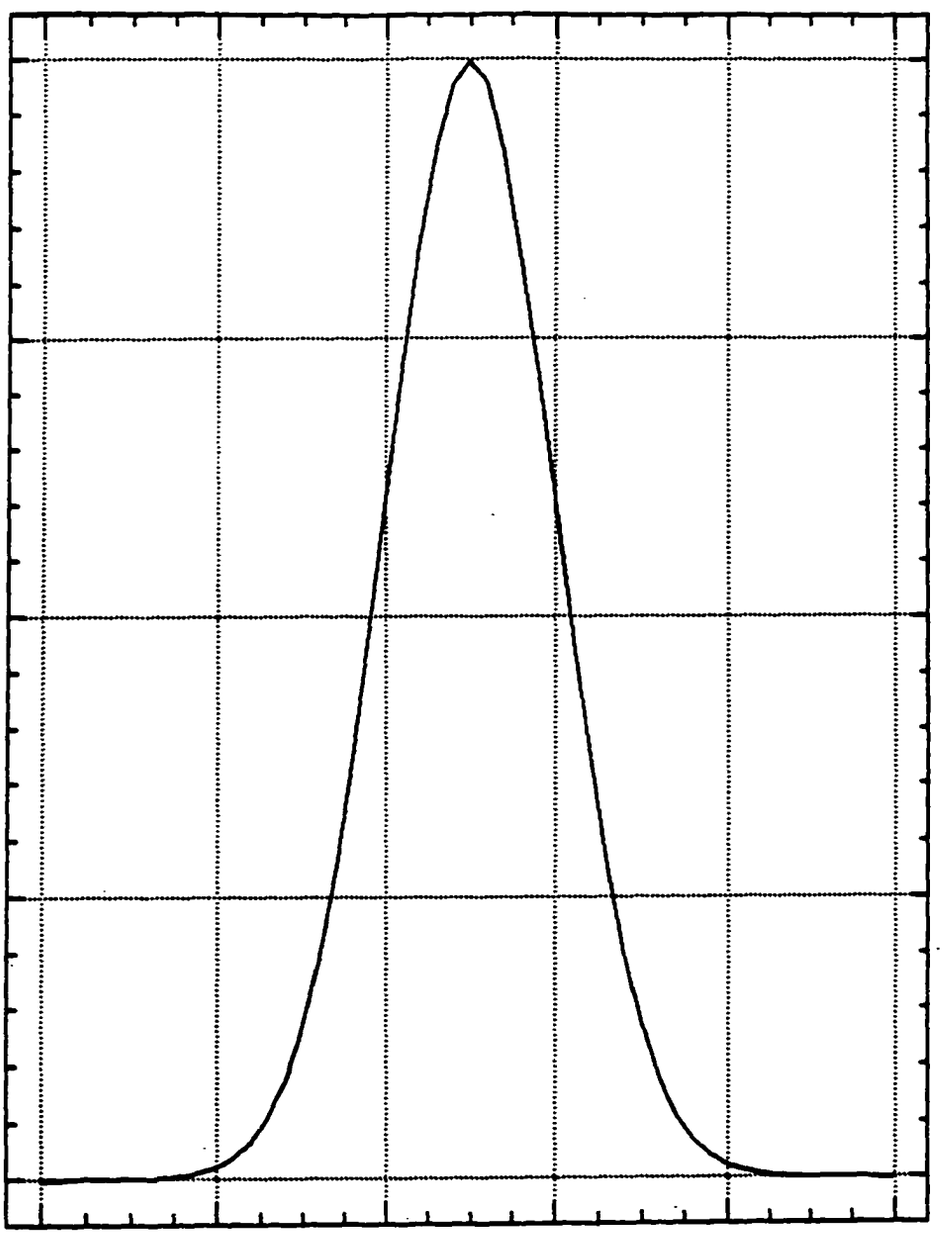

Figure 2.1. Symmetrical Bell-shaped Distribution.

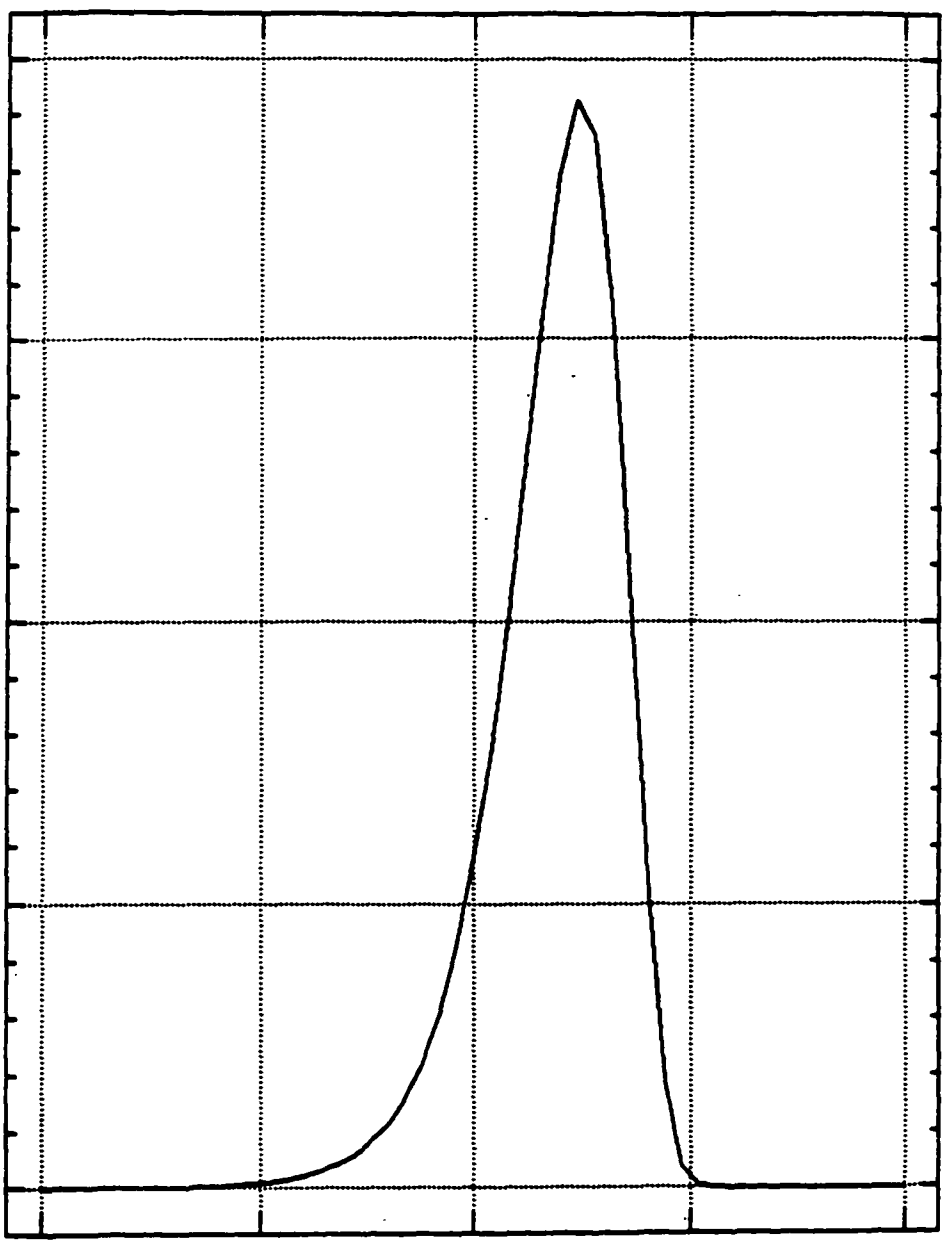

Figure 2.2. Asymmetrical Negatively Skewed Distribution.

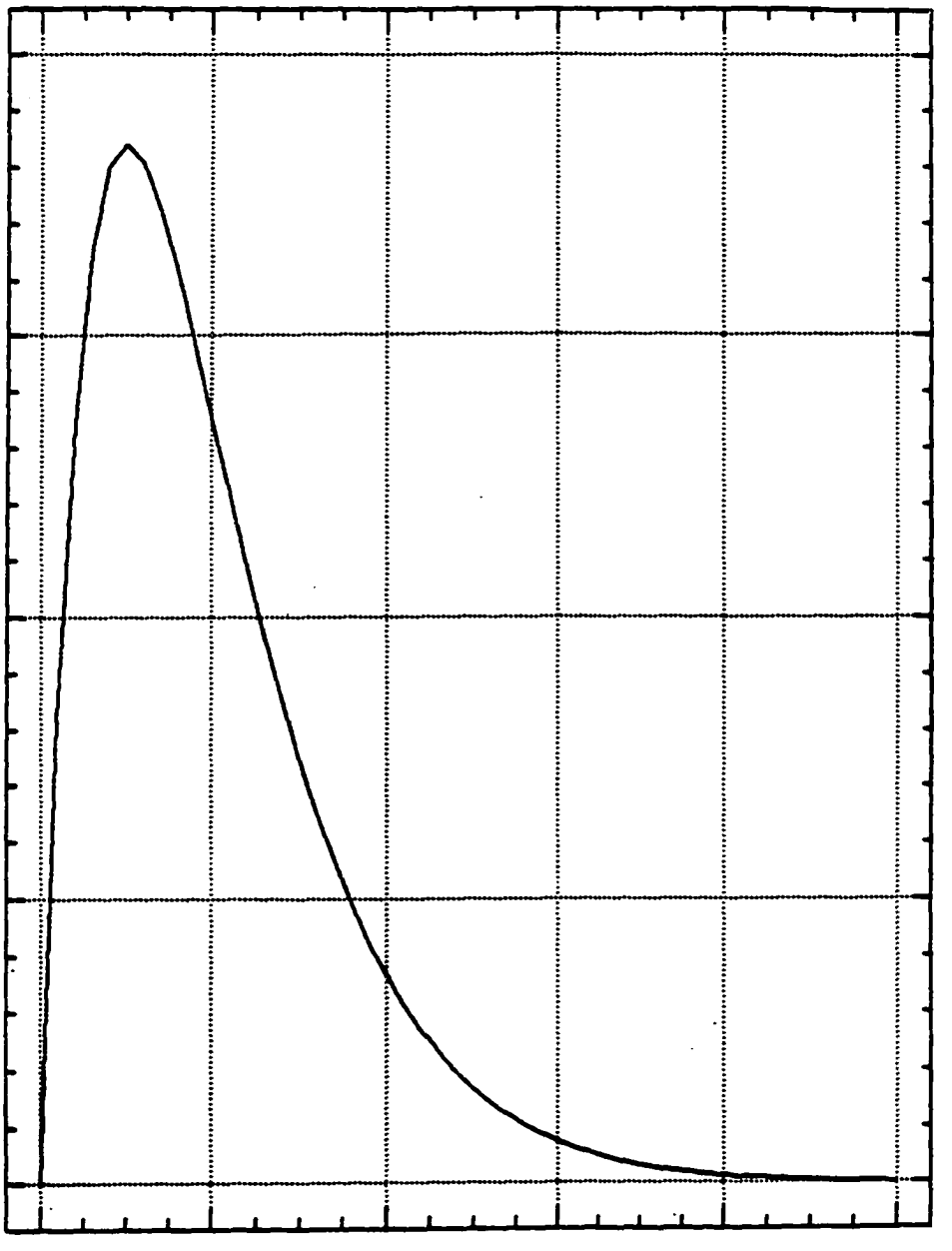

Figure 2.3. Asymmetrical Positively Skewed Distribution.

