University of South Carolina
Scholar Commons

Theses and Dissertations

Fall 2023

Nonparametric Tests of Interaction for the Two-Way Design With Skewed Distributions

Michael Ethan Hornsby Brown

Recommended Citation

Hornsby Brown, M. E.(2023). *Nonparametric Tests of Interaction for the Two-Way Design With Skewed Distributions.* (Doctoral dissertation). Retrieved from https://scholarcommons.sc.edu/etd/7596

NONPARAMETRIC TESTS OF INTERACTION FOR THE TWO-WAY DESIGN WITH SKEWED DISTRIBUTIONS

by

Michael Ethan Hornsby Brown

Bachelor of Science
University of South Carolina Aiken, 2015

Master of Applied Statistics
University of South Carolina, 2019

Submitted in Partial Fulfillment of the Requirements

For the Degree of Doctor of Philosophy in

Educational Psychology and Research

College of Education

University of South Carolina

2023

Accepted by:

Michael Seaman, Major Professor

Christine DiStefano, Committee Member

Angela Starrett, Committee Member

Ryan Carlson, Committee Member

Ann Vail, Dean of the Graduate School

ACKNOWLEDGEMENTS

I would like to extend my most heartfelt appreciation to my advisor, Dr. Michael Seaman, for his wisdom, knowledge, and patience throughout this process. I feel incredibly fortunate to have such an advisor throughout this process. A special thank you to the members of my committee as well for their guidance and feedback in pushing my work to be the best it could be.

I would also like to extend a special thanks to my assistantship advisor, Dr. Vasanthi Rao, for her wisdom, support, motivation, and patience throughout my time as a student and research assistant. These were invaluable resources that had monumental impact.

I am deeply grateful for the patience and support from my husband, Kenny. Thank you to my friends, especially Brent and Brandy, who were supportive from beginning to end. Thank you to my family. Finally, thank you to my grandmother, Barbara, who is not here to see the end of this journey but was and is paramount in my becoming who I am today.

ABSTRACT

The most common parametric procedure used to test main and interaction effects in the two- or more-groups factorial design is the analysis of variance (ANOVA) *F* test. Researchers in the behavioral and social sciences fields require statistical methods that are robust in the presence of deviations from the common parametric ANOVA assumptions of (a) normality, (b) homogeneity of variances among groups, and (c) independence of observations. When there is concern that the parametric assumptions are violated, nonparametric procedures can be employed that do not make as many initial assumptions about the parent populations. Of particular interest in the two-factor design is the test for interaction among the factors. This paper seeks to contribute to the research of nonparametric methods by exploring the properties of various nonparametric tests in detecting and inferring interaction effects when the population distributions are skewed or asymmetrical. A review of rank-based nonparametric tests for interaction is provided to determine what methods have been proposed to test for interaction and how these methods have performed in comparative research studies. This review includes research findings regarding normal scores to determine the potential of a normal scores transformation when testing for interaction. A comparative study of nonparametric methods that have been shown by past research to provide reasonable power and Type I error control is conducted to determine if these methods also perform well when testing for interaction effects using Monte Carlo simulated data with skewed and asymmetric distributions. A second comparative study explores the performance of three novel nonparametric tests for interaction. These three tests are the rank transform test, aligned rank transform test, and McSweeney test with a Van der Waerden normal scores transformation in place of a rank transformation. For the studied designs, the aligned rank transform test, the aligned rank transform test using normal scores, and the McSweeney test using normal scores provide nonparametric tests of interaction that maintain Type I error rate, are as powerful as the ANOVA *F* test when the underlying population is normal, and have more power when the underlying population is not normal.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

F	F statistic
t	t statistic
U	U statistic of the Mann-Whitney test
L	Puri and Sen L *s*tatistic
T^2	Hotelling’s statistic
n	Cell sample size
N	Total sample size
Φ	Standard normal distribution function
χ^2	Chi-square distribution
W	Normal scores chi-square test statistic
μ	Population mean
H_0	Null hypothesis
H_Y	McSweeney statistic
T	Thompson rank transform statistic
α	Type I error rate
SS	Sum of squares
df	Degrees of freedom

LIST OF ABBREVIATIONS

ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
ARE	Asymptotic Relative Efficiency
ART	Adjusted Rank Transform or Aligned Rank Transform
A-V-*F*	Alignment used with Van der Waerden Normal Scores and an *F* statistic
A-V-M	Alignment with Van der Waerden Normal Scores and McSweeney statistic
ENST	Expected Normal Scores Transformation
IPS	Interaction Probability of Superiority
MANOVA	Multivariate Analysis of Variance
QANOVA	Quantile-Based Permutation Analysis of Variance
RNST	Random Normal Scores Transformation
RT	Rank Transform
V-*F*	Van der Waerden Normal Scores Transformation with an *F* statistic

CHAPTER 1

INTRODUCTION

The most common parametric procedure used to test main and interaction effects in the two- or more-groups factorial design is the analysis of variance (ANOVA) *F* test (Jenkins et al., 1984; Sawilowsky, 1990). In this research design, there are participants in all combinations of levels for the two or more categorical predictor variables. For each participant, the level of each predictor variable as well as the numerical response variable of interest is recorded. The goal is to infer if sample group differences in central tendency are representative of true group differences in central tendency of the parent populations. This method of inference is critical in research in the education and social science fields where researchers often want to generalize treatment effects from observed samples to target populations.

The parametric *F* test relies on population assumptions of (a) normality, (b) homogeneity of variances among groups, and (c) independence of observations. There is concern regarding power and Type I error rate robustness of the parametric *F* test in the presence of violations to these assumptions (Bishop, 1976; Blair, 1980; Blair & Higgins, 1980a, 1980b, 1981; Box, 1953, 1954; Bradley, 1968, Brown & Forsythe, 1974; Goodard & Lindquist, 1940; Hornsnell, 1953; Lindquist, 1953; Randolph & Barcikowski, 1989; Rogan & Keselman, 1977; Scheffe, 1959; Snedecor & Cochran, 1980; Tomarkin & Serlin, 1986; Welch, 1937; Wilcox et al., 1986). This concern is exacerbated by the lack of a standard quantitative definition of robustness (Blair & Higgins, 1981; Bradley, 1978). When there is concern that the parametric assumptions are violated, nonparametric procedures can be employed that do not make as many initial assumptions about the parent populations. These nonparametric procedures are critical to applied researchers in the education and social science fields where virtually no measures are found to be normal (Micceri, 1989).

A “nonparametric test” is a statistical test that is not conditioned on specified parameters of the target population of interest in order to make valid inference. Most common nonparametric tests fall into one of two types: (1) tests of a categorical response variable and (2) tests of a ranked response variable. A ranked variable can be viewed as an ordered categorical variable, so these two types could collectively be considered as the analysis of a categorical variable (Nussbaum, 2015; Stevens, 1946). Rank transformations of a numerical response variable provide a way for researchers who suspect the common parametric assumptions have been violated to perform tests of differences in central tendency.

Rank transformations as a data transformation can be used independently or in combination with other methods of data transformation. When transforming two or more sets of scores to ranks for the purpose of performing a statistical test, the sets of scores are commonly combined prior to ranking (Conover & Iman, 1976, 1981; Iman, 1974; Iman and Conover, 1976). While transforming one set of scores to ranks will produce a distribution that is rectangular, when transforming two or more sets of pooled scores to ranks the distributions of the individual set ranks will no longer be necessarily rectangular. Differences in group measures of variance, skew, kurtosis, and bimodality are retained after methods of rank transformation albeit with reduced magnitude (Zimmerman, 2011).

Rank based nonparametric procedures demonstrate potential in two- or more-sample tests of difference in central tendency in the one-factor design. The Mann-Whitney *U* test, the two-sample test of location shift using ranks (also referred to as the Wilcoxon rank sum test), demonstrates performance more powerful than the *t* test with many nonnormal data distributions (Blair, 1980; Blair & Higgins, 1985; Mann & Whitney, 1947; Randles & Wolfe, 1979; Smitley, 1981; Wilcoxon, 1945, 1947). The multiple-group extension of the two-sample Mann-Whitney *U* test for the one-factor design is the Kruskal-Wallis test (Kruskal & Wallis, 1952). The rank-based Kruskal-Wallis test demonstrates power comparable or greater than the parametric one-way ANOVA *F* test when groups have identical distributions shapes from nonnormal distributions (Andrews, 1954; Conover, 1980; Kruskal, 1952; Hodges and Lehmann, 1956).

Statement of the Problem

Researchers in the behavioral and social sciences fields require statistical methods that are robust in the presence of deviations from the common parametric ANOVA assumptions. Real world education and psychological measures often occur having mixed-normal distributions (Bradley, 1978, 1980a, 1980b, 1980c, 1982; Blair, 1980; Micceri, 1989; Still & White, 1981; Tan, 1982). Education and social sciences measures, such as sum scores from measures using Likert scales, often have ceiling and floor effects that result in the need for methods capable of handling censored and truncated distributions (DeWees et al., 2020; Feng et al., 2019; Liu & Wang, 2021; McBee, 2010).

Other frequent departures from normality found in psychometric measures include asymmetry, multimodality, and skew (Bradley, 1977; Micceri, 1989).

For example, Micceri's (1989) meta-analysis analyzed 440 large sample educational and psychological measures and found all to be significantly nonnormal via some class of contamination. Of the 440 measures analyzed by Micceri, only 15.2% had both tail weights at or about normal, 50.2% had at least one tail heavier than normal, and 14.8% had both tail weights less than normal. Of the remainder, 3.2% were concluded as uniform and 16.6% were concluded as Laplace. The Laplace distribution, or double exponential distribution, is a continuous distribution symmetric about the location parameter with a higher peak, fatter tails, and lower shoulders than the normal distribution. From Figure 1, we note the plot of a standard normal distribution: a normal distribution with a mean of 0 and a variance of 1. We can use this to compare against a Laplace distribution with a location parameter of 0 and a scale parameter of 1 (Figure 2).

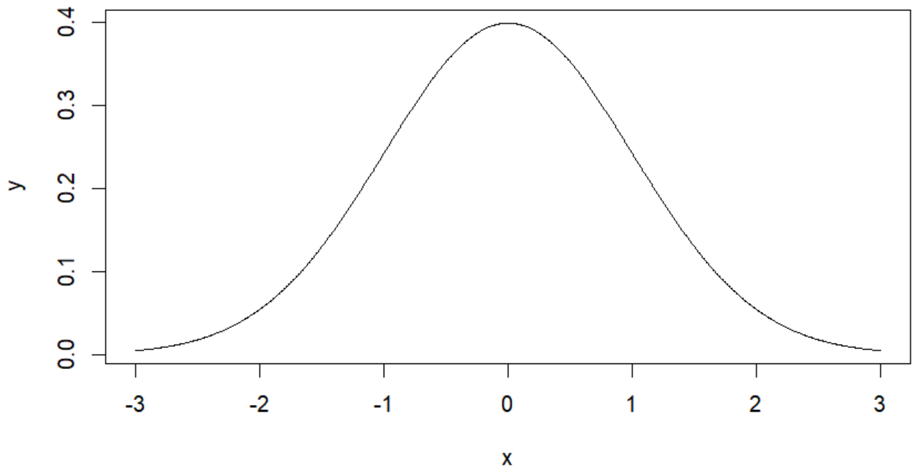

Figure 1.1
Plot of a Standard Normal Distribution

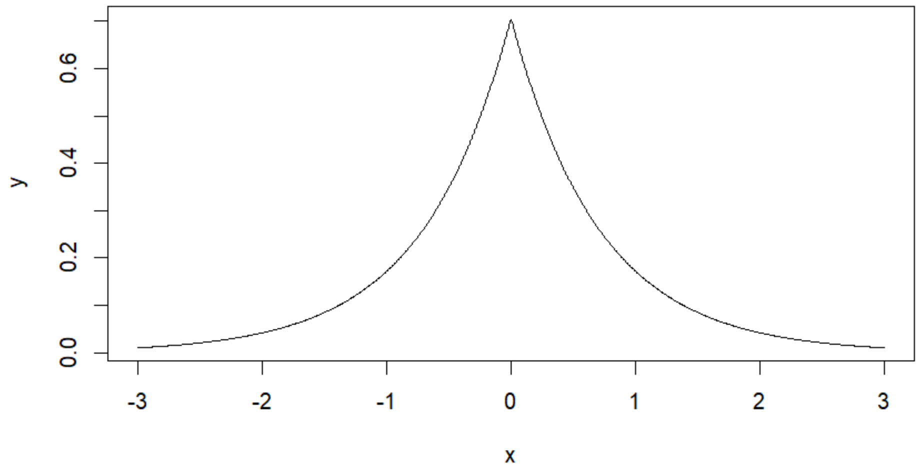

Figure 1.2
Plot of a Laplace(0, 1)

Additionally, of the 440 measures, 40.7% were concluded as moderately asymmetric, 19.5% as extremely asymmetric, and 11.4% as having exponential asymmetry. Of the categories for the measures analyzed by Micceri, nonnormal tail weights and asymmetry is demonstrated most prevalent in measures of criterion mastery, where 60% of measures were found to be distributed as Laplace, 37.1% of measures had extreme asymmetry, and 57.1% of measures had exponential asymmetry. Thus, Micceri suggests the existence of normally distributed data in educational and psychological measures is as improbable as the existence of a unicorn.

Micceri's (1989) findings not only directly conclude the prevalence of the Laplace distribution in education and social science measures but also provide evidence of the prevalence of the logistic distribution by highlighting the frequency of mixed-normal measures and criterion mastery measures. The logistic distribution is often used to approximate mixed-normal distributions, model distributions of subjective quantitative measures, and is prevalent in item response theory analysis (Hellstrom, 1993; Noortgate et al., 2003; Savalei, 2006). The logistic distribution is also a continuous, symmetric distribution with thicker tails and a higher peak than the normal distribution, though the tails are not as thick and the peak not as high as that of the Laplace distribution. For reference, Figure 3 plots a logistic distribution with a location parameter of 0 and a scale parameter of 1.

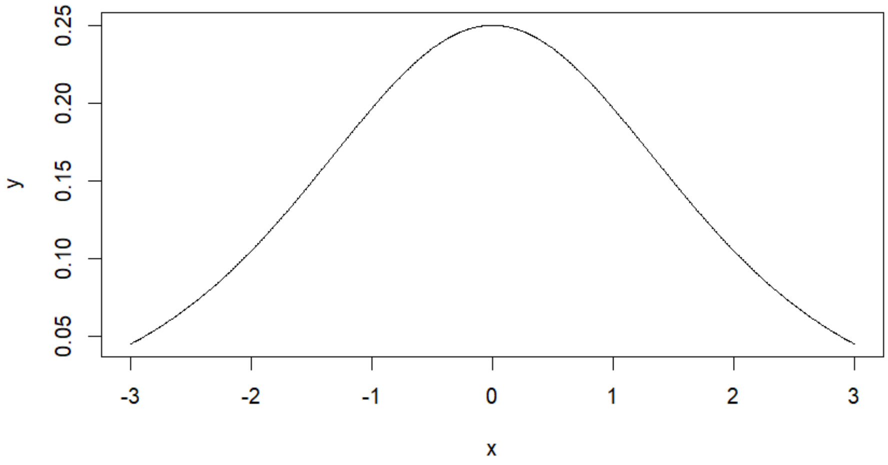

Figure 1.3
Plot of a Logistic(0, 1)

This would imply that researchers in these fields who limit their methods of analysis to those dependent upon the assumption of normality will rarely be using methods as robust and powerful as available alternative nonparametric methods of analysis. Given the prevalence of measures distributed as Laplace, measures distributed as logistic, and measures having varying degrees of asymmetry, methods of analysis capable of handling these variables are crucial for researchers in education and social science fields.

Of particular interest in the two-factor design is the test for interaction. It has been argued that the few proposed nonparametric tests capable of testing for interaction effects are either computationally intensive or less powerful, leaving only the parametric ANOVA *F* test available for researchers to use when testing for interaction effects (Gaito, 1959; Gardner, 1975). Sawilowsky (1990) offers a comprehensive review of rank-based nonparametric tests for interaction proposed in the literature for behavioral and social science researchers. Sawilowsky comprehensively reviews the literature for ten nonparametric tests for interaction. In the time since publication of Sawilowsky's review, there has been no additional study of some of the tests discussed in the review. Simulation studies have been conducted on some of the other tests, though the findings of these studies were not encouraging as the tests yielded lower power than other available methods or were found to be too computationally intensive for the technology available to the average researcher situated in the behavioral and social sciences at the time. Further, this research focused primarily on designs with small cell sample sizes or heterogeneity of variance between groups. To date, no research has been performed to determine the value of these methods when the population distributions have multimodality, skew, and asymmetry: deviations from normality found frequently occurring in the behavioral and social sciences (Micceri, 1989).

Of the methods reviewed by Sawilowsky (1990), the rank transform test and the adjusted rank transform test demonstrated the most potential given their demonstrated power and robustness properties and thus warrant further study. The rank transform test is performed by pooling and ranking all observations prior to calculating the parametric statistic on the ranks (Conover & Iman, 1976, 1981; Iman, 1974; Iman and Conover, 1976). The rank transform test is robust in 2 x 2 designs under the normal distribution but lacks power compared to the parametric *F* test (Sawilowsky, 1989). Further research is needed to determine the power of the rank transform test compared to the parametric *F* test for skewed and asymmetric distributions.

Thompson (1991) conducted research into the asymptotic properties of the rank transform test for 2 x 2 designs and suggested that a chi-square test statistic, rather than an *F* statistic, should be used for the critical value for such tests. This suggestion requires further study.

The adjusted rank transform test is performed by aligning the data prior to pooling and ranking observations. This is a modification of the alignment procedure outlined by Hodges and Lehmann (1962). The usual parametric test statistic is then calculated using aligned and rank-transformed data. The outcomes of simulation studies of the adjusted rank transform test demonstrate strong power and slightly liberal Type I error rates (Blair & Sawilowsky, 1990). The McSweeney test (1967) is to align and rank as well, but a chi-square test statistic, rather than an *F* statistic, is used for a test of the hypothesis. While Sawilowsky does not comprehensively review the McSweeney test, it is mentioned as a reference for other rank-based tests. Studies conducted since Sawilowsky's 1990 review demonstrate that the McSweeney test has strong power, yet at the cost of liberal Type I error rates for small sample sizes (Harwell, 1991; Toothaker & Newman, 1994).