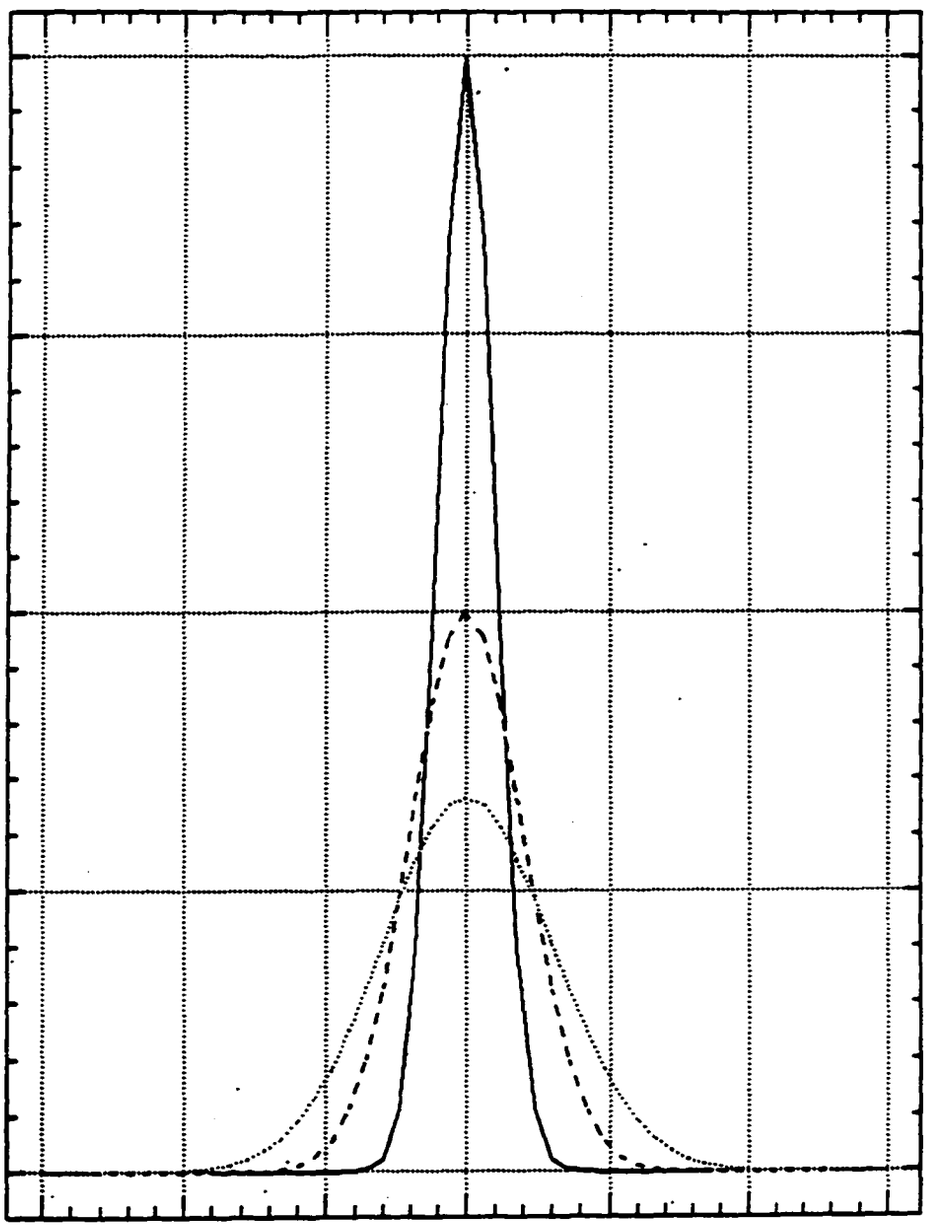

Figure 2.4. Three Distributions with Different Kurtoses.

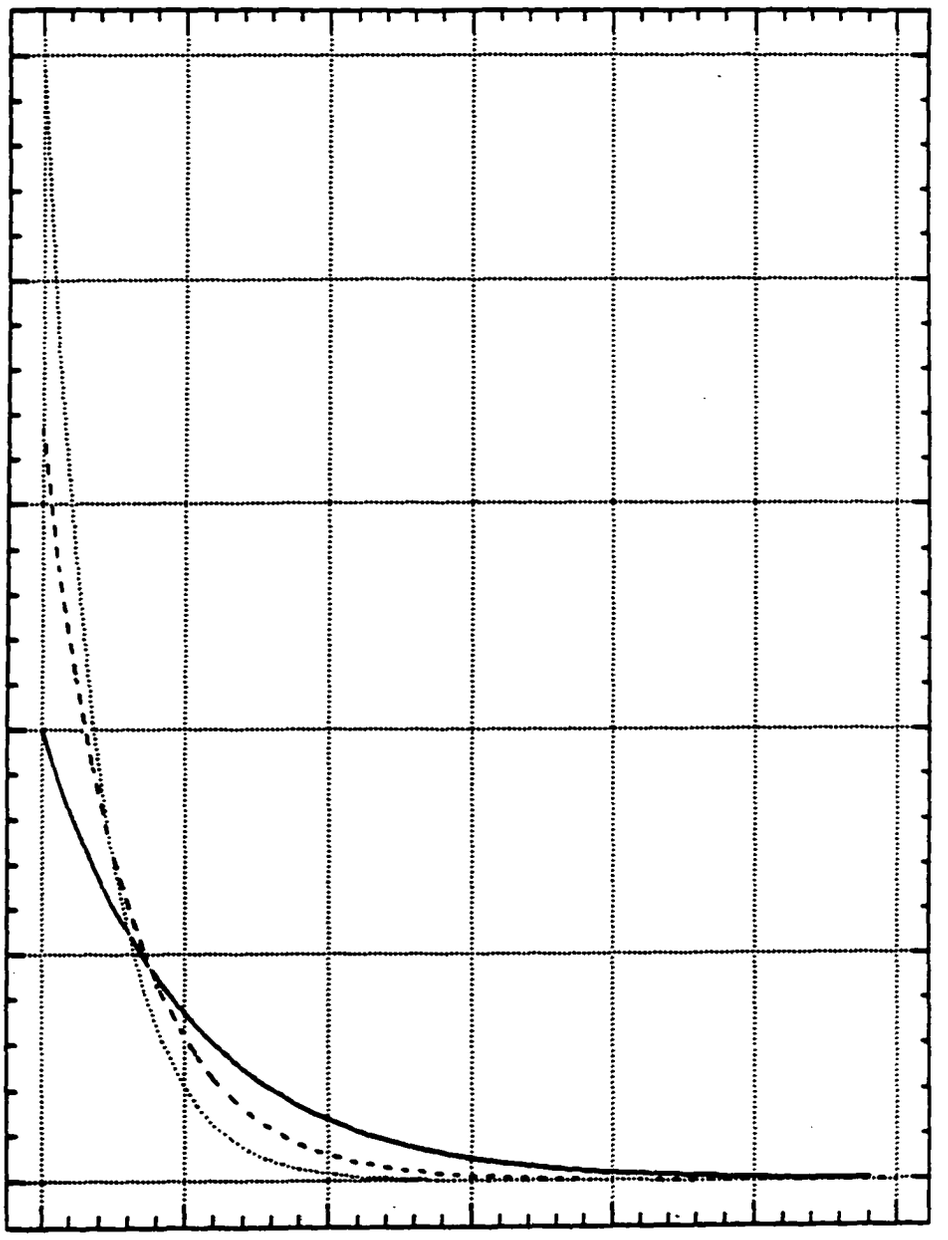

Figure 2.5. Display of Long and Heavy Tails.

2. Five modal categories were defined in the computer analysis, based on occurrences of at least 80 percent frequency of that of the true mode; and intervals between modes were standardized, creating up to four standardized distance values among frequently occurring sample points.

If one standardized distance between modes was greater than two-thirds of a standard deviation, the distribution was categorized bimodal. If more than one standardized distance between modes was greater than two-thirds of a standard deviation, the distribution was categorized multimodal.

A bimodal distribution would likely be found in test scores of a class, for example, where some students studied and some did not study. Thus, the examination would be difficult for some and easy for others. The result is many low grades and many high grades, and very few grades near the mean.

Other Shapes. Rarely does one know the exact density of a distribution. Traditionally, however, there are several types of distributions encountered in theory and in applied behavioral research. The theoretical distributions are uniform, normal, logistic, exponential, double exponential (Laplace), and the cauchy (Hajek, 1969). In practice, Macceri (1989) and Sawilowsky and Hillman (1992) encountered the "discrete mass at zero with gap," which represented sample data from psychometric and ability measures and first-use studies, as well as at least seven other prevalent shapes (e.g., digit preference, extreme asymmetry, multimodal, lumpy).

The Tests

This study compares the Chi Square and a variety of nonparametric statistical tests in the quest for the most powerful test for ordinal categorical data in a 2 x c design for certain situations. The nonparametric tests are the Wilcoxon Rank Sum test (also known as the Mann Whitney *U* test), the Random Normal Scores test, the Savage Scores

test, and the Permutation test. Gajjar, et al. (1992) explains the relevance of these tests to ordinal categorical data:

> The two rows of stratum k represent two independent multinomial populations. Each observation falls into one of c ordinal response categories. Thus x_{jk} is the number of observations, out of a total of m_k , falling into ordered category j for population 1, and x'_{jk} is the number of observations out of a total of m'_k falling into ordered category j for population 2.
>
> The Wilcoxon rank sum test, the Normal scores test, and the Permutation test with arbitrary scores are applicable to such data, and test whether the two populations have the same underlying multinomial distribution within each stratum. The scores, w_1, w_2, ...w_c, are numerical values assigned to the c ordered multinomial response categories (p.4-7).

The Wilcoxon Rank Sum (WRS) or Mann-Whitney *U* Test. The Wilcoxon Rank Sum test was introduced in 1945 (Wilcoxon, 1945; Mann & Whitney, 1947) to test whether two independent groups have been drawn from the same population (*Ho: f(x) = g(x)*). This procedure is a powerful nonparametric alternative to the parametric *t* test and it is one of the most prolific and well-known nonparametric statistical tests used in educational and psychological research today. The Wilcoxon test detects the difference between respective location parameters of two groups, and also conveniently accommodates ordinal categorical data. This test is most powerful with nonnormal distributions; and it is only slightly less powerful than the parametric *t* test under normally distributed populations (Blair, 1980; Sawilowsky, 1990).

To use the Wilcoxon test, N observations are ordered from least to greatest; and, if R denotes the rank of Y in the ordering, the formula for the Wilcoxon test is:

$$W = \sum_{j=1}^{n} R_j$$

In other words, W is the sum of the Y ranks (Hollander & Wolfe, 1973) which tests the hypothesis H_0: $f(x)=g(x)$.

Once the size of the samples is determined, one of three methods to compute the Mann-Whitney U test may be used:

1. When $n_2 < 8$, an exact p-value can be obtained from the U distribution. To apply the U test in small samples, the scores from n_1, the number of cases in the smaller of the two independent groups, and n_2, the number of cases in the larger of the two groups are combined and ranked in order of increasing size. For example, let us suppose that scores were obtained for Curriculum 1 (C_1) and Curriculum 2 (C_2). The scores are combined in a list of increasing order. If there are negative scores, the lowest ranking number is assigned to the largest negative score, retaining each score's identity as either a C_1 or a C_2 score. Now, focus on the C_1 group and count the number of C_2 scores that precede each score in the C_1 group. This number is U. The sampling distribution of U under the H_0 is known and will allow one to determine the exact probability associated with the occurrence under H_0 of a score as extreme as an observed value of U. The researcher only needs to know n_1, n_2 and U to determine the probability under the H_0 associated with the data. Note that the correct statistic is derived from n_2.

2. For samples of $n_2 = 9$ to 20, a rank of 1 should be given the lowest score in the combined group of scores, a rank of 2 is assigned to the next lowest score, and so on. Then, the following formula may be applied (Siegel, 1956):

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

or, equivalently,

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_2$$

where R_1 equals the sum of the ranks assigned to the group whose sample size is n_1 and R_2 equals the sum of the ranks assigned to the group whose sample size is n_2. Each of these formulas yields a different U value. The desired statistic here is the smaller of the two U values. If the observed value of U is larger than $n_1 n_2/2$, it is U'. The simple transformation

$$U = n_1 n_2 - U'$$

where U' represents the larger of the two values, can be employed after computing just one of the above formulas. This procedure will allow the researcher to obtain the proper statistic without computing both formulas above to find the smaller statistic.

3. For samples where $n_2 > 20$, as n_1 and n_2 increase in size, the sampling distribution of U approximates a normal (z) distribution (Mann & Whitney, 1947). This is a relevant point to the discussion about ties for this N below.

The occurrence of ties in behavioral research is likely. When tied scores occur, the Mann-Whitney U test uses the average of the ranks the tied scores would have had ties not occurred. Interestingly, the value of U is only affected if the ties occur between two or more observations involving both groups (Siegel, 1954). Ties affect the variability of the set of ranks, so the correction for ties applies to the standard deviation of the U distribution.

For large samples ($n_2 > 20$), a correction for ties is available for use with the normal curve approximation. The formula for correction for ties also applies to the z expression.

$$\sigma_U = \sqrt{\frac{n_1 n_2}{N(N-1)} \cdot \frac{N_3 - N}{12} - \Sigma T}$$

where $N = n_1 + n_2$, and

$$T = \frac{t_3 - t}{12}$$

where t is the number of observations tied for a given and Ts are summed over all groups of tied observations.

The formula also becomes the new expression for the denominator of z, where

$$z = \frac{U - \mu_U}{\sigma_U} = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{(n_1)(n_2)(n_1+1)}{12}}} = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2}{N(N-1)} \cdot \frac{N_3 - N}{12} - \Sigma T}}$$

One of the reasons the Wilcoxon works well, is that it has a built in correction for ties. It is important to correct for ties in the Wilcoxon Mann Whitney U test because ties affect the power of the statistic and the Type I error rate (Sawilowsky, 1990). The effect of ties diminishes as the sample size increases. As the number of ties increases, however, the effect substantially increases, regardless of sample size.

Normal Scores Test. The Normal Scores test, or Van der Waerden test, is an alternative to the Wilcoxon Rank Sum test to compare two group. It combines both

Normal Scores Test. The Normal Scores test, or Van der Waerden test, is an alternative to the Wilcoxon Rank Sum test to compare two group. It combines both samples (Z_1, ..., Z_N) and operates on the assumption that the samples are independently normally distributed, with 0 mean and unit variance. If only the ranked Zs are preserved, and the original scores are lost, one might wish to reconstruct the original scores. If one considers the random variable, $Z_{(s)}$, which has rank s and is considered to be the sth smallest of a sample of size N from a standard normal distribution, the researcher may expect this random variable to take on an estimated value. The natural estimate is

$$a_N(s) = E\Phi(Z_{(s)}),$$

the expectation of $Z_{(s)}$ when the Zs are a sample from the standard normal distribution Φ (Lehmann, 1975). The expectations are known as Normal Scores, first proposed by Fisher and Yates (1938) to replace the original observations in the standard normal theory tests.

When the reconstructed observations are substituted into the t statistic to obtain a test equivalent to the t statistic based only on the combined ranks of the observations, a Normal Scores test is obtained and the statistic, T, is observed. The T distribution is approximately normal for a large N. The Normal Scores test is a strong competitor to the t test asymptotically and with comparison restricted to shift models. This test compares in power to the Wilcoxon test when F is close to normal. The Normal Scores test is preferable when the distribution ends abruptly, as in the exponential distribution at zero

and the uniform distribution at its endpoints. The Wilcoxon is more powerful for distribution with heavy tails (Lehmann, 1975).

Savage Scores Test. The Savage Scores test, or Exponential Scores test, was introduced by Savage (1956) and compares two sets of measurements for a possible difference in dispersion when the consideration is "length of life or time to failure of some material or complex piece of equipment" (Lehmann, 1975). The exponential distribution is used in this situation and is characterized by the formula

$$\frac{1}{a}e^{-x/a} \quad \text{for } x > 0$$

where a = the X distribution about the endpoint of the exponential distribution, in the case where there are two independent samples, m and n, representing the X and Y scores respectively. The researcher is cautioned not to assume the distribution is exponential, however, and is advised to use a rank test using the following formula to determine the sth smallest (the shortest time of life) of N observations:

$$a_N(s) = E_\varepsilon [Z_{(s)}]$$

This formula represents the expected score under the exponential distribution (note the similarity to the Expected Normal Score above). "Without loss of generality," Lehmann (1973) explains, "a can be taken to equal 1." Subsequently, the remaining scores can be calculated:

$$a_N(s) = \frac{1}{N} + \frac{1}{N-1} + \ldots + \frac{1}{N-s+1}$$

Permutation Test. In 1935, Fisher initiated the permutation model for statistical analysis which was developed by Pitman (1937, 1938). The permutation model is nonparametric because it makes no assumptions about the underlying distribution of the population. There are two kinds of Permutation tests: those based on ranks and those based on original scores (May & Hunter, 1990). Many rank tests are named after their developers, for example, Mann-Whitney, Wilcoxon, Friedman or Kruskal-Wallis. The original scores permutation tests are generally referred to as simply permutation tests, although the Fisher-Pitman test and the Pitman-Welch test have been referred to by name.

The null hypotheses of Permutation tests are generally of the form of no difference between groups, for example,

$$H_o: \bar{x}_1 - \bar{x}_2 = 0.$$

"When there is no known or justifiable relationship between the data at hand and a parent population of data values, the null hypothesis may be operationalized for testing without reference to a hypothetical parent population of cases" (May & Hunter, 1992). The above null hypothesis requires no assumptions about population parameters when sample statistics are used.

A permutation distribution is derived from all possible rearrangements of the scores. A test statistic is computed for each permutation and all possible outcomes form a distribution. Unlike a theoretical sampling distribution which includes outcomes that do not relate to a particular data set, the permutation distribution is derived specifically for the data under consideration. That is why the permutation model is sometimes referred to as the "data-at-hand" model (Haber, 1990).

The proportion of permutations which obtain values of the test statistic equal to or greater than the difference observed for a particular arrangement of scores is the *p*-value of that outcome (May & Hunter, 1990). This value is a ratio of frequencies of

values equal to or more extreme than the one observed to the total number of differences possible for that data set.

Tables of critical values for original scores Permutation tests are obviously not available, as the values depend on a variety of unknown factors (e.g., number of scores, groups, what is to be permuted, etc.). Therefore, the rank score Permutation tests have been utilized to a far greater degree than the original scores tests. Rank Permutation tests (i.e., Wilcoxon, Mann-Whitney) are practical and are used often in psychological and educational settings. Their tables are computable and are broadly applicable because the rank scores (i.e., ranks 1 to *N*) are known before the onset of the study and rank permutation distributions can be computed for widespread use. However, modern computer software packages, such as StatXact-Turbo (Mehta, 1992), allows the researcher to compute permutations for any data set using rank or original scores.

Mehta (1992) provided Permutation tests in the StatXact-Turbo computer software program, and explained:

> Through this powerful feature, StatXact opens up the possibility of computing infinitely many linear rank tests and provides you with the ability to customize your hypothesis tests to particular characteristics of your data (p.4-4).

CHAPTER THREE

Methodology

Investigators have often compared power and robustness properties of statistical tests for normal and nonnormal distributions. However, pseudo-random number generators have not accurately represented distributions encountered in real world situations known to applied researchers. Micceri (1989) addressed this issue in his inquiry of 440 real world distributions characterizing ability and psychometric measures. He explained there is a "need for careful data scrutiny prior to analysis, for purposes of both selecting statistics and interpreting results" (p.161). In a similar approach to Micceri's, data will be collected from the literature for this study. However, this review will canvass studies which compare real treatments with ordinal outcomes.

Method

Sampling Plan

Samples will be collected on an availability basis. Twenty-nine Education and Psychology journals will be canvassed over a four-year period, from 1992 through 1995. Selected journals have been editorially reviewed according to American Psychological Association (1994) writing standards and contain articles dealing with treatment effects, for example, Curriculum I vs. Curriculum II or Therapy I vs. Therapy II. An inclusive list of journals appears in Table 3.1.

EDUCATION	*PSYCHOLOGY*
American Educational Research Journal	American Journal of Community Psychology
Education	American Journal of Family Therapy
Educational Psychologist	American Journal of Psychology
Educational Researcher	Basic and Applied Social Psychology
Harvard Education Review	Developmental Psychology
Journal for the Education of the Gifted	Journal of Applied Psychology
Journal for Research in Mathematics Education	Journal of Clinical Child Psychology
Journal of Negro Education	Journal of Consulting and Clinical Psychology
Journal of Special Education	Journal of Psychology: Interdisciplinary and Applied
Journal of Teacher Education	Psychological Bulletin
Perceptual and Motor Skills	Psychological Reports
Teacher Education and Special Education	Psychological Science
	Psychology and Aging
	Psychology in the Schools
	Professional Psychology: Research and Practice
	Reading Psychology
	School Psychology International

Table 3.1. Education and Psychology Journals Canvassed in the Current Study.