Some of the methods Sawilowsky (1990) reviewed are normal scores-based tests. Studies of these tests suggest strong potential for Type I error rate control and power, but further study is needed. There are three types of normal scores: Terry-Hoeffding normal scores (Hoeffding, 1951; Terry, 1952), Van der Waerden normal scores (Van der Waerden, 1952, 1953, 1956), and Bell-Doksum normal scores (Bell & Doksum, 1965). Using normal scores can result in strong performance in the one-way two- or more-sample test of differences in central tendency (Hodges & Lehmann, 1961; McSweeney, 1967; Penfield & McSweeney, 1968). McSweeney (1967) suggests applying normal scores transformations in place of ranking for both her proposed inferential test as well as for other tests of interaction.

My intent in writing this dissertation is to contribute to the research of nonparametric methods by exploring the properties of various nonparametric tests in detecting and inferring interaction effects when the population distributions are skewed or asymmetrical. The following research questions will guide the investigation:

1. What does published literature regarding proposed nonparametric tests of interaction in the factorial analysis of variance design demonstrate regarding their asymptotic properties, power, and Type I error rates in the presence of nonnormally distributed data, varying cell sample sizes, and heterogeneity of variances?
2. How do nonparametric tests of interaction that have already been demonstrated robust in the 2 x 2 factorial analysis of variance design perform for distributions having varying degrees of skew and asymmetry?
3. Can normal scores be used in combination with alignment-based data transformation procedures to provide powerful and robust nonparametric tests of interaction effects?

I will conduct a comprehensive review of nonparametric tests for interaction to determine what methods have been proposed to test for interaction and how these methods have performed in comparative research studies. I will also review research findings regarding normal scores to determine the potential of a normal scores transformation when testing for interaction. In two empirical articles, I will do a comparative study of nonparametric methods that have been shown by past research to provide reasonable power and Type I error control to determine if these methods also perform well when testing for interaction effects using Monte Carlo simulated data with skewed and asymmetric distributions. In the first empirical article, I will compare power and Type I error rates of four rank-based nonparametric tests of interaction: the rank transform test, the aligned rank transform test, the McSweeney test, and the Thompson rank transform test. In the second empirical article, I will explore the performance of three novel nonparametric tests for interaction. These three tests are the rank transform test, aligned rank transform test, and McSweeney test with a Van der Waerden normal scores transformation in place of a rank transformation. I will highlight strengths and weaknesses of proposed nonparametric tests of interaction for varying cell sample sizes and in the presence and absence of significant main effects.

CHAPTER 2

A REVIEW OF NONPARAMETRIC TESTS OF INTERACTION

In 1990, *Review of Educational Research* published Shlomo Sawilowsky's article titled "Nonparametric Tests of Interaction in Experimental Design." A "nonparametric test" is a statistical test that is not conditioned on specified parameters of the underlying population distribution in order to make valid inferences (Hettmansperger et al., 2000; Marascuilo & Serlin, 1988; Nussbaum, 2015; Sawilowsky, 1990). The most common nonparametric tests can broadly be considered one of three types: (1) tests of a categorical response variable, (2) tests of a ranked response variable, or (3) randomization tests based on the sign of the response variable. Sawilowsky's review focuses on addressing the power, Type I error rate, and robustness of rank-based nonparametric methods. A ranked variable can be viewed as an ordered categorical variable, so these two types could collectively be considered as the analysis of a categorical variable (Nussbaum, 2015; Stevens, 1946). Due to this relationship, this review will primarily focus on rank-based and category-based nonparametric tests.

Sawilowsky offers two primary reasons for reviewing nonparametric tests of interaction. (1) Nonnormally distributed variables are prominent in experimental research, particularly in the fields of the behavioral and social sciences (Blair & Higgins, 1981; Bloom, 1984; Bradley, 1968, 1977; Walberg et al., 1984), and many nonparametric methods do not rely on a condition of normally distributed data for valid inference (Nunnally, 1975, 1978). (2) Researchers often question the necessity and viability of nonparametric methods of analysis. This skepticism is prompted by beliefs that may be incorrect or only partially correct. These include the belief that: (a) common parametric methods of analysis are usually robust to violations of parametric assumptions regarding population distributions (Boneau, 1960; Box, 1954; Glass et al., 1972; Lindquist, 1953), (b) nonparametric methods are less powerful than parametric analogues (Kerlinger, 1973; Nunnally, 1975), and (c) nonparametric methods are not available for addressing data obtained from complicated research and experimental designs (Bradley, 1968). Each of these concerns over the viability of nonparametric methods of analysis warrants further discussion.

There is evidence to suggest nonnormally distributed populations are common in the behavioral and social science fields. A high frequency of mixed-normal distributions in real world data is observed in education and psychological measures (Bradley, 1978, 1980a, 1980b, 1980c, 1982; Blair, 1980; Micceri, 1989; Still & White, 1981; Tan, 1982). Ceiling effects and floor effects in measurement ranges provide for the existence of censored-normal and truncated-normal distributions in education and social science measures (DeWees et al., 2020; Feng et al., 2019; Liu & Wang, 2021; McBee, 2010). These effects occur frequently with the use of Likert scales and other self-reported measures. Other frequent departures from normality found in psychometric measures include asymmetry, multimodality, and skew (Bradley, 1977; Micceri, 1989). The presence of deviations from normality alone does not necessitate research into alternatives to traditional parametric tests without evidence that the performance of these parametric tests suffers in the presence of these deviations.

The concern over robustness of parametric tests to departures from parametric assumptions about data distributions has yielded extensive study. Claims and conclusions regarding the robustness of parametric tests are subject to a degree of variability as there is no commonly agreed upon quantitative definition or standard of robustness (Blair & Higgins, 1981; Bradley, 1978). Consequently, claims of robustness are left to the conclusion of the author rather than the attainment of a standardized performance. Traditional parametric tests of location equality among two or more groups, such as the *t* test and analysis of variance (ANOVA) *F* test, build on population assumptions that include a normally distributed population, independence of observations, and homogeneity of group variances (Fisher, 1922, Hildebrand, 1986; Winer, 1971). The *t* test demonstrates reasonable to strong robustness to minimal departures from normality in some circumstances (Boneau, 1960; Sawilowsky & Blair, 1992), but fails to maintain reasonable robustness in others, such as when distributions are extremely skewed or when normality deviations occur in the presence of small group sample sizes (Blair, 1980; Blair & Higgins, 1980a, 1980b, 1981; Bradley, 1968; Hemelrijk, 1961). Some argue that the ANOVA is robust to various minor deviations from population normality (Baker et al., 1966; Boneau, 1960, 1962; Box, 1954; Cochran, 1947; Cochran & Cox, 1950; Fisher, 1922; Glass et al., 1972; Goodard & Lindquist, 1940; Guilford & Fruchter, 1978; Hack, 1958; Havlicek & Peterson, 1974; Lunney, 1970; Mandeville, 1972; Pearson, 1929; Rider, 1929; Scheffe,1959), while others express skepticism (Blair, 1980; Blair & Higgins, 1980a, 1980b, 1981; Bradley, 1968). Another common factor analyzed in discussion of the robustness question is equality, or lack thereof, in group sample sizes. Error rates for ANOVA inflate with violations to the homogeneity of variance assumption

(Bishop, 1976; Box, 1954; Brown & Forsythe, 1974; Goodard & Lindquist, 1940; Hornsnell, 1953; Lindquist, 1953; Welch, 1937; Wilcox et al., 1986), increasingly so in the presence of unequal group sample sizes (Box, 1953; Randolph & Barcikowski, 1989; Rogan & Keselman, 1977; Scheffe, 1959; Snedecor & Cochran, 1980; Tomarkin & Serlin, 1986). Although there is evidence to suggest lack of robustness for common parametric tests with violations to common parametric data assumptions, viability of nonparametric tests as preferred alternatives requires their demonstrated competitive performance in the presence of these assumption violations.

Early studies demonstrated strong potential in the viability of rank based nonparametric test procedures, often performing more powerfully than the *t* test in two-sample research designs in the presence of nonnormal data (Chernoff & Savage, 1958; Dixon, 1954; Hodges & Lehmann, 1956; Neave & Granger, 1968). The Mann-Whitney *U* test, also referred to as the Wilcoxon rank sum test, has more power than the *t* test with various nonnormal population distributions (Blair, 1980; Blair & Higgins, 1985; Mann & Whitney, 1947; Randles & Wolfe, 1979; Smitley, 1981; Wilcoxon, 1945, 1947). When the normality and homogeneity of variance assumptions are satisfied, a randomization test of the original scores performs as efficiently as the *t* test (Hoeffding, 1952; Lehmann & Stein, 1959). Alterations to the *t* test by first transforming the data has the potential to avoid loss of power when the standard conditions for valid inference are violated. (Rasmussen, 1985, 1986).

The multiple-group extension of the two-sample Mann-Whitney test is the Kruskal-Wallis test (Kruskal & Wallis, 1952). The Kruskal-Wallis test is a rank-based alternative to the one-way ANOVA. The Kruskal-Wallis test has power comparable or greater than the one-way ANOVA when groups have identical distribution shapes, including nonnormal distributions, with asymptotic relative efficiencies (AREs) ranging from 0.864 to infinitely larger than 1. (Andrews, 1954; Conover, 1980; Kruskal, 1952; Hodges and Lehmann, 1956). Similar to ANOVA, the Kruskal-Wallis test demonstrates inflated error rates when the homogeneity of variance assumption is violated, increasingly so in the presence of unequal group sample sizes (Feir-Walsh & Toothaker, 1974; Keselman et al., 1977; Tomarkin & Serlin, 1986). The multiple-group extension of the sign test is the Friedman test, which has much more variable power relative to the ANOVA than the Kruskal-Wallis test but still maintains the ARE above 1 for some distributions (Friedman, 1937; Hollander & Wolfe, 1973; Noether, 1967; Van Elteren & Noether, 1959). The Friedman test also suffers inflated error rates from violating the homogeneity of variance assumption (Harwell & Serlin, 1989b).

Some researchers do not believe that nonparametric methods are available or appropriate to use for more complicated research designs. One reason for this belief is that nonparametric methods, including rank-based procedures, do not capture and use the full scope of information available in the data and consequently result in lower power than parametric procedures (Wilcoxon, 1949). This myth can be dispelled with the demonstratable power studies, though further study of these methods is warranted. Historically, nonparametric tests for interactions have been considered complicated or computationally intensive (Bradley, 1968). As technology advances, these limitations can be iteratively revisited to check if they remain justified. Nonparametric methods testing for an interaction effect using ordinal categorical variables often focus on the median, which provides a test of location shift if the distributions have the same form, or a test of equivalent distributions if distributional form is unknown (Brunner & Neumann, 1984; Kleijnen, 1987; Noether, 1981). Tests of equivalent distributions capture location shifts as well as other distributional differences.

Over thirty years have passed since Sawilowsky (1990) brought nonparametric tests of interaction to the attention of behavioral and social researchers. Thus, the purpose of this literature review is to update the findings from studies of the nonparametric tests of interaction that Sawilowsky presented in his review as well as to discuss nonparametric tests of interaction proposed subsequently. Discussion of the power and Type I error rates of tests includes robustness properties in the presence of deviations from parametric data assumptions such as normality and homoskedasticity. Discussion of test power and Type I error rate in the presence of nonnormal data distributions includes research into robustness of tests in the presence of data having distributions with asymmetry and skew. Research into robustness of tests includes performance in the presence of unequal group sample sizes. Discussion includes research into relative efficiencies and asymptotic relative efficiencies where available. The relative efficiency of two tests is a ratio of the sample size necessary for two tests to achieve the same power level at an equal nominal alpha value. The asymptotic relative efficiency, or Pitman efficiency (Pitman, 1948), is a ratio that compares the relative efficiency of two tests when the sample size is infinitely large and the treatment effect is infinitely small. While real world conditions do not meet the assumptions of infinitely large sample sizes, asymptotic relative efficiency can be a useful metric of the relative power nonparametric tests have to parametric tests in Monte Carlo studies (Blair, 1981; Blair & Higgins, 1981; Hollander & Wolfe, 1973; Randles & Wolfe, 1979; Smitley, 1981). This research also highlights areas where these nonparametric tests still need to be assessed for their use in the behavioral and social sciences.

Nonparametric Methods for Addressing Interaction Reviewed by Sawilowsky

The methods to be discussed in this section were previously discussed in Sawilowsky's review of nonparametric tests for interaction. These methods include: (1) the rank transform test, (2) the adjusted rank transform test, (3) the Puri and Sen *L* test, (4) the McSweeney test, (5) the collapsed-reduced test, (6) the approximate randomization test, (7) the moment approximation test, (8) the random normal and expected normal scores test, (9) the Hettmansperger test, (10) the extended median test, and (11) the small-samples Patel and Hoel test. Sawilowsky dedicates a section of his review for each of these methods barring the McSweeney test which is only referenced with passing mentions while discussing other methods. However, discussion of the McSweeney test falls in logical sequence after discussions of the rank transform test, the adjusted rank transform test, and the Puri and Sen *L* test. This review will include more recent studies of relative efficiency, asymptotic behavior, and other mathematical properties, as well as more recent comparative research into the potential of these methods as alternatives to traditional parametric methods.

Rank Transform Test

Initial exploration into outlining and exploring the viability of the rank transform (RT) method is published in a line of inquiry by Conover and Iman (Conover & Iman, 1976, 1981; Iman, 1974; Iman and Conover, 1976). It should be immediately noted that data can be ranked both within groups and between groups, and that ranking as a data transformation can be used independently or in conjunction with other forms of data transformation (Conover & Iman, 1981). Subsequent to the series of data transformations, the data can then be referred to any inferential statistic. Most commonly, what is referred to as the rank transform test is the method of ranking the entire set of observations from smallest to largest altogether and then performing the traditional parametric analysis of variance using the ranks instead of the scores.

When converting scores to ranks for the purpose of performing a parametric analysis in this way, the two or more sets of scores are combined prior to ranking. While transforming one set of scores to ranks will produce a distribution that is rectangular, when transforming two or more sets of pooled scores to ranks the distributions of the individual set ranks will no longer be necessarily rectangular. Differences in group measures of variance, skew, kurtosis, and bimodality are retained after methods of rank transformation albeit with reduced magnitude (Zimmerman, 2011). The magnitude of the reduction in these measures is sufficient to influence Type I and Type II error rates of the Wilcoxon-Mann-Whitney test. Despite having similar measures of variance, skew, and kurtosis, individual groups do not necessarily retain the same shape after score to rank transformations.

Evidence is provided to suggest both viability for the rank transform method as well as sufficient concern to warrant further study for most research designs. A rank transform Hotelling's T^2 statistic demonstrates robustness to loss of power in repeated measures randomized complete block designs given low sample correlations among the measures (Agresti & Pendergast, 1986; Kepner & Robinson, 1988). Further study demonstrates that the power of the rank transform test can vary substantially when testing for interaction in the presence of main effects (Akritas, 1990). The rank transform analysis of covariance (RT ANCOVA) demonstrates potential with favorable results compared to the parametric ANCOVA (Olejnik & Algina, 1984), but performs with less power compared to some proposed alternative rank transform nonparametric statistics (Harwell & Serlin, 1988; Stephenson & Jacobson, 1988). Good performance is also demonstrated for rank transform multivariate tests of independence for two sets of variables (Habib & Harwell, 1989).

Of particular interest is the viability of the rank transform ANOVA as an alternative to the factorial, parametric ANOVA. Problems with the rank transform test include rejecting null main effects after testing for interaction effects in a 4 x 3 design (Lemmer, 1980), inflated Type I error in the 4 x 3 design in the presence of both nonnull main effects (Blair et al., 1987), and inflated Type I error in the 2 x 2 x 2 design (Sawilowsky et al., 1989). The rank transform ANOVA is demonstrated to be robust in the 2 x 2 layout but lacks power over the parametric ANOVA for larger effect sizes (Sawilowsky, 1989). Given these results, the rank transform ANOVA may not be a reasonable alternative to parametric ANOVA.

Thompson (1991) studied the asymptotic properties of the rank transform ANOVA statistic for testing interactions in the balanced two-way design and proved that in a two-way design with exactly two levels of both main effects (or when there is only one main effect) the test statistic for interaction is asymptotically a chi-square divided by the degrees of freedom. In a two-way design with two levels for both main effects, this means that the interaction statistic has a distribution that is asymptotically a chi-square with one degree of freedom. Thompson also proved that for some effect sizes in two-way designs with more than two levels of the main effects, it is possible for the interaction test statistic to converge to infinity. This would, in turn, cause Type I error rates to converge to 1. This result is consistent with previous studies demonstrating inflated Type I error rates for the rank transform ANOVA. Thompson surmises that for designs larger than the 2 x 2, the use of the rank transform may lead to confusions about effects that are instead, at least in part, due to the nonlinearity of the mapping from data to ranks.

Simulation studies exploring the viability of the rank transform ANOVA have emerged with consistent frequency. The rank transform ANOVA has inflated Type I error and reduced power of the rank transform ANOVA compared to the parametric ANOVA in the 2 x 2 x 2 factorial design for populations which are normal, uniform, or exponential (Sawilowsky, 2000). The rank transform ANOVA demonstrates competitive viability in testing the presence of effects in one-way ANOVA designs and inflated Type I error and reduced power in more complex designs (McKean and Vidmar, 1994). However, Blair, Sawilowsky, and Higgins (1987) found inflated Type I error rates when testing for interaction in the two-way layout in the presence of nonnull main effects. Sawilowsky, Blair, and Higgins (1989) found inflated Type I error rates in the presence of nonnull main effects in the 2 x 2 x 2 layout. Sawilowsky (1985) found Type I error rate and power are dependent on how treatment effects are modelled in the 2 x 2 x 2 layout. Lemmer (1980) found the rank transform ANOVA to demonstrate inflated Type I error for main effects when in the presence of nonnull interaction effects. Remarkably, there seems to be little discussion of cell sample sizes in these studies beyond highlighting that they are based on small samples. McKean and Vidmar (1994) used n = 5 cell sample sizes, Toothaker and Newman (1994) used n = 5 and n = 10 cell sample sizes, and Sawilowsky (2000) used n = 2 and n = 20 cell sample sizes. Focus of the viability of the rank transform ANOVA in this line of inquiry has been situated primarily on population distribution and research design and less on ensuring results remain consistent across varying ranges of sample sizes. Varying this additional facet of research design could potentially warrant further discussions of power and Type I error comparisons.