Articles selected for research will employ ordinal categorical data in a 2 x *c* design. For example, consider the article by Speer (1994) who studied the effect of treatment outcome research on public policy and decision-making processes in mental health. The author studied the change rates of psychotherapy clients in two types of treatments or treatment versus no treatment. A typical data set is displayed in Table 3.2.

	Deteriorated	Unchanged	Improved
Treatment A			
Treatment B			

Table 3.2. A 2 x c Table Displaying Ordinal Categorical Data.

When relevant data sets are not displayed within the journal article contents, a request will be made of each author for permission to include his data in this study. Sample sizes will be determined by each contributing author to this research. A variety of sample sizes will be collected.

Procedures

1. Data will be entered into a Gateway 2000 IBM compatible computer via the Table editor in StatXact-Turbo and will be saved in a file format included in the statistical software package..

2. Data sets will be categorized according to the size of *c* (e.g., 2 - 7) for each 2 x *c* design, and a table will display the number of data sets for each *c* category.

3. A Chi Square test, an asymptotic Wilcoxon and each nonparametric statistical test under consideration will be computed on each data set.

4. A table will be generated to display *p*-values of each test on each data set. Significant findings will be identified with an asterisk.

5. A contingency table will be created comprised of total numbers of significant findings for each test as they relate to the null hypothesis.

6. A Chi Square will be conducted on the contingency table mentioned in procedure number five.

7. A table will depict the number of significant results for each test under each number of ordinal outcome categories.

8. A Permutation test will be performed on the table mentioned in Procedure 7.

9. A histogram of significant findings of each test for each ordinal category will be produced .

10. Interpretations will be developed from the results.

Analysis

Categorizing the data sets according to the number of levels in c will provide a visual aid of the sample collected for this study. It is expected that the number of levels will range from three to seven. If any other ordinal levels occur, they will be categorized as "other."

Next, a Chi Square test (A) will be performed on each data set to replicate Grissom's study. Then, these four nonparametric tests will be run on each data set using the StatXact Turbo statistical software package: the Wilcoxon Rank Sum test (B), the Normal Scores test (C), the Savage Scores test (D), and the Permutation test (E). The *p*-values for each of the five tests computed on each data set will be recorded in a table for reference to facilitate identification of significant outcomes, which will be identified with an asterisk.

Another table depicting the number of results for each test which rejected the null hypothesis and which failed to reject the null will be established. A Chi Square analysis will determine if there is an overall significant difference in results when computers allow analysis of ordinal categorical data through permutations (Table 3.3). If significance is found, individual cell investigations will reveal which test contributed most to the significance.

	Reject	Fail to Reject
Chi Square (A)		
Wilcoxon (B)		
Normal Scores (C)		
Savage Scores (D)		
Permutation (E)		

Table 3.3. Test Results Relating to the Null Hypothesis.

In addition, a contingency table displaying the number of significant findings for each test under each size of *c* for each given alpha level will be formed (Table 3.4), and a Permutation test will be conducted to determine whether the results of this research are significant.

	Level of c					
Test	3	4	5	6	7	Other
A						
B						
C						
D						
E						

Table 3.4. Contingency Table of Significant Results for Each Test under Each Size of *c* for $\alpha = .05$.

Finally, a histogram will be created to further visualize these results. It will depict significant test results for each ordinal category of *c* at a given alpha level (Figure 3.1).

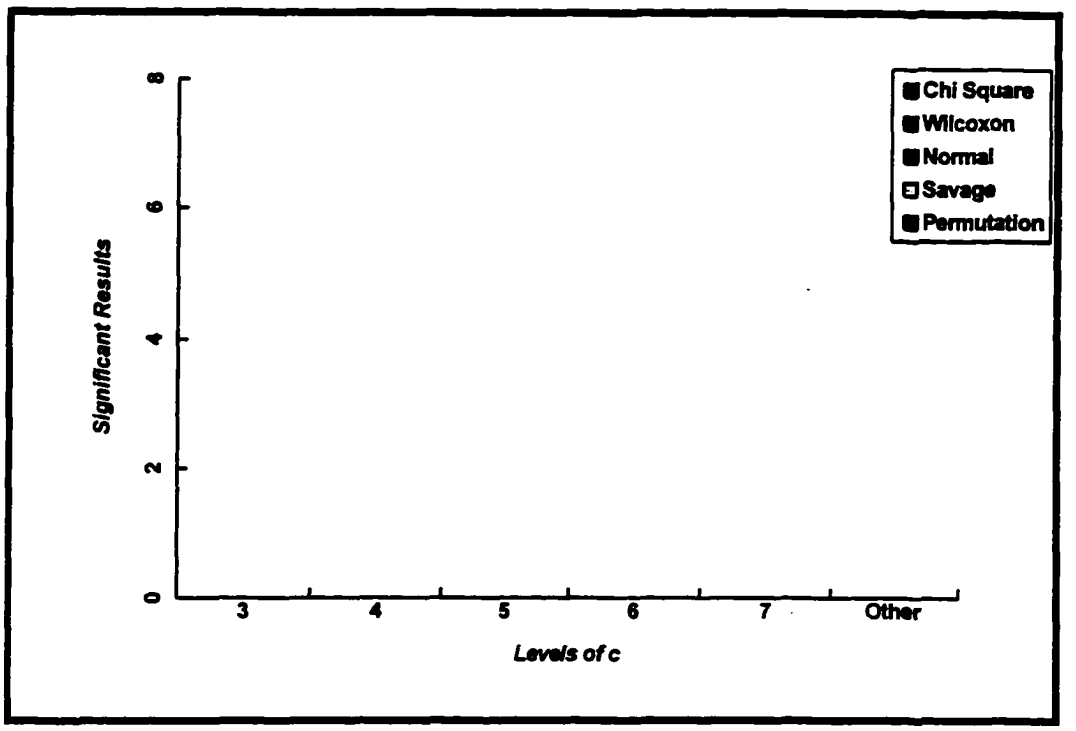

Figure 3.1. Histogram Displaying Significant Results for Each Test Under Each Level of c at $\alpha = 0.05$.

CHAPTER FOUR

Results

A survey of twenty-nine journals over a four-year period yielded 149 ordinal categorical data sets of a 2 x *c* design for this study. Two sample data sets were requested by mail, but these elicited no response. Each sample was evaluated for ordinal levels of outcome data ranging from three to eight or more.

The original table designs were varied. There were four distinct formats encountered in the data. First, Table 4.1 shows a simple two-category design with *c* ordinal levels, such as that used by Grissom (1994) to demonstrate three ordinal categorical outcomes of two marital therapies.

Therapy	Divorced	No Change	Improve
Insight	3	22	4
Behavior	12	13	1

Table 4.1. Frequencies of Three Ordinal Categorical Outcomes of Two Marital Therapies (Grissom, 1994).

Second, a 4 x c design consisting of a table with two distinct sets of two categories was divided into two data sets. For example, Morrow (1995) asked the question, "What are the most frequently observed psychological problems in your school?" Table 4.2 shows four categories of psychological problems: Emotional, Conduct, Attention Deficit Disorder (ADD) and Learning Disability (LD) according to grade levels (used ordinally), K-3, 4-6, 7-8 and 9-12. I created two tables from this

sample. One used Emotional and Conduct as the categories, and the other employed ADD and LD as the categories.

Category	K - 3	4 - 6	7 - 8	9 - 12	Total
Emotional	187	110	23	56	376
Conduct	190	116	226	55	387
ADD	190	118	23	54	385
LD	184	113	23	54	374

Table 4.2. Teacher's Perceptions of Psychological Problems in Students (Morrow, 1995).

Third, nested designs were handled as multiple data sets. For example, Cherian (1992) reported the parental education on an ordinal scale of low, middle and high for boys and girls from families where one or both parents were dead and where both parents were alive (see Table 4..3). This sample also yielded two data sets, both of which engaged the same ordinal levels of parent education, or c. The first data set considered gender (boys and girls), and the second data sets showed the two categories relating to life status of parents.

	Low	Middle	High
One/Both Parents Dead			
Boys	9	89	13
Girls	39	147	11
Both Parents Alive			
Boys	21	121	55
Girls	58	279	178

Table 4.3. Frequencies for Low-, Middle-, and High-Scoring Subjects on Parental Education from Families with One or Both Parents Dead and with Both Parents Alive (Cherian, 1992).

Fourth, other nested designs encouraged the summing of subdivisions (see Table 4.4). For example, Black (1993) compared genuine and simulated suicide notes. The table portrayed several sub-categories of men and several sub-categories of women who were placed in ordinal levels of young, middle-aged or senior age groups. The subdivisions of each gender category were summed to develop a single data set for this study. Note that ordinal levels were not manipulated in this study.

Group	Young (18-29)	Middle-aged (30-64)	Senior (65+)	Total
Men				
Blue collar	4	10	3	17
White Collar	3	18	4	25
Professional	0	7	3	10
Student	3	0	0	3
Homemaker	0	0	0	0
Women				
Blue Collar	0	0	0	0
White Collar	3	3	1	7
Professional	0	2	1	3
Student	2	0	0	2
Homemaker	1	5	4	10
Total	16	45	16	77

Table 4.4. Number of Note Pairs Matched by Age Group, Sex, and Occupational Level (Black, 1993).

In many cases, the samples were capable of many data sets. For example, the sample above could provide a data set comparing two occupations, adding the male and female counterparts to provide cell counts. Although more data sets would ensue, the manipulation of the sample tables in this fashion was not pursued.

Sample sizes for the selected studies ranged from ten to 19,256, representing three to eight ordinal levels of outcome data. A frequency count of data sets containing each number of ordinal levels (i.e., three, four, five, etc.), revealed that most researchers used

three ordinal response levels (columns, *c*), outnumbering all other levels by at least a 3:1 margin. Table 4.5 denotes the frequencies of each ordinal level encountered in this research.

3 Levels	4 Levels	5 Levels	6 Levels	7 Levels	8+ Levels
97	32	15	1	1	3

Table 4.5. Frequency of Levels of *c* for 149 Data Sets.

A Gateway 2000 Pentium P5-133 personal computer with 16 megabytes of memory was used to analyze the tests compared in this study. The software used was StatXact Turbo Statistical Software for Exact Nonparametric Inference developed by Gajjar, Mehta, Patel, and Senchaudhuri (1992). For each data set, the following four nonparametric exact tests were compared for each ordinal situation: the Wilcoxon exact test, the Logrank/Savage Scores exact test, the Normal Scores exact test and the Permutation exact test. The Logrank Test, in StatXact Turbo, specializes to the Savage Scores Test when the data are uncensored. In addition, the Wilcoxon asymptotic test and the Chi Square asymptotic test were performed to replicate the main thrust of Grissom's study (1994), from which this research was prompted. The *p*-values (Table 4.6) for each nonparametric exact test, in addition to the asymptotic p-values for the Wilcoxon and Chi Square test, include an asterisk if the value was significant at $\alpha = 0.0500$. StatXact Turbo software carries out the decimal to four places; thus, p-values of 0.05 with values higher than zero in the third or fourth decimal place were not reported significant for this research. P-value ranges for Monte Carlo tests (employing 20,000 iterations) with 99% confidence level, which were employed when an error message occurred indicating that "The problem (was) too large for the test option," are indicated in parentheses and located in cells of Table 4.6 where the dash is presented.

StatXact 2.0 reported Monte Carlo results in two ways: 1) sometimes the output indicated a p-value of 0.0. When this occurred, the computer output also showed a confidence interval (C.I.) encompassing the true p-value. For example, if the C. I. = 0.0, .0002, as reported for data set number eight, the true p-value was found somewhere within those bounds. This result would be asterisked to indicate significance; 2) sometimes the p-value was followed by a ± value. For example, the Monte Carlo p-value for data set number 41 was .0053, followed by ± .0013. This result meant that the true p-value was found somewhere within the range of .0040 and .0066. Because 0.0500 is within these boundaries, the value was marked significant. Tabled results appear in the following 13 pages without further comment.

Table 4.6. *P*-values for Tests Computed with StatXact Turbo by Gajjar, et al. (1992).

Identification Number	Wilcoxon Exact Test (MO Test)**	Logrank/Savage Scores Exact Test (MO Test)	Normal Scores Exact Test (MO Test)	Permutation Exact Test (MO Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
001	.4225	.4901	.4572	.4415	.4226	.9336
002	.3467	.2980	.3167	.3467	.3180	.8424
003	.4073	.4526	.4073	.4397	.4086	.9208
004	.0920	.0920	.0876	.0876	.0828	.1933
005	.0261*	.0613	.0260*	.0392*	.0215*	.1146
006	.2812	.3907	.3158	.2511	.2813	.0056*
007	.0122*	.0122*	.0122*	.0203*	.0115*	.0207*
008	---- (0.0)*	---- (0.0)*	----- (0.0)*	---- (0.0)*	.0000*	.0000*
009	.0042*	.0012*	.0026*	.0026*	.0035*	.0044*
010	.4658	.4768	.4844	.5071	.4656	.9747
011	.4570	.4390	.4522	.4535	.4567	.9975
012	.5831	.2441	.5831	.5831	.5831	.0123*

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992)

Identification Number	Wilcoxon Exact Test (MO Test)**	Logrank/Savage Scores Exact Test (MO Test)	Normal Scores Exact Test (MO Test)	Permutation Exact Test (MO Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
013	.0024*	.0015*	.0024*	.0024*	.0016*	.0085*
014	.0201*	.0805	.0201*	.0302*	.0231*	.0360*
015	.0001*	.0001*	.0000*	.0001*	.0001*	.0002*
016	.0000*	.0001*	.0000*	.0000*	.0000*	.0001*
017	.0016*	.0018*	.0019*	.0021*	.0016*	.0130*
018	.0160*	.0511	.0195*	.0216*	.0167*	.0953
019	.0004*	.0014*	.0004*	.0004*	.0004*	.0017*
020	.2755	.4146	.3834	.4857	.2760	.0756
021	.2968	.2184	.2662	.2902	.2938	.6945
022	.1475	.2185	.1667	.1864	.1475	.5167
023	.4488	.2234	.4909	.4916	.4469	.1238
024	.0001*	.0012*	.0002*	.0005*	.0001*	.0003*

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992)

Identification Number	Wilcoxon Exact Test (MO Test)**	Logrank/Savage Scores Exact Test (MO Test)	Normal Scores Exact Test (MO Test)	Permutation Exact Test (MO Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
025	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
026	.0135*	.3396	.0501	.0825	.0132*	.0000*
027	.0000*	.0005*	.0000*	.0000*	.0000*	.0000*
028	.0324*	.0178*	.0324*	.0330*	.0316*	.0889
029	.3467	.4884	.3467	.3738	.3494	.3865
030	.0525	.0332*	.0525	.0535	.0514	.1581
031	.4414	.4803	.4414	.4649	.4416	.9315
032	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
033	.4959	.4901	.4971	.5169	.4959	.9996
034	.5769	.3288	.4462	.5769	.5000	.4117
035	.2464	.2715	.2587	.2791	.2397	.8731
036	.3798	.3543	.3744	.4082	.3689	.9813

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992)

Identification Number	Wilcoxon Exact Test (MO Test)**	Logrank/Savage Scores Exact Test (MO Test)	Normal Scores Exact Test (MO Test)	Permutation Exact Test (MO Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
037	.4361	.4448	.4407	.4727	.4295	.9968
038	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
039	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
040	.5000	.4675	.4838	.6169	.4331	.6005
041	---- (.0053)*	---- (.0091)*	---- (.0067)*	.0056*	.0057*	.1342
042	---- (.0163)*	---- (.0231)*	---- (.0171)*	.0161*	.0180*	.2012
043	.1801	.4194	.1775	.3230	.1659	.0807
044	.0263*	.0235*	.0263*	.0330*	.0261*	.1311
045	.1903	.0985	.2053	.2570	.1911	.0941
046	.0117*	.0014*	.0328*	.0574	.0118*	.0002*
047	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
048	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992).

Identification Number	Wilcoxon Exact Test (MO/EX Test)**	Logrank/Savage Scores Exact Test (MO/EX Test)	Normal Scores Exact Test (MO/EX Test)	Permutation Exact Test (MO/EX Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
049	.0289*	.0289*	.0289*	.0294*	.0213*	.1025
050	---- (0.0)*	---- (0.0)*	---- (0.0)*	.0000*	.0000*	.0002*
051	.4254	---- (.1731)	.4773	.4683	.4235	.0161*
052	.0029*	.0290*	---- (.0041)*	.0040*	.0030*	.0238*
053	.0000*	---- (.0001)*	.0000*	.0000*	.0000*	.0000*
054	.0014*	---- (.0001)*	.0020*	.0038*	.0015*	.0023*
055	.0938	.0812	.0914	.1210	.0858	.3271
056	.0000*	.0000*	.0000*	.0000*	.0000*	.0001*
057	.5000	.4619	.5000	.5568	.4881	.9004
058	.2007	.1464	.2007	.2052	.1764	.4380
059	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
060	.1580	.2309	.1580	.1584	.1357	.1049

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992).

Identification Number	Wilcoxon Exact Test (MO Test)**	Logrank/Savage Scores Exact Test (MO Test)	Normal Scores Exact Test (MO Test)	Permutation Exact Test (MO Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
061	.3574	.4266	.3510	.3510	.3562	.2900
062	.2486	.1256	.2070	.2175	.2481	.5384
063	.0400*	.0257*	.0559	.0492*	.0389*	.1791
064	.3997	.4201	.3997	.4228	.3977	.9574
065	---- (.4912)	---- (.4985)	---- (.4918)	.5017	.4942	.9999
066	.4592	.4771	.4585	.4648	.4578	.9228
067	---- (0.0)*	---- (0.0)*	---- (0.0)*	.0000*	.0000*	.0000*
068	.0000*	.0000*	.0000*	.0000*	.0000*	.0363*
069	.3836	.1582	.4673	.4675	.3853	.0363*
070	.3836	.1582	.4673	.4675	.3853	.0363*
071	.3836	.1582	.4673	.4675	.3853	.0363*
072	.3836	.1582	.4673	.4675	.3853	.0363*

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992).

Identification Number	Wilcoxon Exact Test (MO Test)**	Logrank/Savage Scores Exact Test (MO Test)	Normal Scores Exact Test (MO Test)	Permutation Exact Test (MO Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
073	.1562	.1876	.1597	.1703	.1510	.7475
074	.1575	.1779	.1514	.1718	.1535	.7512
075	.0810	.0726	.0806	.0858	.0773	.3301
076	.0025*	.0061*	.0029*	.0045*	.0029	.0111*
077	.0269*	.0333*	.0276*	.0332*	.0266*	.1470
078	.0913	.1182	.0913	.1291	.0705	.2751
079	.3556	.3333	.3556	.3556	.2416	.5988
080	.0007*	.0043*	.0010*	.0011*	.0008*	.0070*
081	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
082	.0454*	.0265*	.0570	.0715	.0453*	.0828
083	.4433	.4234	.4569	.4904	.4398	.9181
084	.0219*	.0037*	.0265*	.0431*	.0221*	.0014*

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992).

Identification Number	Wilcoxon Exact Test (MO Test)**	Logrank/Savage Scores Exact Test (MO Test)	Normal Scores Exact Test (MO Test)	Permutation Exact Test (MO Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
085	.0475*	.0622	.0562	.0583	.0473*	.3469
086	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
087	.0748	.1062	.0748	.0777	.1059	.2591
088	.0136*	.0152*	.0136*	.0164*	.0129*	.0824
089	.2759	.1823	.2759	.2799	.2580	.4065
090	.0136*	.0152*	.0136*	.0164*	.0129*	.0824
091	.5569	.5223	.5223	.5569	.5000	1.0000
092	.5569	.5223	.5223	.5569	.5000	1.0000
093	.1829	.1939	.1829	.2010	.1724	.6344
094	.3696	.3762	.3696	.3911	.3630	.9396
095	---- (0.0)*	---- (0.0)*	---- (0.0)*	---- (0.0)*	.0000*	.0000*
096	---- (0.0)*	---- (0.0)*	---- (0.0)*	---- (0.0)*	.0000*	.0002*

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992).

Identification Number	Wilcoxon Exact Test (MO/EX Test)**	Logrank/Savage Scores Exact Test (MO/EX Test)	Normal Scores Exact Test (MO/EX Test)	Permutation Exact Test (MO/EX Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
097	---- (0.0)*	---- (0.0)*	---- (0.0)*	.0000*	.0000*	.0000*
098	---- (0.0)*	---- (0.0)*	---- (0.0)*	---- (0.0)*	.0000*	.0001*
099	---- (0.0)*	---- (0.0)*	---- (0.0)*	---- (0.0)*	.0000*	.0000*
100	---- (0.0)*	---- (0.0)*	---- (0.0)*	---- (0.0)*	.0000*	.0001*
101	.0000*	---- (0.0)*	.0000*	.0000*	.0000*	.0000*
102	---- (0.0)*	---- (0.0)*	---- (0.0)*	---- (0.0)*	.0000*	.0000*
103	.4691	.4926	.4691	.4838	.4520	.9073
104	.3824	.2651	.3417	.3167	.3814	.7407
105	.4852	.4943	.4878	.4949	.4848	.9999
106	.2804	.2551	.2754	.2776	.2803	.7979
107	.2868	.2580	.2812	.2832	.2866	.7988
108	.3600	.1486	.3265	.3159	.3597	.0681

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992).