Another point of view on the shortcomings of the rank transform ANOVA for the two-way design is provided by Lemmer (1980, 2001). Lemmer noted that the presence of a significant interaction effect can cause significant main effects to be detected even if they are not present. Lemmer proposed not using the rank transform ANOVA to test for main effects until no interaction has been detected or the interaction effect can be demonstrated as small enough to not influence testing for main effects. The goal of this proposed solution is to keep the rank transform ANOVA viable in the two-way design. Another attempt to tackle this angle of keeping the rank transform ANOVA viable in the two-way design is the work of Marden and Muyot (1995) where, in the two-way ANOVA design, alterations are made to the conventional definition of interaction. The traditional null hypothesis of no interaction effect would imply the effect of main effect A is the same for all levels of B and vice versa (p. 1392). Marden and Muyot (1995) built on the work and ideas of de Kroon and van der Laan (1981), Akritas and Arnold (1994), and Patel and Hoel (1973) to redefine interaction into a null hypothesis of main effect A being concordant within all levels of main effect B and vice versa (p. 1392). That is, after the pooled rank transformation is performed, vectors of the ranked observations taken from levels of main effect A will have the same distributions as vectors of the ranked observations from levels of main effect B. Test statistics with asymptotic chi-square distributions are derived.

Application of the rank transform procedure has been explored in designs more complex than the two-way ANOVA as well. Repeated-measures ANOVA on rank-transformed data demonstrates competitive power and Type I error rates to the parametric repeated measures ANOVA (Zimmerman & Zumbo, 1993). The rank transform analysis of covariance (RT ANCOVA) procedure demonstrates robustness to most violations of the usual parametric assumptions while maintaining competitive power (Harwell, 2003; Rheinheimer & Penfield, 2001). Brunner and Dette (1992) derived a test statistic using partial rank transformations for mixed models and found it to reduce to Friedman's statistic if no interactions are present and cell frequencies are equal. Akritas and Brunner (1997) derived rank-based statistics for hypotheses of fixed treatment effects and interactions in mixed models which they found to have liberal Type I error for repeated measures, nested repeated measures, partially nested, and cross-classified repeated measures designs.

Adjusted Rank Transform Test (Aligned Rank Transform Test)

What is referred to as the adjusted rank transform test is also commonly referred to as the aligned rank transform test. Reinach (1965, 1966) provided an early demonstration of the procedure. The motivation of the aligned rank procedure is to adjust the rank transform test so that it does not suffer reduced power and increased Type I error rates in the presence of nonnull main and interaction effects in a factorial ANOVA design. The procedure is performed by first aligning the data, then pooling and ranking the data, and finally referring the aligned and rank-transformed scores to the parametric ANOVA procedure. The goal of aligning is to treat whichever of the two main or interaction effects not currently being tested as nuisance parameters and remove them from the model. When testing for interaction, this means subtracting main effect estimates (means calculated within levels of each main effect) from the observations. This leaves only the interaction effect to be tested. One main effect and the interaction effect can be removed as well in balanced designs to test for the other main effect (Marascuilo & McSweeney, 1977). Simulation study of the adjusted rank transform test demonstrates strong power properties and slightly liberal Type I error rates (Blair & Sawilowsky, 1990).

The process of aligning and ranking the data as a test for interaction in the two-way layout, most often performed by subtracting main effect mean estimates from each observation, is a modification of the aligned rank process proposed by Hodges and Lehmann (1962). It is worthwhile to note that while main effect mean estimates are most often used to align the data prior to ranking, Hodges and Lehmann note that other methods of alignment are also possible (specific alternatives mentioned are trimmed or Winsorized means). Additionally, Hodges and Lehmann note that any test statistic can be used subsequent to the aligning and ranking process. Peterson (2002) explored testing for interaction in the 4 x 3 design with n = 5, 10, 15, and 20 cell sample sizes for various nonnormal distributions (specifically those highlighted by Micceri, 1989) while aligning on alternative effect estimates such as the median, Winsorized means with different degrees of trim, and other estimators to find aligning on the mean and median performed most favorably. While the process of aligning the data based on main effect mean estimates and ranking can be used in combination with different test statistics, further research is needed in aligning the data with different main effect estimators.

Early studies demonstrated promising results of the adjusted rank-transform test (ART test) but often resulted in finding that, when the procedure could not be done by hand, computing power was too laborious for the technology of the time for complex research designs (Conover & Iman, 1976). Consequently, there are increasing simulation studies into the viability of the ART test in recent decades. The ART test has strong power while being robust for populations from normal, uniform, logistic, exponential, double exponential, and lognormal distributions (Blair & Sawilowsky, 1990; Fawcett & Salter, 1984; Groggel,1987; Mansouri & Chang, 1995; Salter & Fawcett, 1985, 1993). The ART test performs with larger power than the parametric test with non-normal distributions likely to be observed in psychological research such as the truncated normal, student, and gamma distributions (Leys & Schumann, 2010). The ART test also holds power advantages over other common nonparametric methods when testing for interaction until both main effects become large (Headrick & Vineyard, 2000; Leys & Schumann, 2010).

Viability of the ART test to detect interaction effects in more complex research designs has been studied as well. The ART test demonstrates strong power when testing for interaction in the two-way design with one observation per cell (Hartlaub et al., 1999). Beasley and Zumbo (2009) found aligned-rank based tests of location shift sensitive and robust to nonnormality when testing for interactions in multiple group repeated measures designs with unequal cell sample sizes. Davis and McKean (1993) extend the ART test for use in multivariate linear models. Shiraishi (1991) derived and outlined aligned rank-based statistics for tests of main effects and interactions in balanced two-way multivariate analysis of variance designs (MANOVA), including a formula for calculating asymptotic relative efficiency which is dependent upon the number of response variables and their multivariate variance-covariance matrix.

Puri and Sen L Test

The Puri and Sen *L* statistic (1985) is a large sample approximation of the Puri and Sen aligned ranks technique (1969). Because the Puri and Sen aligned ranks technique (1969) refers to the exact form of the test statistic, it becomes computationally intensive for larger sample sizes. The large sample approximation of the Puri and Sen *L* statistic is provided to adjust for large samples with the *L* test presented in trace criterion form in Harwell (1990) and Harwell and Serlin (1989a). Calculation of the Puri and Sen *L* statistic first requires that all observations are pooled and rank-transformed. Harwell and Serlin (1989a) present calculation of the *L* statistic as follows:

$$L = (N - 1) \sum_{s=1}^{s} r_s^2 \tag{1}$$

where the r_s^2 [s = 1, 2, … , $S = \min(p, q)$] represent squared canonical correlations. Assuming a multivariate general linear model, *p* is the length of the vector of dependent variables and *q* is the length of the vector of predictor variables. *L* is asymptotically distributed as chi-square with *pq* degrees of freedom and tested as such. An alternative and equivalent calculation for the *L* statistic can be calculated using output from the ANOVA table on the ranked observations using:

$$L = (N - 1) \text{ x } (SS_{\text{Effect}} \ / \ SS_{\text{Total}}) \tag{2}$$

where the calculated *L* statistic is tested using a chi-square distribution with degrees of freedom equal to that of the effect being tested. Harwell (1990) noted the purpose of the model is testing one or more population regression coefficients against zero.

The general trend for the performance of the Puri and Sen *L* statistic is that it performs exceptionally conservatively (i.e., lower than expected Type I error rates) until sample sizes become large (Harwell, 1991; Harwell & Serlin, 1989; Toothaker & Newman, 1994) . For designs with population distributions having varying combinations of symmetry and tail thickness, it can take the *L* statistic sample sizes as large as n = 400 for acceptable Type I error rates to converge and sample sizes as large as n = 100 to demonstrate power advantages over the parametric ANOVA or rank transform test (Harwell & Serlin, 1989a). Harwell (1991) found the *L* statistic to perform more conservatively than the approaches of Hettmansperger (1984) and McSweeney (1967) for normal, double exponential, Cauchy, and chi-square distributions with sample sizes n = 3, 5, 8, and 12. The *L* statistic demonstrates similar ultra-conservatism compared to the rank transform test, Hettmansperger test, and McSweeney test for normal, exponential, and mixed normal distributions with 2 x 2, 2 x 4, and 4x4 designs having cell sizes n = 5 and n = 10 (Toothaker & Newman, 1994). Similar poor performance of the *L* statistic over alternative nonparametric methods is demonstrated in analysis of covariance designs by Olejnik and Algina (1984), Headrick and Vineyard (2000), and Rheinheimer and Penfield (2001).

McSweeney Test

McSweeney's test is similar to the adjusted rank transform test and the Puri and Sen *L* test (McSweeney, 1967). Construction of the McSweeney test requires the data to be aligned and rank transformed. However, rather than using ANOVA with the aligned and rank transformed data, the same statistic calculated in the Puri and Sen *L* test is used. As with the Puri and Sen *L* test, the ANOVA table can be used to calculate the test statistic with the aligned and rank transformed data. To calculate a *p* value, the test statistic is referred to a chi-square distribution with degrees of freedom that would be used for the effect of interest in the parametric analysis of variance. In summary, the difference between the McSweeney test and the Puri and Sen *L* test is the McSweeney test employs the alignment procedure prior to ranking.

Simulation studies of the viability of this test yield mixed results. For normal, mixed-normal, exponential, double exponential, and chi-square population distributions in combination with small cell sample sizes, the McSweeney test demonstrates strong power at the cost of liberal Type I error (Harwell, 1991; Toothaker & Newman, 1994). Kelley et al. (1994) explored viability of the McSweeney test in the 2 x 2 x 2 factorial design for normal, uniform, *t*, and exponential population distributions with sample sizes of n = 7, 21, and 35. Results demonstrate a failure to maintain Type I error rates for null effects in treatment conditions but competitive results when the number of nonnull effects and effect sizes became large. The McSweeney test demonstrates power competitive to the parametric analysis of covariance for designs where parametric assumptions are violated as well (Olejnik & Algina, 1985).

Bradley's Collapsed-Reduced Test

Bradley's (1979) nonparametric technique of testing for interaction is limited to balanced designs and assumes no tied observations. When testing for interactions, the data is entered into a matrix where it is collapsed and reduced until it is subjected to the appropriate nonparametric test for the effect. Researchers can choose from various approaches to entering the data into a matrix and the results of the test will change depending on this choice (Bradley, 1979; Sawilowsky, 1990). The test suffers from high variability even with the same data and *p* values can be altered by researchers who enter the data into the matrix in different ways. The technique is also found to have low power when compared to other readily available nonparametric methods (Kelley et al., 1994). For these reasons, it has not been heavily researched and will not be discussed further in this review.

Approximate Randomization Test

Randomization tests and approximate randomization tests have seen significant increases in research over recent decades. This increase is directly related to the increase in computing power readily available for computing permutations of samples. The randomization test was first introduced by Fisher (1935) and yields efficient results which, at the time, were impractical to perform except with small samples. This is because randomization tests require all possible permutations of the data for each new sample taken (Bradley, 1968). Randomization tests are performed by calculating the test statistic of interest on all possible permutations of the observed scores among conditions and assessing the probability of obtaining a test statistic as large, or larger than, the one observed. Early attempts to mitigate the computation time necessary were to perform randomization tests on ranks (Bradley, 1968; Siegel, 1956). That is because the distribution created by permuting ranks depends only on the sample size so that computation is less intensive. Edgington (1969, 1980) outlined an approximate randomization test where a random sample of possible data permutations is taken to mitigate the computation intensiveness required.

Still and White (1981) were the first to study the use of an approximate randomization test applied to a test for interaction. They demonstrate their approach in the 2 x 2 design and in the 2 x 2 design with repeated measures on one factor. Instead of permuting the scores themselves, a randomization test is performed on aligned scores given that their focus was on testing for interaction. Given *a* levels of the first main effect, *b* levels of the second main effect, and *n* observations per cell, (*abn*)! permutations are performed generating (*abn*)!/(*a*!*b*!*n*!) distinct *F* values. Rejection of the null hypothesis of no effect occurs if the proportion of *F* values equal to or greater than the *F* value for the original observations is less than or equal to *α*. In exploring the efficiency of this technique on normal and nonnormal simulated data in the 2 x 2 design with n = 5 cell sizes, Still and White (1981) conclude that when sampling distributions are unknown, the approximate randomization test is preferable as it is unlikely to ever be less efficient than the parametric *F* test. They note the primary limiting factor is the computational intensiveness needed. Bradbury (1987) found that the approximate randomization test performs better than the parametric *F* for smaller interaction effect sizes, though the *F* statistic closes the power gap with larger effect sizes. Umlauft et al (2017) performed research into rank-based permutation approaches for factorial designs using various Wald-type and Kruskal-Wallis statistics with results demonstrating a permutation-based Wald-type statistic is finitely and asymptotically exact, controls Type I error rate, and yields larger power than non-permutation-based Wald-type statistics based on ranks. Ditzhaus et al. (2021) found similar exactness, robustness, and power results in their research using Wald-type statistics with quantile-based permutation methods (QANOVA) for factorial designs.

There is study into the application of randomization tests for various research designs in a variety of disciplines. The asymptotic behavior of randomization tests of different forms has been compared to resampling techniques such as the bootstrap (Romano, 1989), under approximate symmetry assumptions (Canay et al., 2017), and has been used to construct corresponding confidence intervals (Garthwaite, 1996). Mielke and Berry (1994) and Manly (1995) performed simulation studies comparing the power of randomization tests for two-sample tests of location when the equal variance assumption cannot be assumed, yielding results of inflated Type I error rates that remain competitive to parametric alternatives. Christensen and Zabriskie (2022) studied randomization test robustness for two-sample tests of location when differing parametric assumptions are violated and when the data is subjected to various transformations and found the power of two-tailed permutation tests can equal zero in some designs with samples which are skewed and have unequal sample sizes. The application of randomization tests for making inference is used in a variety of fields, including behavioral studies (Craig & Fisher, 2019; Elliffe & Elliffe, 2019; Peres-Neto & Olden, 2001), biomedical research (Goulden et al., 2010; Helwig, 2019; Ludbrook & Dudley, 1998), econometrics (Kennedy, 1995), politics (Erikson et al., 2010), and ecology (Pillar, 2013; Potvin & Roff, 1993). Reviews of software packages capable of performing these techniques as computing power becomes more readily available to researchers are available as well (Arnholt, 2007; Chen & Dunlap, 1993; Hayes, 1998; LaFleur & Greevy, 2009).

Moment Approximation Test

The Berry and Mielke's (1983) moment approximation test was derived to overcome the intense computation requirement of randomization tests. The test procedure is to calculate a test statistic from an observed data set, derives the moments of the sampling distribution of the test statistic, and calculate a *p* value from the constructed distribution. A simulation by Berry and Mielke found the test to be robust and powerful in a one-way design with three levels and a total sample size of N = 13. However, no further research has been conducted to explore this technique, so no research exists on using it for the test of an interaction. This is likely because the ever-increasing computing power readily available to researchers negates the need to compromise by altering the randomization test approach.

Normal Scores Transformation Methods

Normal scores tests are a combination of a normal scores data transformation and most commonly either a normal or chi-square test statistic compared to their respective distributions. Three methods for performing a normal scores data transformation of the response variable are outlined here. The commonly used chi-square test statistic for the one-way design is also outlined in this section. While performing a normal scores data transformation does not limit a researcher to using a normal or chi-square test statistic, minimal research has explored using a normal scores data transformation in combination with other test statistics for the use of making inference. In summary, what is commonly referred to as a normal scores test is one of three normal scores data transformations in combination with a chi-square test statistic.

There are three methods for making inferences about group differences using normal score transformations: the Bell-Doksum normal scores test (Bell & Doksum, 1965), the Terry-Hoeffding normal scores test (Hoeffding, 1951; Terry, 1952), and the Van der Waerden normal scores test (Van der Waerden, 1952, 1953, 1956). Fisher and Yates (1949) and Bell and Doksum (1965) proposed using a random normal scores transformation, a procedure where ranks of original scores are replaced by randomly drawn normal deviates with corresponding ranks. Bradley (1968) refined this technique by limiting the variates drawn to those of a standard normal distribution, creating a more powerful form of the test statistic. As the deviates drawn in the Bell-Doksum test are random, two researchers analyzing the same data with this test may arrive at different conclusions.

Hoeffding (1951) and Terry (1952) refined this procedure further to replace ranks of the original observations with expected normal scores. As these expected normal scores are constant values dependent only upon sample size, researchers have calculated tables of expected normal order statistics (Harter, 1961) and tables of appropriate chi-square critical values for total sample sizes less than or equal to 20 (Klotz, 1964). Owen (1962) tabled large sample approximation normal theory critical values. Sawilowsky (1990) refers to the data transformation procedure outlined in the Bell-Doksum test as a random normal scores transformation (RNST) and the procedure outlined in the Terry-Hoeffding test as the expected normal scores transformation (ENST).

Expected normal scores transformations are adaptable to any hypothesis in experimental research designs (Conover, 1980; Gibbons, 1985). Lu and Smith (1979) found the ENST robust and powerful compared to the parametric ANOVA *F* in the one-way design. Sawilowsky (1985, 1989) found both the RNST and ENST to be more powerful and robust than the rank transform test but less powerful and robust than the parametric *F* test in the 2 x 2 x 2 design with small sample sizes (n = 2) for various distributions. Feir-Walsh and Toothaker (1974) concluded that the Terry-Hoeffding normal scores test yielded less power than the ANOVA *F* test and the Kruskal-Wallis test for total sample sizes as large as 200. Wiedermann and Alexandrowicz (2011) demonstrated a modified Terry-Hoeffding test is a robust and powerful normal scores test for two-sample paired data when compared to the *t* test.