Identification Number	Wilcoxon Exact Test (MO Test)**	Logrank/Savage Scores Exact Test (MO Test)	Normal Scores Exact Test (MO Test)	Permutation Exact Test (MO Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
109	.2849	.1507	.2637	.2577	.2848	.2359
110	.2980	.1602	.2760	2705	.2979	.2455
111	.3280	.3801	.3326	.3564	.3250	.8453
112	.3993	.4312	.3993	.4459	.3759	.9365
113	.3565	.3815	.3565	.3712	.3211	.7920
114	.1272	.0892	.1584	.1626	.1255	.3264
115	.2536	.2814	.2751	.2982	.2529	.7118
116	.5257	.5041	.5046	.5257	.5000	1.0000
117	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
118	.0051*	.0047*	.0051*	.0059*	.0039*	.0215*
119	.0023*	.0123*	.0030*	.0077*	.0024*	.0039*
120	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992).

Identification Number	Wilcoxon Exact Test (MO Test)**	Logrank/Savage Scores Exact Test (MO Test)	Normal Scores Exact Test (MO Test)	Permutation Exact Test (MO Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
121	.0034*	.0034*	.0034*	.0034*	.0002*	.0006*
122	.0010*	.0013*	.0010*	.0010*	.0008*	.0051*
123	.0526	.0868	.0518	.0530	.0498*	.1404
124	.4368	.4680	.4369	.4680	.4316	.9644
125	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
126	.5581	.4730	.5581	.5581	.5000	.8956
127	.2151	.2697	.2151	.2195	.1824	.3871
128	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
129	.0269*	.0138*	.0251*	.0260*	.0264*	.0779
130	.4672	.2885	.4697	.4695	.4662	.5258
131	.3995	.4954	.3941	.4255	.3985	.5917
132	.3685	.3681	.3667	.4036	.3281	.9015

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992).

Identification Number	Wilcoxon Exact Test (MO Test)**	Logrank/Savage Scores Exact Test (MO Test)	Normal Scores Exact Test (MO Test)	Permutation Exact Test (MO Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
133	.0742	.0131*	.0845	.0690	.0738	.0308*
134	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
135	.3160	.3571	.3160	.3430	.3154	.8505
136	.1665	.2038	.1419	.1419	.1677	.0000*
137	.4699	.4397	.4706	.4949	.4521	.9.810
138	.2250	.3601	.2786	.2913	.2303	.4336
139	.2150	.0809	.1794	.1794	.1944	.2050
140	.3748	.4329	.3748	.4286	.3489	.9040
141	.3605	.2346	.3444	.3527	.3561	.5897
142	.0383*	.0136*	.0208*	.0382*	.0333*	.0689
143	.4740	.3986	.4302	.5000	.4320	.8977
144	.0767	.0367*	.0534	.0785	.0724	.1733

**Monte Carlo Test estimates at 99.00% level of confidence.

Table 4.6 (cont'd.). *P*-values for Exact Tests Computed with StatXact Turbo by Gajjar, et al. (1992).

Identification Number	Wilcoxon Exact Test (MO Test)**	Logrank/Savage Scores Exact Test (MO Test)	Normal Scores Exact Test (MO Test)	Permutation Exact Test (MO Test)	Wilcoxon Asymptotic Test	Chi Square Asymptotic Test
145	.0000*	.0000*	.0000*	.0000*	.0000*	.0002*
146	.0000*	.0000*	.0000*	.0000*	.0000*	.0000*
147	.3776	.4209	.3776	.4160	.3729	.9146
148	.0016*	.0061*	.0016*	.0028*	.0015*	.0092*
149	.0257*	.0302*	.0257*	.0275*	.0196*	.0178*

**Monte Carlo Test estimates at 99.00% level of confidence.

Additional Analyses

Length of time to perform each test ranged from 0.55 seconds to 7 hours, 54 minutes, 2.57 seconds. Tests which took the longest time (i.e., over six hours) produced an error message, "Problem too large for the test option." Also, tables containing large *c* produced the same error message. A total of 51 tests were unable to run to fruition: 13 Wilcoxon exact tests, 17 Logrank/Savage Score exact tests, 14 Normal Scores exact tests and 7 Permutation exact tests (Table 4.7).

Exact Test	Total Runs	Error Messages
Wilcoxon	149	13
Logrank/Savage Scores	149	17
Normal Scores	149	14
Permutation	149	7

Table 4.7. Frequency of Incomplete Computer Runs on Each Exact Test.

Table 4.8 reports the total significant findings for each test as it relates to the null hypothesis with nominal $\alpha = 0.05$. Column figures place the Permutation Test ahead of the others in its ability to find significance, but this test also failed to reject the most times. The Permutation Test ran more times than the Wilcoxon, Logrank/Savage Scores and Normal Scores tests, which makes it difficult to report which test actually out-performed the others. The tests which were compared in this study had similar outcomes, nevertheless, as reported in the Table 4.8 below. Thirty-nine percent of the tests which successfully ran to completion were significant. An exact Chi Square Test performed on

Table 4.8 to test the null hypothesis that there was no difference in how the four tests performed yielded an exact p-value of .9677, retaining the null.

The prevailing question is whether one exact nonparametric test is more powerful than another for varied ordinal outcome levels in applied research. Recall that the Chi Square test is not sensitive to column magnitudes so it is not considered an appropriate test for ordinal categorical data. That test and the asymptotic Wilcoxon replicated the results obtained in Grissom's study (1994) in which those two statistics were compared. Regarding those tests, there are three observations which must be noted. First, the Wilcoxon rejected the null one time when its exact counterpart did not. Second, the Chi Square rejected the null eight times when the exact tests did not. Third, the Chi Square rejected the null seven times when exact tests would not run.

Exact Test	Number of Times Test Unable to Run	Reject the Null Hypothesis	Failed to Reject the Null Hypothesis
Wilcoxon	13	56	80
Logrank/Savage	17	51	81
Normal Scores	14	51	81
Permutation	7	57	85
TOTAL	**51**	**215 (.39)**	**330 (.61)**
Asymptotic Test			
Wilcoxon	0	68	81
Chi Square	0	62	87
TOTAL	**0**	**130 (.23)**	**168 (.77)**

Table 4.8. Frequency of Significant Results for $\alpha = 0.05$ (%).

To summarize, the Wilcoxon performed slightly better than the Chi Square Test overall, and the exact tests had a larger proportion of rejections than did the asymptotic

tests. A discussion regarding significant results for the asymptotic Wilcoxon and Chi Square in this study is in Chapter Five of this paper.

Table 4.9 carries the question of power one step further than Table 4.8. This table describes the number of rejections for each test at each level of *c*. The table displays the frequency and the percent of *significant* results under three through eight or more levels of ordinal outcomes in real educational and psychological data sets for each of the exact tests for ordinal categorical data. Also, observe that for each column, each test has a similar sample size. It can be quickly surmised by looking at the column figures that the four exact tests differed only slightly overall in their effectiveness at each ordinal level. For example, for data sets with three ordinal levels, each of the tests rejected the null approximately the same number of times: the Wilcoxon, 41; the Logrank/Savage Scores test 40; the Normal Scores test 39 times; and the Permutation test 39 times. Similarities were observed for the other ordinal outcome levels as well.

The *overall* frequencies and percentages ( ) for each test are reported in the "Test" column. For example, out of a total of 215 rejections overall, 56 or 26% were observed with the exact Wilcoxon test. Then, for *each* level of *c*, the frequencies for significant results at $\alpha = 0.0500$ are shown. The percentages in the individual columns are based on the frequency of total significant outcomes for that level of *c*. Percentages are reported here to avoid the misconception that one test performs better than another based on frequency alone. One can see that in the column representing three ordinal levels of outcome data, for example, all tests have similar sample sizes and the percentage differential of rejections of the null hypothesis at that level is nil (25% for each of the four tests). These figures support the Chi Square results.

A Permutation exact test (Procedure #8) was not performed on Table 4.9 to determine whether the four tests are equivalent with respect to the levels of *c* because an error message occurred which disallowed the test calculation. The message read: "Module not available (Program specification not found)." The software package ran the

Permutation test only for 2 x *c* data sets, although Mehta commented, "For ordinal categorical data with 5 to 10 categories and balanced group sizes, exact *p*-values are quickly obtained by StatXact with sample sizes as large as 200" (1992, p. 4 -27). It seems that the Permutation test should run if these criteria are met. In personal communication (August 5, 1996), I asked Mehta how to run this test. The reply was to use the "doubly-ordered data" option. This option is not available with this package. Also, the data to be tested is not doubly-ordered. The ordering occurs only on the Y-axis.

Test *f* **(%)**	**3**	**4**	**5**	**6**	**7**	**8+**
Wilcoxon 56 (.26)	41 (.25)	8 (.29)	7 (.27)	0 (.00)	0 (.00)	0 (.00)
Logrank/Savage Scores 52 (.24)	40 (.25)	7 (.27)	5 (.19)	0 (.00)	0 (.00)	0 (.00)
Normal Scores 51 (.24)	39 (.25)	6 (.22)	6 (.23)	0 (.00)	0 (.00)	0 (.00)
Permutation 56 (.26)	39 (.25)	6 (.22)	8 (.31)	1 (1.00)	0 (.00)	2 (1.00)
TOTALS 215 (1.00)	159 (1.00)	27 (1.00)	26 (1.00)	1 (1.00)	0 (.00)	2 (1.00)

Table 4.9. Frequency (%) of Significant Results for Each Test Under Each Level of *c* for $\alpha = 0.0500$.

The histogram below visually depicts what Table 4.9 describes. One can instantly see that the frequencies are similar for each level compared.