Van der Waerden (1952, 1953, 1956) proposed an alternative normal scores transformation. This test uses a rank transformation in combination with the inverse standard normal distribution function. The Van der Waerden normal scores transformation is as follows:

$$z_{ij} = \Phi^{-1}\left(\frac{R(X_{ij})}{N+1}\right) \tag{3}$$

where X_{ij} represents the i^{th} value in the j^{th} group (j = 1, 2, …, k), $R(X_{ij})$ represents the pooled rank of observation X_{ij}, and Φ^{-1} denotes the inverse standard normal distribution function. The chi-square test statistic is calculated as:

$$W = \frac{1}{s^2}\sum_{j=1}^{k} n_j\,\bar{z}_j^2 \tag{4}$$

where n_j represents the sample size for the j^{th} group,

$$s^2 = \frac{1}{N-1}\sum_{j=1}^{k}\sum_{i=1}^{n_j} z_{ij}^2 \tag{5}$$

and

$$\bar{z}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} z_{ij} \tag{6}$$

The null hypothesis that *k* groups yield the same observations is rejected when:

$$W > \chi^2_{\alpha,k-1} \tag{7}$$

McSweeney (1967) studied the fit of the asymptotic chi-square distribution of the normal scores test statistic in the presence of small sample sizes for the one-way c-sample design, concluding that both the Terry-Hoeffding and Van der Waerden normal scores test statistics with small sample sizes are well-approximated by the chi-square distribution with k – 1 degrees of freedom. Transforming original scores to normal scores for conducting a one-way two- or more-sample test of differences in central tendency can result in equivalent or larger power than traditional parametric *t* and *F* tests, the Wilcoxon rank sum test (also called the Mann-Whitney *U* test), and the Kruskal-Wallis test while still controlling the Type I error rate at or near the nominal level (Hodges & Lehmann, 1961; Keselman & Toothaker, 1973; Kruskal & Wallis, 1952; Mann & Whitney, 1947; McSweeney, 1967; Penfield, 1994; Penfield & McSweeney, 1968; Thompson et al., 1966; Van der Laan, 1964; Van der Laan & Oosterhoff, 1965, 1967; Wilcoxon, 1945, 1947). This power increase is shown to be larger for large- or heavy-tailed distributions (Curtis & Marascuilo, 1992; Lu & Smith, 1979; Padmanabhan, 1977). Zimmerman (1996) demonstrated via simulation that rank transformations, including the Van der Waerden normal scores transformation, can be used to reduce group variance heterogeneity in two-sample tests of location difference.

Hettmansperger Test

Hettmansperger (1984) outlined an aligned-rank based test which can be used to test for significant interactions. First, the data are aligned by subtracting mean main effect estimates and then ranked akin to the aligned rank test. At this point, the ranks are standardized. From the aligned, ranked, and standardized observations, the Hettmansperger test is to calculate the interaction sums of squares. Hettmansperger offers two critical values to test against. The first is a chi-square critical value with degrees of freedom equal to that of the interaction effect being tested. The second is an *F* critical value with degrees of freedom 1 equal to that which would be used in the parametric analysis and degrees of freedom 2 equal to the within degrees of freedom. The latter is suggested for samples of smaller size due to more liberal Type I error rates of the chi-square critical value.

Toothaker and Newman (1994) found the Hettmansperger test to maintain power comparable to other nonparametric methods but demonstrate liberal Type I error rates with small sample sizes. The study used data from normal, exponential, and mixed normal population distributions in 2 x 2, 2 x 4, and 4 x 4 designs with small cell sample sizes of n = 5 and n = 10. The test has similar Type I error rates as other nonparametric methods, such as the Puri and Sen *L* and McSweeney test, but has more power than these methods for normal, double exponential, Cauchy, and chi-square distributions. It also has larger power than the parametric ANOVA *F* test for all but the normal distribution (Harwell, 1991). Headrick and Vineyard (2000) compared the test to the Puri and Sen *L* and the adjusted rank transform test across a variety of treatment effect sizes, conditional distributions, and cell sample sizes of n = 10 and n = 20 and recommended the Hettmansperger test as an alternative to the parametric ANOVA provided cell sample sizes are at least 10. The Hettmansperger test also demonstrates potential in analysis of covariance research designs with performance stronger than several other nonparametric techniques (Rheinheimer & Penfield, 2001).

Extended Median Test

Shoemaker's extended median test (1986) is based on the proportions of the count of observations in each cell of one main effect that are above and below the median of observations based on the level of the other main effect within that cell. The test statistic is distributed as a Pearson chi-square statistic with degrees of freedom equal to the interaction effect degrees of freedom using parametric ANOVA. Mohebbi and Shoemaker (1990) extend upon this proposed median test in the analysis of variance layout providing evidence it performs with more power than the parametric ANOVA for highly skewed distributions assuming cell sizes of at least 10. Freidlin and Gastwirth (2000) argue that the median test should be retired from general use given its low power compared to the parametric ANOVA for normally distributed data and less power than the Wilcoxon rank sum test or modified Wilcoxon rank sum test (Fligner & Policello, 1981) for nonnormal designs.

Small-Samples Patel and Hoel Test

The Patel and Hoel technique (1973) is difficult to compute in applied situations. The technique is a nonparametric test for interaction that demonstrated promise at the time of publication for large sample sizes and heavy tailed distributions. The Patel and Hoel technique is a test of interaction in the 2 x 2 design where a null hypothesis of H_0: μ = 0 is tested where:

$$\mu = P(X_{12} \leq X_{11}) - P(X_{22} \leq X_{21}) \tag{8}$$

Krauth (1988) refined a small-sample modification of the Patel and Hoel technique that is not distribution free, assumes no tied observations, and is easier to compute. The test requires critical values from a hypergeometric distribution. There are no further studies of this test, perhaps due to the complexity of this method.

Nonparametric Tests of Interaction Published After 1990

The methods to be discussed in this section are novel nonparametric tests of interaction published subsequent to Sawilowsky's review of nonparametric tests for interaction. These methods include: (1) Gao and Alvo's Row-Column Test and (2) De Neve and Thas's Interaction Probability of Superiority. This review will include calculations of relative efficiency, asymptotic behavior, and other mathematical properties, as well as comparative research into the potential of these methods as alternatives to traditional parametric methods.

Gao and Alvo Row-Column Test

Gao and Alvo (2005a) construct a novel composite linear rank-based nonparametric statistic to test for interaction effects in the two-factor design. In calculating the test statistic, each observation is first given two ranks: one respective to its value relative to all observations in the same level of factor A and the other respective to its value relative to all observations in the same level of factor B. Two vectors are created containing the sums of factor level rank scores within each cell. The proposed test statistic is a composite of the variance-covariance matrices calculated from these two vectors. Under the null hypothesis of no interaction, the proposed test statistic follows an asymptotic chi-square distribution with $(I-1)(J-1)$ degrees of freedom where I is the number of levels of main effect A and J is the number of levels of main effect B. Gao and Alvo present simulations to compare results of the proposed row-column test to the ANOVA F test and aligned rank transform test in the 4 x 3 design using n = 10 and n = 20 cell sample sizes for normal, contaminated normal, and Cauchy distributions. Results demonstrate competitive power of the row-column test as the cost of liberal Type I error rates ranging from 0.058 to 0.099.

Gao and Alvo note the test can be extended to unbalanced designs with different cell sample sizes using weighted row and column ranks. Gao and Alvo (2005b) subsequently expand upon using the composite linear weighted rank-based nonparametric statistic to test for main, interaction, and nested effects. Asymptotic relative efficiency (ARE) of the proposed test relative to the ANOVA *F* test is calculated revealing values of 0.908, 1.046, and 1.092 for normal, logistic, and double-exponential distributions, respectively, when testing for interaction effects. De Neve and Thas (2017) conclude the Gao and Alvo row-column test does not control Type I error rates for small (n = 5) cell sample sizes in balanced designs and in unbalanced designs for normally distributed data. Type I error rates are concluded as acceptable for n = 10, but the test considerably lacks power in the presence of heteroskedasticity in balanced designs and when there are significant main effects in unbalanced designs.

De Neve and Thas Interaction Probability of Superiority

De Neve and Thas (2017) propose a rank-based test that builds upon the score-type test of location shift for interaction proposed by Bhapkar and Gore (1974). De Neve and Thas highlight a traditional measure of an interaction effect:

$$\alpha = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) \tag{9}$$

From this, two transformed outcomes are defined: $Z = Y_{11} - Y_{21}$ and $Z^* = Y_{12} - Y_{22}$. The statistic proposed to quantify the interaction effect, named the interaction probability of superiority (IPS), is defined as:

$$\beta = \mathrm{P}(Z \leq \mathrm{Z}^*) \tag{10}$$

Under the null hypothesis of no interaction ($\alpha = 0$), it is noted that $\beta = 0.5$. Using an estimate of β calculated from the observed data, an asymptotic chi-square test statistic is constructed having $(K - 1)(L - 1)$ degrees of freedom where K is the number of levels of main effect A and L is the number of levels of main effect B. Asymptotic relative efficiencies of the test relative to the ANOVA *F* are calculated for the following distributions: 0.91 for uniform, 0.99 for normal, 1.07 for logistic, 1.24 for Laplace, 1.59 for exponential, and 3.81 for log-normal.

For balanced designs with n = 5 and n = 10 cell sample sizes, De Neve and Thas (2017) show the IPS test to control Type I error rates, perform with power competitive to the ANOVA *F* test and other nonparametric tests, and be robust in the presence of heteroskedasticity and varying error distributions. The test yields similar results in unbalanced designs with the exception of inflated Type I error rates in the presence of outliers.

Summary

For the thirteen methods discussed to test for the presence of an interaction effect, design factors are highlighted for which each method has been studied to evaluate Type I error and power properties for the factorial design in Table 1. The design factors included in Table 1 are nonnormal distributions, heteroskedasticity, unequal cell sample sizes, and designs more complex than the factorial design. Power and Type I error results are summarized for each method in the presence of varying design factors.

Table 2.1

Design factors studied for each nonparametric test of an interaction effect

Tests:	Nonnormal Distributions	Heteroskedasticity	Unequal Cell Sample Sizes	More Complex Designs
Rank Transform	Studied for use with normal, uniform, exponential, and mixed-normal distributions yielding varying results	Differences in group measures of variance, skew, kurtosis, and bimodality are retained after rank transformation albeit with reduced magnitude		Extended for use with ANCOVA, MANOVA, and repeated-measures ANOVA
Aligned (Adjusted) Rank Transform	Competitive or larger power than ANOVA for normal, truncated normal, exponential, student's *t*, uniform, logistic, exponential, double exponential, lognormal, and gamma distributions		Robust to nonnormality with unequal cell sample sizes	Extended for use in MANOVA and repeated measures designs

Table 2.1

Design factors studied for each nonparametric test of an interaction effect

Tests:	Nonnormal Distributions	Heteroskedasticity	Unequal Cell Sample Sizes	More Complex Designs
Puri and Sen *L*	Performs conservatively for power and Type I error compared to ANOVA, rank transform, Hettmansperger, and McSweeney until samples become large for normal, mixed-normal, chi-square, exponential double exponential, and Cauchy			Similar conservatism is demonstrated in analysis of covariance design
McSweeney	Strong power at the cost of liberal Type I error rates for small sample sizes with normal, mixed-normal, uniform, student's *t*, chi-square, exponential, and double exponential			Power competitive to parametric ANCOVA
Bradley's Collapsed-Reduced	Lacks power compared to parametric and nonparametric alternatives			Limited to balanced designs

Table 2.1

Design factors studied for each nonparametric test of an interaction effect

Tests:	Nonnormal Distributions	Heteroskedasticity	Unequal Cell Sample Sizes	More Complex Designs
Approximate Randomization	Strong power but computationally intensive, competitive or larger power compared to ANOVA	Type I error rates inflate in the presence of unequal sample variances		
Moment Approximation				
Normal Scores Transformation	Can result in power increases compared to *t* and *F* tests, including large or heavy-tailed distributions	Normal scores transformations can be used to reduce group variance heterogeneity		
Hettmansperger	Liberal Type I error rates and power competitive to ANOVA *F*, Puri and Sen *L*, and McSweeney for normal, mixed-normal, chi-square, exponential double exponential, and Cauchy			Demonstrat-es potential in analysis of covariance research designs
Extended Median	Less power than ANOVA *F* test for normal distribution			
Small-Samples Patel and Hoel				

Table 2.1

Design factors studied for each nonparametric test of an interaction effect

Tests:	Nonnormal Distributions	Heteroskedasticity	Unequal Cell Sample Sizes	More Complex Designs
Gao and Alvo Row-Column	AREs of 0.908 for normal, 1.046 for logistic, and 1.092 for double exponential compared to ANOVA when testing for an interaction effect	Considerably lacks power in the presence of heteroskedasticity	Considerably lacks power with significant main effects in unbalanced designs	
De Neve and Thas Interaction Probability of Superiority	Controls Type I error rates and has power competitive to ANOVA *F*	Robust to heteroskedasticity	Robust to Type I error for unbalanced designs except in the presence of outliers	

Regarding testing for interactions in behavioral and social science research, recommendations for which test to use to detect interaction effects has changed over time. When one can safely assume the common parametric assumptions are satisfied, the parametric ANOVA is the most powerful option. When these parametric assumptions are not satisfied, several alternative nonparametric methods are available. While the rank transform test demonstrates strong power, the adjusted rank transform demonstrates equivalent or larger power and better robustness in Type I error rates for most designs. The McSweeney test demonstrates strong power as well but with slightly liberal Type I error rates. Normal scores data transformations demonstrate potential in competitive power for the normal distribution and mitigating the effects of heteroskedasticity between groups. Using normal scores data transformations with other nonparametric statistics has seen minimal research. A significant, recent change is the increasing emergence of randomization tests. These demonstrate power equal to or larger than the parametric *F* test and are distribution free. The computational power commonly available to most researchers proves this technique a competitive option for designs with small sample sizes.

A researcher needing a nonparametric test of interactions has competitive options from which to choose. Given that the performance of each option often depends on the design, the effect of interest, and the sample size, the above review can be used to select a method based on the particular study characteristics. The recommendation of a nonparametric test over a parametric one requires the nonparametric test to be robust to Type I error rate, maintain competitive power properties when normality can safely be assumed, and yield an increase in power for multiple types of nonnormally distributed data. No nonparametric test of interaction discussed in this review has been researched thoroughly enough to conclude it is always a preferable alternative to the common parametric ANOVA *F* test, that is, the above conditions can always be safely concluded as having been met. However, several of the discussed tests of interaction demonstrate potential in having met one or more of these conditions. Further study is needed to determine if the discussed methods that demonstrate this potential meet the conditions necessary to be considered preferable to the ANOVA *F* test or if any of the other methods could potentially be refined to address demonstrated properties of conservatism or liberal Type I error rates.

Discussion

The contribution of comprehensive theoretical and simulation research into these proposed nonparametric tests lacks momentum. While there is promising simulation research into how these tests compare for populations of various nonnormal distributions, it is noteworthy that little of this research contributes to the study of using these nonparametric tests for the other deviations from normality (asymmetry, multimodality, and skew) noted by Micceri (1989) as common in behavioral and social science research. Nonparametric tests have been proposed in the literature that may prove robust and competitive to parametric tests. The properties of these tests must still be thoroughly explored.

The so-called parametric method of analysis of variance is widely used in two-factor designs to detect and make inferences about the interaction of the two factors in their influence on a response variable. The method can be shown to be the most powerful when the underlying populations adhere to the conditions for valid inference. This is often not the case. Thus, the analysis of variance will often be less powerful or yield higher Type I error rates than alternative nonparametric methods. Researchers should carefully consider the nature of the populations from which they draw samples before selecting an inferential method for analyzing two-factor interactions.

CHAPTER 3

NONPARAMETRIC TESTS OF INTERACTION FOR THE TWO-WAY DESIGN WITH SKEWED DISTRIBUTIONS

Abstract

This study compares the ANOVA *F* test and four nonparametric competitors with two-factor designs for empirical Type I error rate and power when testing for an interaction of the factors. I perform simulations of 2 x 2 designs with cell sizes of 5, 10, 30, and 50. I use samples having distributions commonly found in the educational and social sciences including skew normal, skew logistic, and asymmetric Laplace distributions each having four levels of skew and asymmetry effects. The ANOVA *F* is robust for Type I error but lacks power compared to nonparametric alternatives for skew logistic and asymmetric Laplace samples. The aligned rank transform test demonstrates competitive Type I error and considerable power advantages over ANOVA *F* for the skew logistic and asymmetric Laplace distributions. The McSweeney test performs with larger power than ANOVA *F* in similar situations at the cost of more liberal Type I error.

Keywords: Nonparametric Statistics, Rank Transform, Aligned Rank Transform

Introduction

In 1990, *Review of Educational Research* published Shlomo Sawilowsky's article titled "Nonparametric Tests of Interaction in Experimental Design." A "nonparametric test" is a statistical test that is not conditioned on specified parameters of the underlying population distribution in order to make valid inferences (Hettmansperger et al., 2000; Marascuilo & Serlin, 1988; Nussbaum, 2015; Sawilowsky, 1990). The most common nonparametric tests can broadly be considered one of two types: (1) tests of a categorical response variable and (2) tests of a ranked response variable. A ranked variable can be viewed as an ordered categorical variable, so these two types could collectively be considered as the analysis of a categorical variable (Nussbaum, 2015; Stevens, 1946). Sawilowsky's review focuses on addressing the power, Type I error rate, and robustness of rank-based nonparametric methods.