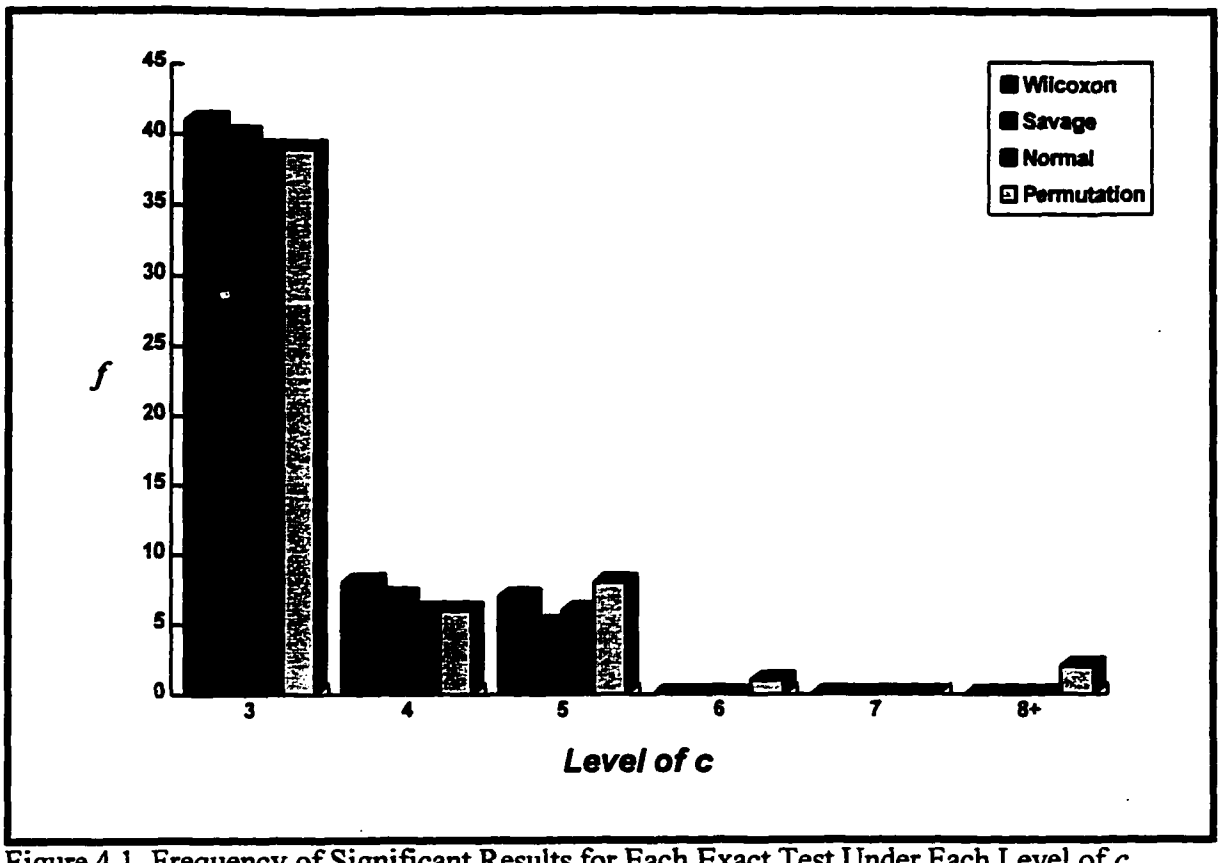

Figure 4.1. Frequency of Significant Results for Each Exact Test Under Each Level of c for $\alpha = 0.0500$.

CHAPTER FIVE

Conclusions

Overview

The debate over theoretical approaches in analyzing data in education and psychology has rendered strong support for nonparametric methods for small, skewed or sparse data. Also, Mehta (1992) emphasized this by computing exact p-values for nonparametric statistics enable the researcher to confidently report reliable results.

The intent of the present study was to compare four nonparametric tests for ordinal categorical data found in real world studies in order to determine the appropriateness of each test in various ordinal situations. Ordinal outcome levels varied from three to eight or more columns. The authentic data sets in education and psychology produced for this study were concrete evidence of nonnormal data found in real world situations. Even large data sets were skewed, justifying the use of nonparametric methods. In addition to a comparison of tests, a modern, user-friendly computer software program demonstrated that the hand-calculations of the past are obsolete and burdensome; and that exact tests can be computed quickly.

Discussion

The Data Sets

The 149 data sets collected exceeded the 30 - 50 data sets anticipated. This finding confirms the theory that ordinal categorical data is frequently used in education and psychology. However, data sets were not as straight-forward as expected. For example, although the four data formats exhibited in Tables 4.1 through 4.4 of Chapter Four of this paper provided appropriate data sets, sometimes manipulation of the data was necessary.

Table 4.5 indicated that most researchers employed three ordinal outcome levels. This result implies that there was an interest on the researcher's part in knowing if there was a change and whether it was a positive or negative one. Degree of change was not as important to these researchers as simply knowing if there was a change and which direction the change occurred. This finding, although not related to the present study, raises the question of whether researchers are sensitive to cut-off points between categories of ordinal data. For example, Researcher A might encompass improved and greatly improved in one category called "Improved," whereas Researcher B might understand improved and greatly improved to be different levels of outcomes, assigning each to separate categories. This question would extend itself to areas of survey research as well, such as market research, where Likert scales are used extensively.

Ordinal Levels

The data consumed six ordinal categories from three to eight or more. This study proposed that there might be a difference in results when a particular test for a particular ordinal outcome size was chosen. This study showed there was nearly no difference in the results when any of the four tests were selected for any number of ordinal levels. Table 4.9 presents both the frequencies and the percentages of times that each test rejected the null at each ordinal level. For example, at the level representing three categories, the level most researchers chose for their studies, there was a difference of only one time that two of the tests did not reject and the other two rejected, and the *percent* of time each of the four tests rejected was virtually the same, 25%. At the level of four, the Wilcoxon outperformed the other tests by a small margin. At the level of five outcome categories, the Permutation test slightly excelled, as it did also at the levels of six (only one sample) and eight or more categories (only two samples). None of the tests rejected the null hypothesis that there was no difference in results of the tests at seven ordinal levels; but, in this case, there was only one sample available.

The Tests

The four nonparametric tests used in this analysis were appropriate for the type of data collected, but the question of which test was most powerful for each situation remains unanswered. All tests provided similar p-values from exact distributions of the sample data, so each one could be said to be "best." Table 4.6 revealed that approximately 39% of the 545 tests which ran were significant. The results (Table 4.7) showed very little difference in numbers of significant outcomes between the tests. In fact, there were times when the proportion of significant outcomes for the tests at various levels of *c* were identical.

The four tests used in this study were exact tests. In other words, they yielded exact p-values.

Limitations

Exact tests, although powerful for analyzing small, skewed, sparse, or heavily tied data, may not always provide the most benefit to the researcher. For example, exact statistics may take considerable time to compute for large data sets, and the benefits of exact inference may be out-weighed by the time restrictions of the research. Mehta (1992) wrote:

> The sample size limits of computational feasibility for exact linear rank tests and odds ratio estimates are difficult to establish. They depend simultaneously on the speed and memory of your computer, the type of scores used, the degree of imbalance in the two groups, the number of ties in the data, and the amount of time you are willing to wait for an answer (p. 4-27).

It was noted in Chapter Four that time to run the exact tests ranged from 0.55 seconds to 7 hours, 54 minutes, 2.57 seconds. Anticipation of the likelihood that data sets will adapt to exact inference quickly is beneficial. Sometimes, in this research, it took nearly eight

hours before receiving an error message. Ultimately, data sets with many ordinal levels proved to take the longest time to analyze.

The question of when a researcher can, or should, dispense with exact methods, opting for asymptotic inference instead was addressed by Mehta (1992):

> ...with highly imbalanced ordinal categorical data, even a sample size as large as 30,000 might not suffice for the asymptotic theory to apply. Fortunately, the more imbalanced the data,, the more efficient the exact algorithms become, so obtaining the exact p-value remains feasible even with large sample sizes (p. 4-28).

Monte Carlo

An assumption of this study, not previously noted, was that all tests would run with StatXact Turbo. As the research and analysis progressed, however, it became evident that some tests would not run for certain large data sets. Given this limitation, I telephoned Mehta (personal conversation, July 5, 1996), co-author of StatXact, to gain advice regarding this problem. Mehta's recommendation was that, for large data sets, the Monte Carlo option should be selected. Mehta (1992) previously stated:

> The Monte Carlo option is a valid alternative to computing either the exact or asymptotic p-values. It is useful for those situations where the data set is too large for an exact p-value computation, yet too sparse to rely on the asymptotic theory. The method consists of sampling three way collections of s 2 x c contingency tables from Γ, repeatedly, and computing an unbiased point estimate and a 99% confidence interval for the exact p-value based on the sample. By increasing the size of the Monte Carlo sample, one can make the width of the confidence interval arbitrarily small and thus effectively guarantee that the Monte Carlo point estimate is accurate to any number of decimal places (p.4-16).

Mehta (1996), later, reiterated this concept, explaining that samples of tables are selected at random from a predefined, true distribution to obtain a range of exact *p*-values for a particular number of iterations. The larger the number of iterations, the closer one will observe the exact *p*-value for the particular table at hand. One may continue sampling to close the range of *p*-values. For example, if the range of exact *p*-values obtained from running the Monte Carlo option for 10,000 iterations is 0.046 to 0.052, one might choose to run more iterations to obtain a smaller range, zeroing in on the exact distribution for the data set, especially if one's alpha level was set at 0.05.

Most Monte Carlo exact test results were significant. In other words, a confidence interval (C.I.) contained a range of p-values within which $\alpha = 0.0500$ occurred. If Mehta's theory about the accuracy of Monte Carlo is correct, then nearly half of the total tests run were significant. The Chi Square for Table 4.8, however, was not conducted to include the Monte Carlo tests and it was not significant at .9677. Consequently, it must be concluded that the observed number of rejections was not significantly different than the expected number. The exact tests, in other words, did not produce a significant number of rejections in this research.

Another limitation of the software was that the planned Permutation test on Table 4.9 (Procedure #8) would not run for an R x *c* table. The module was not available. Therefore, those results could not be obtained.

Cytel Software has since published an updated Windows version of StatXact called StatXact 3.0 for Windows 3.X and 95. This version computes many more exact tests and has expanded other analysis capabilities. As of the time this dissertation was completed, version 2.0 was integrated with the Statistical Package for the Social Sciences (SPSS). Version 3.0 was not available in SPSS, nor would the 2.0 version run under Windows 95 and SPSS 7.0.

Asymptotic Tests

The asymptotic Wilcoxon and the Chi Square tests were calculated for each of the data sets to replicate Grissom's (1994) study. The software had no problem running asymptotic tests, which sometimes took only fractions of a second. Comparatively speaking, these tests performed as well as the exact tests. The tests reported that from 80 to 87 times they failed to reject the null. The Chi Square, as expected due to its generalized approximation of analysis, failed to reject the most number of times the null hypothesis of no difference in the tests' ability to find significance in the data sets. A question arose regarding the significant outcomes of the Chi Square: why did the Chi Square reject the null eight times when the exact tests did not? Even the asymptotic Wilcoxon rejected only one time when the exact tests did not, and it failed to reject the null 81 times when its exact test counterpart failed to reject 80 times. These results could indicate Type I errors. The Chi Square test is not appropriate for ordinal data; but, unfortunately, it is still used in this situation in social and behavioral survey research.

In applied studies, researchers must consider the purpose of the research before selecting a test. If, for example, one is analyzing survival data, the Logrank, or Savage Scores Test for uncensored data, would be appropriate. The Chi Square, we know, is inappropriate for ordinal data, because it is insensitive to degrees of magnitude in the outcome. Although the tests all seemed effective for this research, considering the purpose of the research will lead the researcher to the most appropriate test.

Implications for Future Research

This research indicates that all of the exact nonparametric tests selected perform well for ordinal categorical data. There was no difference in performance when ordinal levels varied. Also, the software, StatXact Turbo, by Mehta is a useful software package for analyzing data with exact tests, and it is simple to learn and use. The newest version,

StatXact 3.0 for Windows, is even more advanced. In this age of technological evolution, there is no excuse for researchers to hand-calculate formulas as Grissom (1994) suggested with the Mann-Whitney U test. In addition, researchers can avoid the challenges of peer review that arise in reporting approximate or asymptotic results, because these are exact tests.

There are several suggestions for future research. First, this study should be repeated utilizing only small samples. The finding that asymptotic tests performed as well as the exact tests must be examined further. If a study using only small samples provided similar results, then the proponents of exact tests must be challenged. Therapy journals or special education journals may be useful sources of data. Application to smaller samples might produce more interesting outcomes.

Second, an extension of this study to R x *c* data sets would be of interest. Do these statistics behave the same way for larger tables?

Third, consider the distributions of data as they relate to the power of these tests. Nonnormal data produces irregular distributional shapes. A view of the test results categorized according to distributional properties of the data would perhaps give some insight to application of appropriate tests for certain distributions.

REFERENCES

American Psychological Association. (1994). Publication Manual of the American Psychological Association (4th ed.). Washington, D.C.: American Psychological Association.

Anastasi, A. (1992). What counselors should know about the use and interpretation of psychological tests. Journal of Counseling and Development, 70, 610-615.

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

Black, S. T. (1993). Comparing genuine and simulated suicide notes: A new perspective. Journal of Consulting and Clinical Psychology, 61(4), 699-702.

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 Wilcoxon's rank-sum statistic to that of Student's t statistic under various nonnormal distributions. Journal of Educational Statistics, 5, 309-335.

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.

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

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

Cherian, V. I. (1992). Relation of parental education and life status to academic achievement by Xhosa children. Psychological Reports, 71, 947-956.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. & Cohen, P. (1975). Applied multiple regression/correlation analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum.

Cohen, P. (1983). To be or not to be: Control and balancing of Type I and Type II errors. Evaluation and Program Planning, 5, 247-253.

Daniel, W. W. & Terrell, J. C. (1995). Business statistics for management and economics (7th ed.). Boston: Houghton Mifflin.

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. Journal of Counseling and Clinical Psychology, 62, 75-82.

Edgington, E. S. & Khuller, P. L. V. (1992). A randomization test computer program for rends in repeated-measures data. Educational and Psychological Measurement, 52, 93-95.

Fisher, R. A. & Yates, F. (1938). Statistical tables for biological, agricultural and medical research (1st ed.). Edinburgh: Oliver & Boyd.

Gajjar, Y., Mehta, C. R., Patel, N. & Senchaudhuri, P. (1992). StatXact-Turbo statistical software for exact nonparametric inference user manual. MA: Cytel Software.

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.

Grissom, R. J. (1994). Statistical analysis of ordinal categorical status after therapies. Journal of Consulting and Clinical Psychology, 62, 281-284.

Haber (1990). Comments on *The test of homogeneity for 2 x 2 contingency tables: A review of and some personal opinions on the controversy* by G. Camilli. Psychological Bulletin, 108, 146-149.

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

Hammond, K. R. & Householder, J. E. (1963). Introduction to the statistical method: Foundations and use in the behavioral sciences. New York: Knopf.

Hillman, S.B. & Sawilowsky, S. S. (1992). A comparison of methods in distinguishing levels of substance abuse. Journal of Clinical Child Psychology, 21, 348-353.

Hollander, M. & Wolfe, D. A. (1973). Nonparametric statistical methods. New York: Wiley.

Hunter, M. A. & May, R. B. (1990). Some myths concerning parametric and nonparametric tests. In B. Zumbo (Chair), Alternatives to classical statistical procedures. Symposium conducted at the annual meetings of the Canadian Psychological Association, Ottowa. (Abstract in Canadian Psychology, 31, 238).

Jenkins, S. J., Fuqua, D. R., & Froehle, T. C. (1984). A critical examination of use of non-parametric statistics in the *Journal of Counseling Psychology.* Perceptual and Motor Skills, 59, 31-35.

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

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

Lehmann, E. L. (with D'Abrera, H. J. M.). (1975). Nonparametrics: Statistical methods based on ranks. San Francisco: Holden-Day.

Levine, D. M., Ramsey, P. P., & Berenson, M. L. (1995). Business statistics for quality and productivity. Englewood Cliff, NJ: Prentice Hall.

Mann, H. B. & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18, 50-60.

May, R. B. & Hunter, M. A. (1990). Some advantages of permutation tests. In B. Zumbo (Chair), Alternatives to classical statistical procedures. Symposium conducted at the annual meetings of the Canadian Psychological Association, Ottawa. (Abstract in Canadian Psychology, 31, 238). [Companion paper to *Some myths concerning parametric and nonparametric tests* by M. Hunter and R. May, 1990].

McCall, R. B. (1980). Fundamental statistics for psychology (3rd ed.). San Diego: Harcourt Brace Jovanovich.

Mehrens, W. A. (1978). Rigor and reality in counseling research. Measurement and Evaluation in Guidance, 11, 8-13.

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

Morrow, Van L. (1995). Teachers evaluate psychological problems and personal counseling needs of students. Education, 116 (1), 130-139.

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

Omer, H. & Dar, R. (1992). Changing trends in three decades of psychotherapy research: The flight from theory into pragmatics. Journal of Consulting and Clinical Psychology, 60, 88-93.

Petrinovich, L. (1979). Probabilistic functionalism: A conception of research method. American Psychologist, 34, 373-390.

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any populations. Journal of the Royal Statistical Society (Series B), 4, 119-130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any populations: II. Journal of the Royal Statistical Society (Series B), 4, 225-232.

Ripstra, C. C. (1974). The quality of experimental methodology in counseling and counselor education. (Unpublished doctoral dissertation, Michigan State University).

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? Journal of Consulting and Clinical Psychology, 58, 646-656.

Runyon, R. P. & Haber, A. (1991). Fundamentals of Behavioral Statistics (7th ed.). New York: McGraw-Hill.

Savage, R. (1956). Contributions to the theory of rank order statistics: The two-sample case. Annals of Mathematical Statistics, 27, 590-615.

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

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. & Hillman, S. B. (1992). Power of the independent samples t test under a prevalent psychometric measure distribution. Journal of Consulting and Clinical Psychology, 60, 240-243.

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

Speer, D. C. (1994). Can treatment research inform decision makers? Nonexperimental method issues and examples among older outpatients. Journal of Consulting and Clinical Psychology, 62, 560-568.

Strategy Plus, Inc. (1993). Execustat 3.0 (Student edition). Belmont, CA: Duxbury Press.

Thoreson, C. E. (1969). Relevance and research in counseling. Review of Educational Research, 39, 263-281.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics, 1, 80-82.

ABSTRACT

COMPARATIVE PROPERTIES OF NONPARAMETRIC STATISTICS FOR THE ANALYSIS OF THE 2 x *c* LAYOUT FOR ORDINAL CATEGORICAL DATA

by

MARGARET A. POSCH

December 1996

Advisor: Dr. Shlomo Sawilowsky

Major: Evaluation and Research

Degree: Doctor of Philosophy

Researchers in Education and Psychology often encounter skewed, sparse, or small data sets, called nonnormal, which are ordinal categorical in nature. A proper approach to analysis of these data kindled a debate which has concluded, for the most part, that nonparametric statistical tests outperform their parametric counterparts for many situations. Yet, researchers continue to use normal theory statistics to solve problems with nonnormal data. This paper discusses the virtues of nonparametric statistics for analysis of nonnormal data and applies four nonparametric tests appropriate for ordinal categorical data sets: the Wilcoxon Mann-Whitney U test, the Normal Scores test, the Savage Scores test, and the Permutation test. The properties of these tests are compared for various situations of sample size, alpha levels, and number of ordinal categories in outcome data. Twenty-nine journals from the fields of Education and Psychology were reviewed over a four-year period to collect real data sets.

In addition to identifying the most appropriate test statistic for each situation, this paper brings the researcher well into the 1990s through its advocacy of modern, user-friendly computer software programs which educators and psychologists will not find intimidating and which will provide quick, simple, and accurate analysis of their data. Specifically, the paper promotes the use of StatXact Turbo 2.0 (Gajjar, et al., 1992) for

use with ordinal categorical data. This program produces exact permutations of p-values for each data set, providing the researcher with reliable results, an issue of paramount importance when prescribing treatment for the welfare of human subjects.

AUTOBIOGRAPHICAL STATEMENT
MARGARET A. POSCH

WORK EXPERIENCE

Wayne State University, Detroit, MI

1995 - present:	Research Assistant, College of Education
1994 - 1995	Graduate Research Assistant, College of Education
	Graduate Assistant, School of Business Administration
1993 - 1994	Part-time Student Assistant, College of Education
	Part-time Faculty, School of Business Administration

The Prudential Grosse Pointe Real Estate Company, Grosse Pointe, MI

1991 - 1993	REALTOR Associate

Lakeshore Public Schools, St. Clair Shores, MI

1967 - 1969	Elementary Teacher

Farmington Public Schools, Farmington, MI

1966 - 1967	Elementary Teacher

EDUCATION

Wayne State University	**Ph.D.**, Evaluation and Research (Statistics)
1992 - 1996	G.P.A. = 3.88
	Doctoral Dissertation: Comparative Properties of Nonparametric Statistics for the Analysis of the 2 x *c* Layout for Ordinal Categorical Data.
Wayne State University	**M.Ed.**, Educational Evaluation and Research
1982 Graduation	G.P.A. = 4.00
	Master's Project: Computerizing the Normal Curve (BASIC programming language)

University of Detroit Mercy **B.A.**, Elementary Education/Speech and Drama
1966 Graduation

PUBLICATIONS/REPORTS

Posch, M. & Wolf-Branigan, M. (1994). Qualitative analysis of self-determination among residents of a long-term substance abuse program. Journal of Rehabilitation and Recovery: The Salvation Army/Western Territory USA, 2 (4), 33-38.

Sawilowsky, Ehrich, Posch & Okafor. (1994). Qualitative analysis of open-ended responses to structured interview questions: quantitative analysis of structured interview questions; quantitative analysis of self-determination scale. (Report to Salvation Army Harbor Light Center: responses to project client interviews regarding self-determination, Detroit, MI).