Sawilowsky offers two primary reasons for reviewing nonparametric tests of interaction. (1) Nonnormally distributed variables are prominent in experimental research, particularly in the fields of the behavioral and social sciences (Blair & Higgins, 1981; Bloom, 1984; Bradley, 1968, 1977; Walberg et al., 1984), and many nonparametric methods do not rely on a condition of normally distributed data for valid inference (Nunnally, 1975, 1978). (2) Researchers often question the necessity and viability of nonparametric methods of analysis. This skepticism is prompted by beliefs that may be incorrect or only partially correct. These include the belief that: (a) common parametric methods of analysis are usually robust to violations of parametric assumptions regarding population distributions (Boneau, 1960; Box, 1954; Glass et al., 1972; Lindquist, 1953), (b) nonparametric methods are less powerful than parametric analogues (Kerlinger, 1973;

Nunnally, 1975), and (c) nonparametric methods are not available for addressing data obtained from complicated research and experimental designs (Bradley, 1968). Each of these concerns over the viability of nonparametric methods of analysis warrants further discussion.

There is evidence to suggest nonnormally distributed populations are common in the behavioral and social science fields. A high frequency of mixed-normal distributions in real world data is observed in education and psychological measures (Bradley, 1978, 1980a, 1980b, 1980c, 1982; Blair, 1980; Micceri, 1989; Still & White, 1981; Tan, 1982). Ceiling effects and floor effects in measurement ranges provide for the existence of censored-normal and truncated-normal distributions in education and social science measures (DeWees et al., 2020; Feng et al., 2019; Liu & Wang, 2021; McBee, 2010). These effects occur frequently with the use of Likert scales and other self-reported measures. Other frequent departures from normality found in psychometric measures include asymmetry, multimodality, and skew (Bradley, 1977; Micceri, 1989). Micceri's (1989) meta-analysis analyzed 440 large sample educational and psychological measures and found all to be significantly nonnormal via some class of contamination. Of the 440 measures analyzed by Micceri, only 15.2% had both tail weights at or about normal, 50.2% had at least one tail heavier than normal, and 14.8% had both tail weights less than normal. Of the remainder, 3.2% were concluded as uniform and 16.6% were concluded as Laplace. Additionally, of the 440 measures, 40.7% were concluded as moderately asymmetric, 19.5% as extremely asymmetric, and 11.4% as having exponential asymmetry. Of the categories for the measures analyzed by Micceri, nonnormal tail weights and asymmetry is demonstrated most prevalent in measures of criterion mastery, where 60% of measures were found to be distributed as Laplace, 37.1% of measures had extreme asymmetry, and 57.1% of measures had exponential asymmetry. Thus, Micceri suggests the existence of normally distributed data in educational and psychological measures is as improbable as the existence of a unicorn. Micceri's findings not only directly conclude the prevalence of the Laplace distribution in education and social science measures but also provide evidence of the prevalence of the logistic distribution by highlighting the frequency of mixed-normal measures and criterion mastery measures. The logistic distribution is often used to approximate mixed-normal distributions, model distributions of subjective quantitative measures, and is prevalent in item response theory analysis (Hellstrom, 1993; Noortgate et al., 2003; Savalei, 2006). This would imply that researchers in these fields who limit their methods of analysis to those dependent upon the assumption of normality will rarely be using methods as robust and powerful as available alternative nonparametric methods of analysis. Given the prevalence of measures distributed as Laplace, measures distributed as logistic, and measures having varying degrees of asymmetry, methods of analysis capable of handling these variables are crucial for researchers in education and social science fields. The presence of deviations from normality alone does not necessitate research into alternatives to traditional parametric tests without evidence that the performance of these parametric tests suffers in the presence of these deviations.

The concern over robustness of parametric tests to departures from parametric assumptions about data distributions has yielded extensive study. Claims and conclusions regarding the robustness of parametric tests are subject to a degree of variability as there is no commonly agreed upon quantitative definition or standard of robustness (Blair &

Higgins, 1981; Bradley, 1978). Consequently, claims of robustness are left to the conclusion of the author rather than the attainment of a standardized performance. Traditional parametric tests of location equality among two or more groups, such as the *t* test and analysis of variance (ANOVA) *F* test, build on population assumptions that include a normally distributed population, independence of observations, and homogeneity of group variances (Fisher, 1922, Hildebrand, 1986; Winer, 1971). Another common factor analyzed in discussion of the robustness question is equality, or lack thereof, in group sample sizes. The *t* test demonstrates reasonable to strong robustness to minimal departures from normality in some circumstances (Boneau, 1960; Sawilowsky & Blair, 1992), but fails to maintain reasonable robustness in others, such as when distributions are extremely skewed or when normality deviations occur in the presence of small group sample sizes (Blair, 1980; Blair & Higgins, 1980a, 1980b, 1981; Bradley, 1968; Hemelrijk, 1961). Some argue that the ANOVA is robust to various minor deviations from population normality (Baker et al., 1966; Boneau, 1960, 1962; Box, 1954; Cochran, 1947; Cochran & Cox, 1950; Fisher, 1922; Glass et al., 1972; Goodard & Lindquist, 1940; Guilford & Fruchter, 1978; Hack, 1958; Havlicek & Peterson, 1974; Lunney, 1970; Mandeville, 1972; Pearson, 1929; Rider, 1929; Scheffe,1959), while others express skepticism (Blair, 1980; Blair & Higgins, 1980a, 1980b, 1981; Bradley, 1968). Error rates for ANOVA inflate with violations to the homogeneity of variance assumption (Bishop, 1976; Box, 1954; Brown & Forsythe, 1974; Goodard & Lindquist, 1940; Hornsnell, 1953; Lindquist, 1953; Welch, 1937; Wilcox et al., 1986), increasingly so in the presence of unequal group sample sizes (Box, 1953; Randolph & Barcikowski, 1989; Rogan & Keselman, 1977; Scheffe, 1959; Snedecor & Cochran, 1980; Tomarkin &

Serlin, 1986). Although there is evidence to suggest lack of robustness for common parametric tests with violations to common parametric data assumptions, viability of nonparametric tests as preferred alternatives requires their demonstrated competitive performance in the presence of these assumption violations.

Early studies demonstrated strong potential in the viability of rank based nonparametric test procedures, often performing more powerfully than the *t* test in two-sample research designs in the presence of nonnormal data (Chernoff & Savage, 1958; Dixon, 1954; Hodges & Lehmann, 1956; Neave & Granger, 1968). The Mann-Whitney *U* test, also referred to as the Wilcoxon rank sum test, has more power than the *t* test with various nonnormal population distributions (Blair, 1980; Blair & Higgins, 1985; Mann & Whitney, 1947; Randles & Wolfe, 1979; Smitley, 1981; Wilcoxon, 1945, 1947). When the normality and homogeneity of variance assumptions are satisfied, a randomization test of the original scores performs as efficiently as the *t* test (Hoeffding, 1952; Lehmann & Stein, 1959). Alterations to the *t* test by first transforming the data has the potential to avoid loss of power when the standard conditions for valid inference are violated. (Rasmussen, 1985, 1986).

The multiple-group extension of the two-sample Mann-Whitney test is the Kruskal-Wallis test (Kruskal & Wallis, 1952). The Kruskal-Wallis test is a rank-based alternative to the one-way ANOVA. The Kruskal-Wallis test has power comparable or greater than the one-way ANOVA when groups have identical distribution shapes, including nonnormal distributions, with asymptotic relative efficiencies (AREs) ranging from 0.864 to infinitely larger than 1. (Andrews, 1954; Conover, 1980; Kruskal, 1952; Hodges and Lehmann, 1956). Similar to ANOVA, the Kruskal-Wallis test demonstrates inflated error rates when the homogeneity of variance assumption is violated, increasingly so in the presence of unequal group sample sizes (Feir-Walsh & Toothaker, 1974; Keselman et al., 1977; Tomarkin & Serlin, 1986). The multiple-group extension of the sign test is the Friedman test, which has much more variable power relative to the ANOVA than the Kruskal-Wallis test but still maintains the ARE above 1 for some distributions (Friedman, 1937; Hollander & Wolfe, 1973; Noether, 1967; Van Elteren & Noether, 1959). The Friedman test also suffers inflated error rates from violating the homogeneity of variance assumption (Harwell & Serlin, 1989).

Some researchers do not believe that nonparametric methods are available or appropriate to use for more complicated research designs. One reason for this belief is that nonparametric methods, including rank-based procedures, do not capture and use the full scope of information available in the data and consequently result in lower power than parametric procedures (Wilcoxon, 1949). This myth can be dispelled with the demonstratable power studies, though further study of these methods is warranted. Historically, nonparametric tests for interactions have been considered complicated or computationally intensive (Bradley, 1968). As technology advances, these limitations can be iteratively revisited to check if they remain justified. Nonparametric methods testing for an interaction effect using ordinal categorical variables often focus on the median, which provides a test of location shift if the distributions have the same form, or a test of equivalent distributions if distributional form is unknown (Brunner & Neumann, 1984; Kleijnen, 1987; Noether, 1981). Tests of equivalent distributions capture location shifts as well as other distributional differences.

Over thirty years have passed since Sawilowsky (1990) brought nonparametric tests of interaction to the attention of behavioral and social researchers. In this article, Monte Carlo simulation results are presented for five statistical methods to test for interaction effects: (a) the parametric ANOVA *F* test, (b) the rank transform test, (c) the aligned rank transform test, (d) the McSweeney test, and (e) the rank transform test using a chi-square critical value building on the theoretical advancements in research published by Thompson (1991). While simulation studies have researched the performance of these tests for various nonnormal distributions, there is minimal research into these methods in the presence of skew and asymmetry. This study addresses this gap by exploring the properties of nonparametric tests in the presence of skewed distributions.

Tests for Interaction

In this section I briefly describe each of the four methods included in this study. The ANOVA *F* test is also included as a reference. Further details about each method can be found in the cited references.

Rank Transform Test

Initial exploration into outlining and exploring the viability of the rank transform (RT) method is published in a line of inquiry by Conover and Iman (Conover & Iman, 1976, 1981; Iman, 1974; Iman and Conover, 1976). The steps for perform a rank transform test in the two-way design is to pool the observations, rank the observations, and calculate the parametric ANOVA *F* statistic on the ranks for the effect of interest (Conover & Iman, 1981). Significance is determined by referring the calculated *F* statistic to the parametric ANOVA numerator and denominator degrees of freedom.

When converting scores to ranks for the purpose of performing a parametric analysis in this way, the two or more sets of scores are combined prior to grouping. While transforming one set of scores to ranks will produce a distribution that is rectangular, when transforming two or more sets of pooled scores to ranks the distributions of the individual set ranks will no longer be necessarily rectangular. Differences in group measures of variance, skew, kurtosis, and bimodality are retained after methods of rank transformation albeit with reduced magnitude (Zimmerman, 2011). Of particular interest is the viability of the rank transform ANOVA as an alternative to the factorial, parametric ANOVA. Problems with the rank transform test include rejecting null main effects after testing for interaction effects in a 4 x 3 design (Lemmer, 1980), inflated Type I error in the 4 x 3 design in the presence of both nonnull main effects (Blair et al., 1987), and inflated Type I error in the 2 x 2 x 2 design (Sawilowsky et al., 1989). The rank transform ANOVA is demonstrated to be robust in the 2 x 2 layout but lacks power over the parametric ANOVA for larger effect sizes (Sawilowsky, 1989). The rank transform test is included here for further examination of its performance using data having asymmetry and skew.

Aligned Rank Transform Test

The motivation of the aligned rank procedure is to adjust the rank transform test so that it does not suffer reduced power and increased Type I error rates in the presence of nonnull main and interaction effects in a factorial ANOVA design. The procedure is performed by first aligning the data, then pooling and ranking the data, and finally referring the aligned and rank-transformed scores to the parametric ANOVA procedure. The goal of aligning is to treat whichever of the two main or interaction effects not currently being tested as nuisance parameters and remove them from the model. When testing for interaction, this means subtracting main effect estimates (means calculated within levels of each main effect) from the observations. This leaves only the interaction effect to be tested. One main effect and the interaction effect can be removed as well in balanced designs to test for the other main effect (Marascuilo & McSweeney, 1977). Reinach (1965, 1966) provides an early demonstration of the procedure.

The process of aligning and ranking the data as a test for interaction in the two-way layout, most often performed by subtracting main effect mean estimates from each observation, is a modification of the aligned rank process proposed by Hodges and Lehmann (1962). Simulation study of the aligned rank transform (ART) test demonstrates strong power properties and slightly liberal Type I error rates (Blair & Sawilowsky, 1990). The ART test has strong power while being robust for populations from normal, uniform, logistic, exponential, double exponential, and lognormal distributions (Blair & Sawilowsky, 1990; Fawcett & Salter, 1984; Groggel,1987; Mansouri & Chang, 1995; Salter & Fawcett, 1985, 1993). The ART test performs with larger power than the parametric test with non-normal distributions likely to be observed in psychological research such as the truncated normal, student, and gamma distributions (Leys & Schumann, 2010). The ART test also holds power advantages over other common nonparametric methods when testing for interaction until both main effects become large (Headrick & Vineyard, 2000; Leys & Schumann, 2010).

The McSweeney Test

The McSweeney H_Y statistic is an asymptotically chi-square test statistic used to test the hypothesis of no effect based on the ranks of aligned observations (McSweeney, 1967). Prior to calculating the McSweeney test statistic, the data must be pooled, aligned, and ranked. Taking the aligned and ranked data, the McSweeney statistic is calculated by multiplying the effect sums of squares by the total degrees of freedom and dividing by the total sums of squares. This statistic is tested against a chi-square distribution with degrees of freedom equal to what the effect degrees of freedom would be in a parametric analysis.

Simulation studies of the viability of this test yield mixed results. For normal, mixed-normal, exponential, double exponential, and chi-square population distributions in combination with small cell sample sizes, the McSweeney test demonstrates strong power at the cost of liberal Type I error (Harwell, 1991; Toothaker & Newman, 1994). Kelley et al. (1994) explored viability of the McSweeney test in the 2 x 2 x 2 factorial design for normal, uniform, *t*, and exponential population distributions with sample sizes of n = 7, 21, and 35. Results demonstrate a failure to maintain Type I error rates for null effects in treatment conditions but competitive results when the number of nonnull effects and effect sizes became large.

The Thompson T Test

Calculation of the test statistic for the Thompson *T* test when used for interaction effects is equal to the test statistic calculated in the rank transform test. However, the test statistic is referred to a chi-square distribution with one degree of freedom instead of being referred to an *F* distribution. Thompson (1991) studied the asymptotic properties of the rank transform ANOVA statistic for testing interactions in the balanced two-way design, proving that in a two-way design with exactly two levels of both main effects (or when there is only one main effect) the test statistic for interaction is asymptotically a chi-square divided by the degrees of freedom. In a two-way design with two levels for both main effects, this means that the interaction statistic has a distribution that is asymptotically a chi-square with one degree of freedom. There are no empirical studies employing this procedure to test for interaction effects.

Methods

Monte Carlo simulation methods were employed to compare Type I error rates and power properties of the following tests for two-factor main and interaction effects: the ANOVA *F* test, the rank transform test (RT), the aligned rank transform test (ART), the aligned rank McSweeney H_Y test, and the rank transform Thompson *T* test. Usage of the parametric analysis of variance *F* test assumes the residuals are normally distributed, an assumption Micceri (1989) found violated in all observed cases in education and social sciences measures. Thus, simulation of residuals in this study employs distributions that Micceri observed as frequently occurring in education and social science measures in his meta study, including the Laplace distribution, the logistic distribution, and distributions having mild to severe measures of asymmetry.

Measures of skew and asymmetry lying several standard deviations above the main body of the distribution are common in psychometric and achievement measures (Micceri, 1989; Walberg et al., 1984). Of the 440 measures observed by Micceri (1989), 49.1 percent had extreme to exponential tail weights and 30.9 percent had extreme to exponential asymmetry (note, the exponential distribution has a skewness of 2). Additionally, when using small sample sizes, the value of skewness present in the sample may increase due to potential sample-population mismatch (Tipton et al., 2017). Thus, in comparing the varying tests in the presence of increasingly severe measures of asymmetry, each of three different skewed distributions in combination with four different levels of the skew parameter is used to simulate the residuals. A skew-normal distribution was used with a mean of 0, a variance of 1, and skew values of -3, -2, -1, 0, 1, 2, and 3. A skew value of 0 for the skew-normal distribution indicates no skew (a standard normal) with increasing severities of left and right skew as the skew parameter deviates negatively or positively, respectively. A skewed logistic distribution was used with a mean of 0, a variance of 1, and skew values of 0.25, 0.33, 0.5, 1, 2, 3, and 4. A skew value of 1 for the skewed logistic distribution indicates no skew (a symmetric logistic) with increasing severities of left skew for skew parameters less than 1 and right skew for skew parameters larger than 1. An asymmetric Laplace distribution was used with a mean of 0, a variance of 1, and skew values of 4, 3, 2, 1, 0.5, 0.33, and 0.25. A skew value of 1 for the asymmetric Laplace distribution indicates no skew (a symmetric Laplace) with increasing severities of left skew for skew parameters larger than 1 and right skew for skew parameters smaller than 1.

These twenty-one distributions were crossed with four levels of cell sample size (5, 10, 30, and 50) and all possible combinations of the presence or absence of each main and interaction effect in the two-factor design. It is common for researchers who are considering nonparametric methods to select a method based on asymptotic relative efficiency, though this ratio assumes large sample sizes and is not realistically representative of research study conditions faced by researchers in educational and social science fields. Thus, the total sample sizes selected for this study, ranging from N = 20 to N = 200, are more representative of these common research conditions. Additionally, using a range of smaller to larger cell sample sizes will reveal power and robustness properties that occur as a result of potential sample-population mismatch that can occur by random chance in small sample sizes (Tipton et al., 2017).

The nonnull effect size was chosen for each cell sample size such that the power to detect each effect would be 80% under the standard normal theory, that is, for n = 5 the nonnull effect size simulated is 0.67, for n = 10 the nonnull effect size is 0.46, for n = 30 the nonnull effect size is 0.26, and for n = 50 the nonnull effect size is 0.20. Because the designs simulated are two-way designs with two levels of each main effect, these effect sizes result in 80% power under the standard normal theory for their respective cell sample sizes for both main and interaction effects given each effect has 1 degree of freedom. This raw effect size was added or subtracted from each cell based on level of each main and interaction effect. As normality with a mean of 0 and a variance of 1 is assumed in generation of the effect sizes, these raw effect sizes are standardized effect sizes for the normal distribution as well. A nominal α of 0.05 is used. In total, this creates 672 configurations of residual distributions, sample sizes, and effect size combinations that were simulated with 5,000 replications to obtain the Type I error rate and power. The R programming language (R Core Team, 2021) was used in conjunction with the 'sn', 'glogis', and 'LaplacesDemon' packages to simulate the data. Table 1 summarizes the design factors used.

Table 3.1
Design factors and levels of each design factor used in simulation

Design Factor:	Levels:
Residual distributions	3
Severities of skew for each residual distribution	7
Cell sample size	4
Null and nonnull main and interaction effect combinations	8

Using 5,000 simulations creates a 95% margin of error of 0.006 for the Type I error estimates and a 95% margin of error of 0.011 for the power estimates under standard normal theory. The 95% margin of error of 0.006 for Type I error estimates is close to the stringent error band interval of 0.005 proposed by Bradley (1978) for examining the robustness of Monte Carlo simulated Type I error rates for hypothesis testing. This error band interval becomes ±0.005 when a nominal α of 0.05 is used. Bradley also proposed a second more liberal error band that becomes ±0.02 when a nominal α of 0.05 is used, and that is the band adopted in this study to indicate whether a method is robust. That is, any proposed hypothesis test with an empirical Type I error rate greater than 0.07 is deemed as not being robust to violations of symmetry or adequate sample size assumptions for that simulated design. Thus, power comparisons for the proposed hypothesis tests will only be considered for those tests maintaining an empirical Type I error rate that is less than or equal to 0.07 for each simulated design. Proposed testing methods will be called slightly conservative or slightly liberal if they maintain an empirical Type I error rate which deviates less than 0.01 below and above 0.05, respectively. Methods will be called conservative or liberal if their Type I error rate deviates 0.02 or more from 0.05. Both main and interaction effects are estimated and recorded, yet the primary empirical measures of interest refer to performance of a test for detecting an interaction. Results are provided in table format for all simulated designs.

Results

These results summarize the comparison of the testing methods ability to detect an interaction. The focus is on the empirical Type I error rate and the power results when testing for interaction in the presence and absence of main effects for all three distributions.

Skew-normal Distribution

The only times any of the included methods resulted in liberal rejection of the null hypothesis in the skew-normal distribution designs were when n = 5. When n = 5, the Thompson *T* statistic often demonstrates an empirical Type I error rate greater than 0.07 in the presence of only one nonnull main effect. When n = 5 or n = 10 and both main effects are nonnull, the rank transform test and Thompson *T* test performed demonstrably conservative as shown in Tables 2 and 3. This conservatism for these two tests becomes negligible when n becomes 30 or larger. When n becomes 30 or larger, all proposed testing methods are competitive with negligible differences in power, that is, power differences less than 0.01. Notably, the ANOVA *F* test performed marginally better than the remaining tests with larger cell sample sizes and less extreme levels of skew. However, the aligned rank transform test performs competitively to the ANOVA *F* for power and Type I error regardless of nonnull main effect presence and in the presence of extreme skew as shown in Tables 3, 4, and 5. This suggests that for all cell sample sizes and levels of skew severity for the skew-normal distribution, the aligned rank transform test performs competitively to the ANOVA *F* test for power and Type I error rate. Tables 3, 4, and 5 also demonstrate that the McSweeney H_Y test performs with competitive power to the ANOVA *F* test regardless of whether nonnull main effects are present and in the presence of extreme skew. Type I error rates of the McSweeney H_Y test remain competitive to the ANOVA *F* for all cell sample sizes apart from n = 5 where the McSweeney H_Y test has error rates between 0.06 and 0.07. This suggests that for all but n

= 5 cell sample sizes and for all levels of skew severity for the skew-normal distribution, the McSweeney H_Y test performs competitively compared to the ANOVA F test. These results remain consistent in designs not represented in the provided tables.

Table 3.2
Type I Error and Power Results for Skew-normal with two nonnull main effects and a skew parameter of 1

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0554	0.0482	0.0518	0.0488	0.92	0.921	0.9286	0.919
Rank Transform	0.0092	0.0354	0.0512	0.0482	0.7094	0.8706	0.9148	0.9102
ART	0.0582	0.05	0.0526	0.0456	0.914	0.9128	0.9166	0.9142
McSweeney	0.0698	0.054	0.0546	0.0466	0.9264	0.9168	0.9178	0.9154
Thompson	0.017	0.0406	0.0536	0.0498	0.7662	0.8852	0.9178	0.9116

Note. ART refers to the aligned rank transform test.

Table 3.3
Type I Error and Power Results for Skew-normal with two nonnull main effects and a skew parameter of 3

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0486	0.0538	0.049	0.0528	0.9886	0.9902	0.9896	0.9914
Rank Transform	0.0048	0.0258	0.0424	0.0502	0.743	0.9602	0.9894	0.9912
ART	0.0558	0.0536	0.0484	0.0544	0.9852	0.9908	0.9926	0.992
McSweeney	0.0676	0.0566	0.0496	0.055	0.9878	0.9914	0.993	0.9926
Thompson	0.008	0.0312	0.0438	0.0524	0.7978	0.9656	0.9898	0.9918

Note. ART refers to the aligned rank transform test.

Table 3.4

Type I Error and Power Results for Skew-normal with one null and one nonnull main effect and a skew parameter of 3

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0464	0.0472	0.057	0.0496	0.9874	0.9872	0.9896	0.9904
Rank Transform	0.0434	0.0512	0.0568	0.0476	0.9824	0.985	0.9908	0.9918
ART	0.0534	0.0486	0.0526	0.0468	0.9884	0.9898	0.9922	0.9918
McSweeney	0.0632	0.0526	0.054	0.048	0.9906	0.9902	0.9922	0.992
Thompson	0.0606	0.0568	0.0588	0.0486	0.9876	0.9878	0.9916	0.992

Note. ART refers to the aligned rank transform test.

Table 3.5

Type I Error and Power Results for Skew-normal with two null main effects and a skew parameter of 3

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.048	0.0534	0.051	0.0468	0.9882	0.9876	0.99	0.9898
Rank Transform	0.0528	0.0538	0.0518	0.045	0.9844	0.988	0.992	0.9922
ART	0.055	0.053	0.0514	0.046	0.9882	0.988	0.992	0.9924
McSweeney	0.066	0.0566	0.0522	0.0464	0.9906	0.9892	0.9924	0.9924
Thompson	0.0692	0.0606	0.0532	0.0462	0.9894	0.9898	0.9924	0.9922

Note. ART refers to the aligned rank transform test.

Skewed Logistic Distribution

Once again, any of the proposed testing methods result in liberal rejection only when the cell sample size equals 5 and at least one main effect is null. For designs with n = 5 cell sample size and one nonnull main effect, the Thompson *T* statistic yielded an empirical alpha greater than 0.07 for multiple skew severities. However, once the cell sample size became 10 or larger, the aligned rank transform test, the McSweeney H_Y test, and the Thompson *T* test performed more powerfully than the parametric ANOVA for all cell sample sizes and severities of the skew parameter while maintaining empirical Type I error rates within the acceptable range as shown in Tables 6, 7, 8, and 9. For larger sample sizes, power advantages of the nonparametric tests compared to the ANOVA *F* test range from 2-5 percentage points, increasing as the severity of the skew parameter increases. These results remained consistent in the presence and absence of nonnull main effects. Additionally, these results remain consistent in designs not represented in the provided tables.

Table 3.6
Type I Error and Power Results for Skew Logistic with two nonnull main effects and a skew parameter of 1

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0448	0.0476	0.0458	0.0484	0.3566	0.336	0.344	0.331
Rank Transform	0.0408	0.0516	0.0448	0.0458	0.3318	0.343	0.3606	0.3554
ART	0.0524	0.0488	0.046	0.0468	0.3766	0.3522	0.3644	0.3564
McSweeney	0.0636	0.053	0.0472	0.0476	0.4116	0.3618	0.3698	0.3592
Thompson	0.058	0.058	0.0464	0.0472	0.381	0.3686	0.367	0.3608

Note. ART refers to the aligned rank transform test.

Table 3.7

Type I Error and Power Results for Skew Logistic with two nonnull main effects and a skew parameter of 2

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0462	0.048	0.0482	0.0474	0.4756	0.47	0.4584	0.4632
Rank Transform	0.039	0.0402	0.0502	0.0514	0.4182	0.4712	0.4864	0.4972
ART	0.054	0.0452	0.0536	0.0506	0.4982	0.489	0.4866	0.5058
McSweeney	0.0658	0.048	0.0548	0.0514	0.5316	0.4986	0.4898	0.5068
Thompson	0.0556	0.049	0.0536	0.0528	0.479	0.4968	0.4942	0.5038

Note. ART refers to the aligned rank transform test.

Table 3.8

Type I Error and Power Results for Skew Logistic with two nonnull main effects and a skew parameter of 3

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0488	0.052	0.0554	0.0452	0.52	0.499	0.5068	0.5082
Rank Transform	0.0352	0.047	0.056	0.0456	0.4496	0.5138	0.541	0.5544
ART	0.0524	0.0504	0.0572	0.0454	0.5456	0.5368	0.549	0.5558
McSweeney	0.0622	0.055	0.058	0.0462	0.5814	0.5458	0.5524	0.5584
Thompson	0.05	0.0564	0.058	0.0474	0.5074	0.5344	0.5482	0.5582

Note. ART refers to the aligned rank transform test.

Table 3.9

Type I Error and Power Results for Skew Logistic with two nonnull main effects and a skew parameter of 4

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.05	0.0454	0.0512	0.0508	0.5578	0.5436	0.5274	0.5228
Rank Transform	0.039	0.0508	0.052	0.051	0.4838	0.5478	0.5686	0.5716
ART	0.0582	0.0546	0.0532	0.0512	0.5836	0.582	0.572	0.576
McSweeney	0.0692	0.0574	0.0548	0.0518	0.6206	0.591	0.576	0.5792
Thompson	0.0544	0.0572	0.0538	0.0528	0.545	0.5746	0.575	0.5758

Note. ART refers to the aligned rank transform test.

Asymmetric Laplace Distribution

It is again only when the cell sample size equals 5 that any of the proposed testing methods resulted in liberal rejection of the null hypothesis. Furthermore, it is only the McSweeney H_Y statistic and Thompson T statistic that become liberal. The McSweeney H_Y statistic exceeded the liberal Type I error rate threshold of 0.07 for multiple asymmetry severity levels and combinations of null and nonnull main effects. The Thompson T statistic yielded Type I error rate above 0.07 when there was one nonnull main effect for n = 5 cell sample sizes. These methods remained competitive and within the acceptable empirical Type I error band interval once the cell sample size became 10. For small cell sample sizes, the rank transform method performs most powerfully when there is an absence of both main effects and extreme levels of asymmetry while never resulting in a liberal rejection for Type I error rates as shown in Table 9. For small cell sample sizes and the presence of both nonnull main effects, the McSweeney H_Y test performs most powerfully in the presence of extreme asymmetry as shown in table 10.

The McSweeney H_Y test performs powerfully for either case of main effects in the presence of less extreme asymmetry parameters as shown in Tables 11 and 12. Tables 11 and 12 also capture that for smaller sample sizes and less extreme asymmetry parameters, the rank transform test and rank transform *T* test perform powerfully where there are no main effects but lose considerable power when there are main effects. For larger cell sample sizes, all nonparametric methods remain competitive with data from an asymmetric Laplace distribution as demonstrated in Tables 10, 11, 12, and 13. The ANOVA *F* statistic, however, performs with noticeably less power compared to the nonparametric tests for larger cell sample sizes. For larger cell sample sizes, the power advantage of the nonparametric tests range from 12-17 percentage points, increasing as the severity of the asymmetry parameter increases. results remain consistent in designs not represented in the provided tables.

Table 3.10
Type I Error and Power Results for Asymmetric Laplace with two null main effects and an asymmetry parameter of 1/4

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.045	0.045	0.0478	0.0532	0.1892	0.1932	0.1766	0.165
Rank Transform	0.0536	0.0474	0.0506	0.0522	0.2942	0.3254	0.3436	0.3458
ART	0.0578	0.0522	0.052	0.0516	0.254	0.2858	0.3132	0.321
McSweeney	0.066	0.054	0.0526	0.0516	0.2808	0.2918	0.3156	0.3214
Thompson	0.073	0.0528	0.0526	0.0536	0.3378	0.3478	0.3518	0.3498

Note. ART refers to the aligned rank transform test.

Table 3.11

Type I Error and Power Results for Asymmetric Laplace with two nonnull main effects and an asymmetry parameter of 1/4

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0472	0.0482	0.0528	0.0518	0.2098	0.1872	0.1732	0.171
Rank Transform	0.049	0.05	0.0542	0.05	0.2352	0.2654	0.3036	0.3228
ART	0.0554	0.0518	0.0538	0.0524	0.2734	0.2808	0.3094	0.3234
McSweeney	0.0658	0.054	0.0544	0.0524	0.3018	0.2866	0.3096	0.324
Thompson	0.0672	0.0574	0.0564	0.0512	0.2826	0.284	0.311	0.3258

Note. ART refers to the aligned rank transform test.

Table 3.12

Type I Error and Power Results for Asymmetric Laplace with two null main effects and an asymmetry parameter of 1

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0446	0.052	0.0488	0.048	0.7962	0.8014	0.8026	0.7894
Rank Transform	0.0492	0.0532	0.0484	0.0462	0.8428	0.8856	0.9046	0.9134
ART	0.0504	0.0552	0.0484	0.0492	0.8362	0.8742	0.8996	0.9116
McSweeney	0.0622	0.057	0.0496	0.0492	0.855	0.8792	0.9006	0.9128
Thompson	0.068	0.06	0.051	0.0478	0.8746	0.8962	0.9066	0.915

Note. ART refers to the aligned rank transform test.

Table 3.13
Type I Error and Power Results for Asymmetric Laplace with two nonnull main effects and an asymmetry parameter of 1

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0468	0.0486	0.0436	0.0514	0.8004	0.8112	0.8012	0.804
Rank Transform	0.0148	0.0366	0.0464	0.0536	0.6284	0.7884	0.8728	0.8974
ART	0.0558	0.0488	0.0464	0.0554	0.845	0.8794	0.9026	0.9174
McSweeney	0.0642	0.0506	0.047	0.0566	0.8666	0.8846	0.9046	0.918
Thompson	0.025	0.0444	0.0492	0.0546	0.6916	0.8078	0.8772	0.8986

Note. ART refers to the aligned rank transform test.

Standard Normal Distribution

Tables 14, 15, and 16 provide results for the standard normal distribution with 2 null, 1 null and 1 nonnull, and 2 nonnull main effects, respectively. These tables are provided to give a reference of how the nonparametric tests compare to the ANOVA *F* test when the normality assumption can safely be assumed. As shown in Table 16, the conservatism demonstrated by the rank transform test and the Thompson *T* test for small cell sample sizes in the presence of two nonnull main effects is present for the standard normal distribution. The McSweeney test demonstrates liberal Type I error rates for small cell sample sizes but converges to similar Type I error rates and power to that of the aligned rank transform test as cell sample sizes increase. The aligned rank transform test maintains acceptable Type I error rates and has similar power to the ANOVA *F* test for small cell sample sizes. The ANOVA *F* test performs with up to 2 percentage points of larger power at the largest cell sample size (n = 50, N = 200). These results suggest that when normality can be safely assumed, the aligned rank transform test performs competitively to the ANOVA *F* test.

Table 3.14
Type I Error and Power Results for Standard Normal with two null main effects

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0488	0.055	0.0536	0.049	0.8044	0.7952	0.8	0.8
Rank Transform	0.0522	0.055	0.0518	0.0512	0.7908	0.7824	0.778	0.7792
ART	0.0516	0.0556	0.052	0.0512	0.8006	0.7814	0.7816	0.7772
McSweeney	0.0624	0.0592	0.0538	0.0516	0.8228	0.7904	0.7856	0.7792
Thompson	0.0682	0.0626	0.0552	0.0516	0.8234	0.802	0.785	0.7832

Note. ART refers to the aligned rank transform test.

Table 3.15
Type I Error and Power Results for Standard Normal with one null and one nonnull main effects

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0504	0.0518	0.0522	0.051	0.7972	0.813	0.807	0.8058
Rank Transform	0.0496	0.0496	0.0544	0.051	0.77	0.781	0.7952	0.7808
ART	0.0538	0.0532	0.054	0.0526	0.7916	0.7956	0.7936	0.7794
McSweeney	0.066	0.0558	0.055	0.0534	0.8202	0.8034	0.799	0.7808
Thompson	0.0678	0.0576	0.0566	0.0522	0.8106	0.8036	0.8002	0.7842

Note. ART refers to the aligned rank transform test.

Table 3.16

Type I Error and Power Results for Standard Normal with two nonnull main effects

	Null interaction effect				Nonnull interaction effect			
Cell size:	5	10	30	50	5	10	30	50
ANOVA	0.0538	0.051	0.049	0.0562	0.8092	0.7982	0.795	0.7972
Rank Transform	0.0276	0.0436	0.0474	0.0546	0.6378	0.7548	0.7722	0.7766
ART	0.0574	0.0502	0.0474	0.055	0.8008	0.7844	0.777	0.7784
McSweeney	0.0666	0.053	0.0496	0.0554	0.8264	0.7916	0.7808	0.7826
Thompson	0.0402	0.0514	0.0486	0.0556	0.6986	0.7756	0.7796	0.78

Note. ART refers to the aligned rank transform test.

Discussion

Micceri (1989) highlights the need for tests capable of handling the prevalence and range of variability in the skew and asymmetry present in the fields of the behavioral and social sciences. The recommendation of a nonparametric test over a parametric one requires the nonparametric test to be robust to Type I error rate, maintain competitive power properties when normality can safely be assumed, and yield an increase in power for multiple types of nonnormally distributed data. In testing for two-factor interactions, the simulation results presented in this study suggest that there is not a uniformly most powerful test that maintains acceptable Type I error rates for all designs. However, there are situations in which some tests can be clearly ruled out as viable, particularly if the presence of null or nonnull main effects can be confirmed. Furthermore, given that nonparametric methods demonstrate competitive empirical power compared to the ANOVA *F* test even when normality and symmetry assumptions are met and competitive to larger power when normality and symmetry are not safe assumptions, there is no need to use the traditional parametric test and gamble that the assumptions for valid inference are satisfied. The results provided demonstrate that an alternative nonparametric test exists with little to no risk.

Results suggest, particularly with regard to cell sample size and method, increased data scrutiny can contribute to considerable increases in power. Because of the strong conservative often demonstrated by the rank transform test and the Thompson *T* test in the presence of two nonnull main effects for small sample sizes and the similar power properties compared to the ART test for larger cell sample sizes, these two tests can be dismissed as viable nonparametric tests of interaction. When these two tests demonstrated minor power increases over the ART test, they also demonstrated increases in Type I error rate. When data is asymmetrically Laplace distributed, the ART test maintains Type I error rate within the liberal rejection region and always demonstrates larger power than the ANOVA *F* test, that is, power increases of 7-17 percentage points depending upon cell sample size. This power advantage increases as the asymmetry effect increases suggesting the ART test is preferable for the asymmetric Laplace distribution. The ART test performs similarly for the skew logistic distribution. The ART test maintains Type I error rates within the liberal rejection region and performs with larger power (2-5 percentage points depending upon cell sample size) than the ANOVA *F* for all sample sizes and levels of skew. This power advantage over the ANOVA *F* also increases as the skew parameter increases. While the McSweeney test also performs with larger power than the ANOVA *F* once cell sample sizes become 10 or more, the ART test maintains a more conservative Type I error rate with competitive power. These results suggest the ART is preferable for the skew logistic design as well. Admittedly, the ANOVA *F* remains robust and competitive for the standard and skew-normal distribution for all cell sample sizes and degrees of skew. However, as the ART test is competitive in both Type I error rates and power for the standard and skew-normal as well, there appears minimal disadvantages to using it compared to all other testing methods when there is uncertainty about the population distribution. These results highlight the danger of incorrectly assuming normality when using the ANOVA *F* test. Simulations results presented indicate using the ART in place of the ANOVA *F* test can mitigate power loss from incorrectly assuming normality by as much as 17 percentage points.

Restated, the recommendation of a nonparametric test over a parametric one requires the nonparametric test to be robust to Type I error rate, maintain competitive power properties when normality can safely be assumed, and yield an increase in power for multiple types of nonnormally distributed data. The presented results demonstrate that that the aligned rank transform test is robust to Type I error rate, maintains competitive power properties when normality can be safely assumed, and yields an increase in power compared to the ANOVA *F* test for all simulated designs except for the standard normal distribution with large cell sample sizes where the ANOVA *F* test yields up to 2 percentage points in larger power. This suggests that unless normality is a truly safe assumption, which would necessitate a cell sample size on the larger end of the sizes simulated in this research, the aligned rank transform test becomes the preferable test of interaction effect. If no assumption can be made regarding underlying distribution, or when cell sample sizes are small regardless, the aligned rank transform test meets all of the required criteria to be the recommended test given its power advantages for nonnormally distributed data.

This research employs simulations of residuals following only normal, logistic, and Laplace distributions with varying severities of skew. I studied the nonparametric tests based on ranks using seven levels of skew severity for each of the three mentioned distributions. This study includes only a two-factor design with two levels of each main effect. This line of research naturally reveals several subsequent lines of inquiry. As Micceri (1989) acknowledges, there should also be acknowledgement in the field of psychometric measures of the prevalence of multimodality, lumpiness, and other frequently occurring deviations from normality in population distributions apart from skew. Additionally, further study is needed regarding the suggested nonparametric tests in the presence of designs having more than two levels of each main effect, having unbalanced cell sample sizes, and in designs where cell residuals have different distributions including but not limited to different severities of skew. Admittedly, there is research into efficacy comparisons of some parametric methods under some of these conditions (Akritas, 1990). Further research is needed in calculation of effect sizes for significant nonparametric test results when using nonparametric tests based on ranks.

This research supports the use of nonparametric tests in testing for the presence of an interaction location shift by demonstrating their larger power than the parametric ANOVA *F* test for the nonnormal distributions. Given this increased power, the use of these nonparametric tests may provide for a more accurate way of estimating effect sizes as opposed to calculating mean differences. Given the results demonstrated in this study and the prevalence of nonparametric techniques introduced, a proportionate increase in efficacy comparisons for these methods is needed to ensure that robust testing methods which require fewer distributional assumptions than traditional parametric tests do not fall through the cracks and their potential left untapped.

References

Akritas, M. G. (1990). The rank transform method in some two-factor designs. *Journal of the American Statistical Association*, *85*(409), 73–78. https://doi.org/10.1080/01621459.1990.10475308

Andrews, F. C. (1954). Asymptotic behavior of some rank tests for analysis of variance. *Annals of Mathematical Statistics*, *25*, 724–736.

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

Bishop, T. (1976). *Heteroscedastic ANOVA, MANOVA and multiple comparisons* [Doctoral dissertation, Ohio State University].

Blair, R. C. (1980). *A comparison of the power of the two independent means t test to that of the wilcoxon's rank-sum test for samples of various populations* [Doctoral dissertation, University of South Florida].

Blair, R. C., & Higgins, J. J. (1980a). A comparison of the power of the t test and the Wilcoxon statistic when samples are drawn from a certain mixed normal distribution. *Evaluation Review*, *4*, 645–656.

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

Blair, R. C., & Higgins, J. J. (1981). A note on the asymptotic relative efficiency of the Wilcoxon rank-sum test relative to the independent means t test under mixtures of two normal distributions. *British Journal of Mathematical and Statistical Psychology, 34*(1), 124–128. https://doi.org/10.1111/j.2044-8317.1981.tb00623.x

Blair, R. C., & Higgins, J. J. (1985). A comparison of the power of the paired samples rank transform statistic to that of Wilcoxon’s signed rank statistic. *Journal of Educational Statistics*, *10*(4), 368–383.

Blair, R. C., & Sawilowsky, S. S. (1990). American Educational Research Association. In *A test for interaction based on the rank transform*. Boston, MA.

Blair, R. C., Sawilowsky, S. S., & Higgins, J. J. (1987). Limitations of the rank transform statistic in tests for interactions. *Communications in Statistics - Simulation and Computation*, *16*(4), 1133–1145. https://doi.org/10.1080/03610918708812642

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16. https://doi.org/10.3102/0013189x013006004

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

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

Box, G. E. P. (1953). Non-normality and tests of variances. *Biometrika*, *40*, 318–355.

Box, G. E. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, *25*(2), 290–302. https://doi.org/10.1214/aoms/1177728786

Bradley, J. V. (1968). *Distribution-free statistical tests*. Prentice-Hall.

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician, 31*(4), 147–150. https://doi.org/10.2307/2683535

Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical Psychology*, *31*, 144–154.

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

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

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

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

Brown, M. B., & Forsythe, A. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, *16*, 129–132.

Brunner, E., & Neumann, N. (1984). Rank tests for the 2 x 2 split plot design. *Metrika*, *31*, 233–243.

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

Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, *3*, 27–38.

Cochran, W. G., & Cox, G. M. (1950). *Experimental designs*. Wiley.

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). John Wiley.

Conover, W. J., & Iman, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs. *Communications in Statistics - Theory and Methods*, *5*(14), 1349–1368. https://doi.org/10.1080/03610927608827447

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, *35*(3), 124–129. https://doi.org/10.2307/2683975

DeWees, T. A., Mazza, G. L., Golafshar, M. A., & Dueck, A. C. (2020). Investigation into the effects of using normal distribution theory methodology for Likert scale patient-reported outcome data from varying underlying distributions including floor/ceiling effects. *Value in Health*, *23*(5), 625–631.

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

Fawcett, R. F., & Salter, K. C. (1984). A Monte Carlo study of the F test and three tests based on ranks of treatment effects in randomized block designs. *Communications in Statistics - Simulation and Computation*, *13*(2), 213–225. https://doi.org/10.1080/03610918408812368

Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores, and Kruskal-Wallis test under violations of assumptions. *Educational and Psychological Measurement*, *34*, 789–799.

Feng, Y., Hancock, G. R., & Harring, J. R. (2019). Latent growth models with floor, ceilings, and random knots. *Multivariate Behavioral Research*, *54*(5), 751–770.

Fisher, R. A. (1922). *Statistical Methods for Research Workers*. Oliver & Boyd.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

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

Goodard, R. H., & Lindquist, E. F. (1940). An empirical study of the effects of heterogeneous within groups variance upon certain F-tests of significance in analysis of variance. *Psychometrica*, *5,* 263–274.

Groggel, D. J. (1987). A Monte Carlo Study of rank tests for block designs. *Communications in Statistics - Simulation and Computation*, *16*(3), 601–620. https://doi.org/10.1080/03610918708812607

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). McGraw-Hill.

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

Harwell, M. R. (1991). Completely randomized factorial analysis of variance using ranks. *British Journal of Mathematical and Statistical Psychology, 44*(2), 383–401. https://doi.org/10.1111/j.2044-8317.1991.tb00970.x

Harwell, M. R., & Serlin, R. C. (1989). Annual Meeting of the American Educational Research Association. In *An empirical study of the Friedman test under covariance heterogeneity*. San Francisco.

Havlicek, L. L., & Peterson, N. L. (1974). Robustness of the t-test in a guide for researchers on effect of violations of assumptions. *Psychological Reports*, *34*, 1095–1114.

Headrick, T. C., & Vineyard, G. (2000). *An empirical investigation of four tests for interaction in the context of factorial analysis of covariance.* Southern Illinois University at Carbondale.

Hellstrom, A. (1993). The normal distribution in scaling subjective stimulus difference: Less “normal” than we think? *Perception & Psychophysics*, *54*(1), 82–92.

Hemelrijk, J. (1961). Experimental comparison of Student’s and Wilcoxon’s two sample test. *Quantitative Methods in Psychology*. Interscience.

Hettmansperger, T. P., McKean, J. W., & Sheather, S. J. (2000). Robust nonparametric methods. *Journal of the American Statistical Association*, *95*(452), 1308–1312.

Hildebrand, D. K. (1986). *Statistical thinking for behavioral scientists*. Duxbury Press.

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

Hodges, J. L., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics*, *33*(2), 482–497. https://doi.org/10.1214/aoms/1177704575

Hoeffding, W. (1952). The large sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, *23*, 169–192.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric methods*. John Wiley.

Hornsnell, G. (1953). The effect of unequal group variances on the F-test for the homogeneity of group means. *Biometrika*, *40*, 128–136.

Iman, R. L. (1974). A power study of a rank transform for the two-way classification model when interaction may be present. *Canadian Journal of Statistics*, *2*(1-2), 227–239. https://doi.org/10.2307/3314695

Iman, R. L., & Conover, W. J. (1976). *A comparison of several rank tests for the two-way layout* (SAND76-0631). Albuquerque, NM: Sandia Laboratories.

Kelley, D. L., Sawilowsky, S. S., & Blair, R. C. (1994). Midwestern Educational Research Association. In *Comparison of ANOVA, McSweeney, Bradley, Harwell-Serlin, and Blair-Sawilowsky tests in the balanced 2x2x2 layout.* Chicago, IL.

Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd ed.). Holt, Rinehart, and Winston.

Keselman, H. J., Rogan, J. C., & Feir-Walsh, B. J. (1977). An evaluation of some nonparametric tests for location equality. *British Journal of Mathematical and Statistical Psychology*, *30*, 213–221.

Kleijnen, P. C. (1987). *Statistical tools for simulation practitioners*. Marcel Dekker.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *Annals of Mathematical Statistics*, *23*, 525–545.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Lehmann, E. L., & Stein, C. (1959). *Testing statistical hypotheses*. John Wiley.

Lemmer, H. H. (1980). Some empirical results on the two-way analysis of variance by ranks. *Communications in Statistics - Theory and Methods*, *9*(14), 1427–1438. https://doi.org/10.1080/03610928008827972

Leys, C., & Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology*, *46*(4), 684–688. https://doi.org/10.1016/j.jesp.2010.02.007

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and Education.* Houghton.

Liu, Q., & Wang, L. (2021). T-test and ANOVA for data with ceiling and/or floor effects. *Behavior Research Methods, 53*(1), 264–277.

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement*, *7*, 263–269.

Mandeville, G. K. (1972). A new look at treatment differences. *American Educational Research Journal*, *9*, 311–321.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Mansouri, H., & Chang, G.-H. (1995). A comparative study of some rank tests for interaction. *Computational Statistics & Data Analysis*, *19*(1), 85–96. https://doi.org/10.1016/0167-9473(93)e0045-6

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Brooks/Cole.

Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences*. Freeman.

McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit model. *Gifted Child Quarterly*, *54*(4), 314–320.

McSweeney, M. T. (1967). *An Empirical Study of Two Proposed Nonparametric Tests for Main Effects and Interaction* [Doctoral dissertation, University of California, Berkeley].

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

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

Noether, G. E. (1967). *Elements of nonparametric statistics*. Wiley.

Noether, G. E. (1981). Comment. *American Statistician*, *35*(3), 129–132.

Noortgate, W. V. den, Boeck, P. D., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational & Behavioral Statistics*, *28*(4), 369–386.

Nunnally, J. C. (1975). *Introduction to statistics for psychology and education*. McGraw-Hill.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Nussbaum, E. M. (2015). *Categorical and nonparametric data analysis: Choosing the best statistical technique*. Routledge.

Pearson, E. S. (1929). The analysis of variance in cases of nonnormal variation. *Biometrika*, *23*, 259–286.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric tests*. John Wiley.

Randolph, E. A., & Barcikowski, R. S. (1989). 11th Annual Meeting of the Mid-Western Educational Research Association. In *Type I error rate when real study values are used as population parameters in a Monte Carlo study*. Chicago.

Rasmussen, J. L. (1985). The Power of Student's t and Wilcoxon W statistics. *Evaluation Review*, *9*(4), 505–510.

Rasmussen, J. L. (1986). An evaluation of parametric and nonparametric tests on modified and nonmodified data. *British Journal of Mathematical and Statistical Psychology*, *39*, 213–220.

Reinach, S. G. (1965). A nonparametric analysis for a multiway classification with one element per cell. *South African Journal of Agricultural Science*, *8*, 941–960.

Reinach, S. G. (1966). *Distribution-free methods in experimental design* [Doctoral dissertation, University of Pretoria].

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

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. *American Educational Research Journal*, *14*, 493–498.

Salter, K. C., & Fawcett, R. F. (1985). A robust and powerful rank test of treatment effects in balanced incomplete block designs. *Communications in Statistics - Simulation and Computation*, *14*(4), 807–828. https://doi.org/10.1080/03610918508812475

Salter, K. C., & Fawcett, R. F. (1993). The art test of interaction: A robust and powerful rank test of interaction in factorial models. *Communications in Statistics - Simulation and Computation*, *22*(1), 137–153. https://doi.org/10.1080/03610919308813085

Savalei, V. (2006). Logistic approximation to the normal: the KL rationale. *Psychometrika*, *71*(4), 763–767.

Sawilowsky, S. S. (1989). Rank transform: The bridge is falling down. In American Educational Research Association. San Francisco.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research*, *60*(1), 91–126. https://doi.org/10.3102/00346543060001091

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. https://doi.org/10.1037/0033-2909.111.2.352

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics, 14*(3), 255–267. https://doi.org/10.3102/10769986014003255

Scheffe, H. (1959). *The analysis of variance*. Wiley.

Smitley, W. D. S. (1981). *A comparison of the power of the two independent means t test and the Mann-Whitney U test* [Doctoral dissertation, University of South Florida].

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

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680.

Still, A. W., & White, A. P. (1981). The approximate randomization test as an alternative to the F test in analysis of variance. *British Journal of Mathematical and Statistical Psychology, 34*(2), 243–252. https://doi.org/10.1111/j.2044-8317.1981.tb00634.x

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

Thompson, G. L. (1991). A note on the rank transform for interactions. *Biometrika, 78*(3), 697–701. https://doi.org/10.1093/biomet/78.3.697

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2017). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review, 41*(5), 472-505.

Tomarkin, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*(1), 90–99.

Toothaker, L. E., & Newman, D. (1994). Nonparametric competitors to the two-way ANOVA. *Journal of Educational and Behavioral Statistics*, *19*(3), 237–273. https://doi.org/10.2307/1165296

Van Elteren, P., & Noether, G. E. (1959). The asymptotic efficiency of the Xr2 test for a balanced incomplete block design. *Biometrika*, *46*, 465–477.

Walberg, J. H., Strykowski, B. F., Rovai, E., & Hurg, S. S. (1984). Exceptional performance. *Review of Educational Research*, *54*, 87–112.

Welch, B. L. (1937). The significance of the difference between two means when the population Variances are Unequal. *Biometrika*, *29*, 350–362.

Wilcox, R. R., Charlin, V., & Thompson, K. (1986). New Monte Carlo results on the Robustness of the ANOVA, F, W, and F* statistics. *Communications in Statistics - Simulations and Computation*, *15,* 933–944.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–82. https://doi.org/10.2307/3001968

Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics*, *3*, 119–122.

Wilcoxon, F. (1949). *Some rapid approximate statistical procedures*. American Cyanamid.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). McGraw-Hill.

Zimmerman, D. W. (2011). Inheritance of properties of normal and non-normal distributions after transformation of scores to ranks. *Psicologica*, *32*, 65–85.

CHAPTER 4

NONPARAMETRIC TESTS OF INTERACTION USING VAN DER WAERDEN NORMAL SCORES FOR THE TWO-WAY DESIGN WITH SKEWED DISTRIBUTIONS

Abstract

This study compares the ANOVA *F* test and three nonparametric competitors with two-factor designs for empirical Type I error rate and power when testing for an interaction of the factors. These nonparametric competitors are the rank transform test, aligned rank transform test, and McSweeney test with a Van der Wearden normal scores data transformation in place of the rank transformation. I perform simulations of 2 x 2 designs with cell sizes of 5, 10, 30, 50, 100, 200, and 500. I employ samples having distributions commonly found in the educational and social sciences including skew normal, skew logistic, and asymmetric Laplace distributions each having four levels of skew and asymmetry effects. The ANOVA *F* is robust for Type I error but lacks power compared to nonparametric alternatives for skew logistic and asymmetric Laplace samples. The aligned rank transform and McSweeney tests with Van der Waerden normal scores data transformations demonstrate competitive Type I error and considerable power advantages over ANOVA *F* for the skew logistic and asymmetric Laplace distributions.

Keywords: Nonparametric, Rank Transform, Aligned Rank Transform, Normal Scores

Introduction

Researchers in the behavioral and social sciences fields require statistical inference methods that are robust in the presence of deviations from the common parametric analysis of variance (ANOVA) assumptions. Real world education and psychological measures often occur having mixed-normal distributions (Bradley, 1978, 1980a, 1980b, 1980c, 1982; Blair, 1980; Micceri, 1989; Still & White, 1981; Tan, 1982). Education and social sciences measures, such as Likert scales and other self-reported measures, often have ceiling and floor effects that result in the need for methods capable of handling censored and truncated distributions (DeWees et al., 2020; Feng et al., 2019; Liu & Wang, 2021; McBee, 2010). Other frequent departures from normality found in psychometric measures include asymmetry, multimodality, and skew (Bradley, 1977; Micceri, 1989). Micceri's (1989) meta-analysis analyzed 440 large sample educational and psychological measures and found all to be significantly nonnormal via some class of contamination. Of the 440 measures analyzed by Micceri, only 15.2% had both tail weights at or about normal, 50.2% had at least one tail heavier than normal, and 14.8% had both tail weights less than normal. Of the remainder, 3.2% were concluded as uniform and 16.6% were concluded as Laplace. Additionally, of the 440 measures, 40.7% were concluded as moderately asymmetric, 19.5% as extremely asymmetric, and 11.4%as having exponential asymmetry. Of the categories for the measures analyzed by Micceri, nonnormal tail weights and asymmetry is demonstrated most prevalent in measures of criterion mastery, where 60% of measures were found to be distributed as Laplace, 37.1% of measures had extreme asymmetry, and 57.1% of measures had exponential asymmetry. Thus, Micceri suggests the existence of normally distributed data in educational and psychological measures is as improbable as the existence of a unicorn. Micceri's findings not only directly conclude the prevalence of the Laplace distribution in education and social science measures but also provide evidence of the prevalence of the logistic distribution by highlighting the frequency of mixed-normal measures and criterion mastery measures. The logistic distribution is often used to approximate mixed-normal distributions, model distributions of subjective quantitative measures, and is prevalent in item response theory analysis (Hellstrom, 1993; Noortgate et al., 2003; Savalei, 2006). This would imply that researchers in these fields who limit their methods of analysis to those dependent upon the assumption of normality will rarely be using methods as robust and powerful as available alternative nonparametric methods of analysis. Given the prevalence of measures distributed as Laplace, measures distributed as logistic, and measures having varying degrees of asymmetry, methods of analysis capable of handling these variables are crucial for researchers in education and social science fields.

The concern over robustness of parametric tests to departures from parametric assumptions about data distributions has yielded extensive study. Of particular interest in the two-factor design is the test for interaction among the factors. It has been argued that the few proposed nonparametric tests capable of testing for interaction effects are either computationally intensive or less powerful, leaving only the parametric ANOVA *F* test available for researchers to use when testing for interaction effects (Gaito, 1959; Gardner, 1975). To address this problem, McSweeney (1967) proposed a nonparametric test that can be used to test for the presence of both main and interaction effects. The McSweeney test (McSweeney, 1967) uses an alignment data transformation. The goal of aligning is to treat whichever of the two main or interaction effects not currently being tested as nuisance parameters and remove them from the model (Hodges & Lehmann, 1962). When testing for interaction, this means subtracting main effect estimates (means calculated within levels of each main effect) from the observations. This leaves only the interaction effect to be tested. Prior to calculating the McSweeney statistic, the data must be aligned, pooled, and ranked. Using the aligned and ranked data, the McSweeney statistic is calculated using:

$$H_Y = (N-1)\left(\frac{SS_{Effect}}{SS_{Total}}\right) \tag{1}$$

The null hypothesis of no effect for the McSweeney test is rejected when:

$$H_Y > \chi^2_{\alpha, df_{Effect}} \tag{2}$$

Simulation studies of the viability of this test yielded mixed results. For normal, mixed-normal, exponential, double exponential, and chi-square population distributions in combination with small cell sample sizes, the McSweeney test demonstrates strong power at the cost of liberal Type I error (Harwell, 1991; Hornsby Brown, 2023; Toothaker & Newman, 1994). Kelley et al. (1994) explored the viability of the McSweeney test in the 2 x 2 x 2 factorial design for normal, uniform, *t*, and exponential population distributions with sample sizes of n = 7, 21, and 35. Results demonstrate a failure to maintain Type I error rates for null effects in treatment conditions but competitive results when the number of nonnull effects and effect sizes became large. Hornsby Brown (2023) explored using the McSweeney test in the 2 x 2 design for the normal, logistic, and Laplace distributions with varying severities of skew and cell sample sizes of n = 5, 10, 30, and 50. Results demonstrate strong power but liberal Type I error rates for small sample sizes. McSweeney (1967) makes two suggestions that have yet to be researched: (1) derive a new test for interaction based on normal scores or conduct normal scores tests for interaction that retain the advantage of removing by alignment the influence of main effects, (2) modify the proposed McSweeney test to perform a normal scores transformation in place of a rank transformation.

There are three methods for making inferences about group differences using normal score transformations: the Bell-Doksum normal scores test (Bell & Doksum, 1965), the Terry-Hoeffding normal scores test (Hoeffding, 1951; Terry, 1952), and the Van der Waerden normal scores test (Van der Waerden, 1952, 1953, 1956). Fisher and Yates (1949) and Bell and Doksum (1965) proposed using a random normal scores transformation, a procedure where ranks of original scores are replaced by randomly drawn normal deviates with corresponding ranks. Bradley (1968) refined this technique by limiting the variates drawn to those of a standard normal distribution, creating a more powerful form of the test statistic. As the deviates drawn in the Bell-Doksum test are random, two researchers analyzing the same data with this test may arrive at different conclusions.

Hoeffding (1951) and Terry (1952) refined this procedure further to replace ranks of the original observations with expected normal scores. As these expected normal scores are constant values dependent only upon sample size, researchers have calculated tables of expected normal order statistics (Harter, 1961) and tables of appropriate chi-square critical values for total sample sizes less than or equal to 20 (Klotz, 1964). Owen (1962) tabled large sample approximation normal theory critical values. Sawilowsky (1990) refers to the data transformation procedure outlined in the Bell-Doksum test as a random normal scores transformation (RNST) and the procedure outlined in the Terry-Hoeffding test as the expected normal scores transformation (ENST).

Expected normal scores transformations are adaptable to any hypothesis in experimental research designs (Conover, 1980; Gibbons, 1985). Lu and Smith (1979) found the ENST robust and powerful compared to the parametric ANOVA *F* in the one-way design. Sawilowsky (1985, 1989) found both the RNST and ENST to be more powerful and robust than the rank transform test but less powerful and robust than the parametric *F* test in the 2 x 2 x 2 design with small sample sizes (n = 2) for various distributions. Feir-Walsh and Toothaker (1974) concluded that the Terry-Hoeffding normal scores test yielded less power than the ANOVA *F* test and the Kruskal-Wallis test for total sample sizes as large as 200. Wiedermann and Alexandrowicz (2011) demonstrated a modified Terry-Hoeffding test is a robust and powerful normal scores test for two-sample paired data when compared to the *t* test.

Van der Waerden (1952, 1953, 1956) proposed an alternative normal scores transformation. This test uses a rank transformation in combination with the inverse standard normal distribution function. The Van der Waerden normal scores transformation is as follows:

$$z_{ij} = \Phi^{-1}\left(\frac{R(X_{ij})}{N+1}\right) \tag{3}$$

where X_{ij} represents the i^{th} value in the j^{th} group (j = 1, 2, …, k), $R(X_{ij})$ represents the pooled rank of observation X_{ij}, and Φ^{-1} denotes the inverse standard normal distribution function. The chi-square test statistic is calculated as:

$$W = \frac{1}{s^2}\sum_{j=1}^{k} n_j\,\bar{z}_j^2 \tag{4}$$

where n_j represents the sample size for the j^{th} group,

$$s^2 = \frac{1}{N-1}\sum_{j=1}^{k}\sum_{i=1}^{n_j} z_{ij}^2 \tag{5}$$

and

$$\bar{z}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} z_{ij} \tag{6}$$

The null hypothesis that *k* groups yield the same observations is rejected when:

$$W > \chi^2_{\alpha,k-1} \tag{7}$$

McSweeney (1967) studied the fit of the asymptotic chi-square distribution of the normal scores test statistic in the presence of small sample sizes for the one-way c-sample design, concluding that both the Terry-Hoeffding and Van der Waerden normal scores test statistics with small sample sizes are well-approximated by the chi-square distribution with k – 1 degrees of freedom. Transforming original scores to normal scores for conducting a one-way two- or more-sample test of differences in central tendency can result in equivalent or larger power than traditional parametric *t* and *F* tests, the Wilcoxon rank sum test (also called the Mann-Whitney *U* test), and the Kruskal-Wallis test while still controlling the Type I error rate at or near the nominal level (Hodges & Lehmann, 1961; Keselman & Toothaker, 1973; Kruskal & Wallis, 1952; Mann & Whitney, 1947; McSweeney, 1967; Penfield, 1994; Penfield & McSweeney, 1968; Thompson et al., 1966; Van der Laan, 1964; Van der Laan & Oosterhoff, 1965, 1967; Wilcoxon, 1945, 1947). This power increase is shown to be larger for large- or heavy-tailed distributions (Curtis & Marascuilo, 1992; Lu & Smith, 1979; Padmanabhan, 1977). Zimmerman (1996) demonstrated via simulation that rank transformations, including the Van der

Waerden normal scores transformation, can be used to reduce group variance heterogeneity in two-sample tests of location difference.

Minimal research has been performed in modifying the McSweeney test by using a normal scores data transformation in place of the rank transformation after the alignment procedure or in using the aligned normal scores data transformation to test for interaction effects. Van der Laan (1964) hypothesizes that normal scores tests might perform better than rank transformation tests for samples from asymmetric distributions. Weber (1972) explores replacing the rank transformation data step in the McSweeney test with a Terry-Hoeffding normal scores data transformation. Results of the test outlined by Weber indicate power competitive to the ANOVA *F* test when testing for interaction effects in the 2 x 3 design with n = 10 cell sample sizes but Type I error rates slightly larger than nominal α levels. No research has been performed in modifying the McSweeney test by using a Van der Waerden normal scores data transformation in place of the rank transformation data step. This research aims to address this gap by conducting a study that explores using the alignment data transformation in combination with the Van der Waerden normal scores transformation while using the analysis of variance *F* and the McSweeney H_y test statistics to test for interaction effects for skewed and asymmetrical distributions.

Methods

Monte Carlo simulation methods are employed to compare Type I error rates and power properties of the following four test procedures for two-factor main and interaction effects: (1) the ANOVA *F* test , (2) a Van der Waerden normal scores transformation tested with an *F* statistic (V-*F*), (3) an alignment data transformation in combination with a Van der Waerden normal scores transformation tested with an *F* statistic (A-V-*F*), and (4) an alignment transformation in combination with a Van der Waerden normal scores transformation using the McSweeney H_Y statistic (A-V-M). It should be noted here that method 4 (A-V-M), an alignment transformation in combination with a Van der Waerden normal scores transformation tested by the McSweeney statistic, is an equivalent test to performing a Van der Waerden normal scores transformation with use of a chi-square test statistic subsequent to performing an alignment data transformation. Usage of the parametric analysis of variance *F* test assumes the residuals are normally distributed, an assumption Micceri (1989) found violated in all observed cases in education and social sciences measures. Thus, simulation of residuals in this study employs distributions that Micceri observed as frequently occurring in education and social science measures in his meta study, including the Laplace distribution, the logistic distribution, and distributions having mild to severe measures of asymmetry.

Measures of skew and asymmetry lying several standard deviations above the main body of the distribution are common in psychometric and achievement measures (Micceri, 1989; Walberg et al., 1984). Of the 440 measures observed by Micceri (1989), 49.1 percent had extreme to exponential tail weights and 30.9 percent had extreme to exponential asymmetry (note, the exponential distribution has a skewness of 2). Additionally, when using small sample sizes, the value of skewness present in the sample may increase due to potential sample-population mismatch (Tipton et al., 2017). Thus, in comparing the varying tests in the presence of increasingly severe measures of asymmetry, each of three different skewed distributions in combination with four different levels of the skew parameter is used to simulate the residuals. A skew-normal distribution is used with mean of 0, a variance of 1, and skew values of 0, 1, 2, and 3. A skewed logistic distribution is used with a mean of 0, a variance of 1, and skew values of 1, 2, 3, and 4. An asymmetric Laplace distribution is used with a mean of 0, a variance of 1, and skew values of 1, 0.5, 0.33, and 0.25. A skew-normal distribution with a skew value of 0, a skew logistic with a skew of 1, and an asymmetric Laplace with an asymmetric value of 1 are symmetric distributions. As the skew parameters are altered, the skew becomes increasingly larger to the right.

These twelve distributions are crossed with seven levels of cell sample size (5, 10, 30, 50, 100, 200, and 500) and all possible combinations of the presence or absence of each main and interaction effect in the two-factor design. It is common for researchers who are considering nonparametric methods to select a method based on asymptotic relative efficiency, though this ratio assumes large sample sizes and is not realistically representative of research study conditions faced by researchers in educational and social science fields. Thus, the total sample sizes selected for this study, ranging from N = 20 to N = 2000, are more representative of these common research conditions. Additionally, using a range of smaller to larger cell sample sizes will reveal power and robustness properties that occur as a result of potential sample-population mismatch that can occur by random chance in small sample sizes (Tipton et al., 2017). The nonnull effect size is chosen for each cell sample size such that the power to detect each effect would be 80% under the standard normal theory, that is, for n = 5 the nonnull effect size simulated is 0.67, for n = 10 the nonnull effect size is 0.46, for n = 30 the nonnull effect size is 0.26, for n = 50 the nonnull effect size is 0.20, for n = 100 the nonnull effect size is 0.14, for n = 200 the nonnull effect size is 0.10, and for n = 500 the nonnull effect size is 0.06.

Because the designs simulated are two-way designs with two levels of each main effect, these effect sizes result in 80% power under the standard normal theory for their respective cell sample sizes for both main and interaction effects given each effect has 1 degree of freedom. This raw effect size is added or subtracted from each cell based on level of each main and interaction effect. As normality with a mean of 0 and a variance of 1 is assumed in generation of the effect sizes, these raw effect sizes are standardized effect sizes for the normal distribution as well. A nominal α of 0.05 is used. In total, this creates 672 configurations of residual distributions, sample sizes, and effect size combinations that are simulated with 5,000 replications to obtain the Type I error rate and power. The R programming language (R Core Team, 2021) was used in conjunction with the 'sn', 'glogis', and 'LaplacesDemon' packages to simulate the data. Table 1 summarizes the design factors used.

Table 4.1
Design factors and levels of each design factor used in simulation

Design Factor:	Levels:
Residual distributions	3
Severities of skew for each residual distribution	4
Cell sample size	7
Null and nonnull main and interaction effect combinations	8

Using 5,000 simulations creates a 95% margin of error of 0.006 for the Type I error estimates and a 95% margin of error of 0.011 for the power estimates under standard normal theory. The 95% margin of error of 0.006 for Type I error estimates is close to the stringent error band interval of 0.005 proposed by Bradley (1978) for examining the robustness of Monte Carlo simulated Type I error rates for hypothesis testing. This error band interval becomes ±0.005 when a nominal α of 0.05 is used.

Bradley also proposed a second more liberal error band that becomes ±0.02 when a nominal α of 0.05 is used, and that is the band adopted in this study to indicate whether a method is robust. That is, any proposed hypothesis test with an empirical Type I error rate greater than 0.07 is deemed as not being robust to violations of symmetry or adequate sample size assumptions for that simulated design. Thus, power comparisons for the proposed hypothesis tests will only be considered for those tests maintaining an empirical Type I error rate that is less than or equal to 0.07 for each simulated design. Proposed testing methods will be called slightly conservative or slightly liberal if they maintain an empirical Type I error rate which deviates less than 0.01 below and above 0.05, respectively. Methods will be called conservative or liberal if their Type I error rate deviates 0.02 or more from 0.05. Both main and interaction effects are tested and recorded, yet the primary empirical measures of interest refer to performance of a test for detecting an interaction. Results are provided in table format for all simulated designs.

Results

These results contain the comparison of the testing methods ability to detect an interaction. Restated, the testing methods are: (1) the ANOVA *F* test , (2) a Van der Waerden normal scores transformation tested with an *F* statistic (V-*F*), (3) an alignment data transformation in combination with a Van der Waerden normal scores transformation tested with an *F* statistic (A-V-*F*), and (4) an alignment transformation in combination with a Van der Waerden normal scores transformation using the McSweeney H_Y statistic (A-V-M). The focus is on the empirical Type I error rate and the power results when testing for interaction in the presence and absence of main effects for all three distributions.

Skew-normal Distribution

No method is rejected for Type I error rates being too liberal for any combination of significant main effects, skew severity, and cell sample size. While no Type I error rates for any method exceed the liberal error band of 0.07, Type I error rates often fall between 0.05 and 0.07 for smaller cell sample sizes. For the n = 5 cell sample size, methods V-*F*, A-V-*F*, and the ANOVA *F* test often result in similar Type I error rates with method A-V-M having a Type I error rate around 1 percentage point larger. Hornsby Brown (2023) notes strong conservative results of the rank transform test when testing for an interaction effect in the presence of two nonnull main effects. Replacing the rank transformation with a Van der Waerden normal scores transformation (method V-*F*) does not mitigate this problematic conservatism sufficiently enough to consider this a viable test as shown in tables 2 and 3. While method V-*F* is not viable for this reason, favorable results are noted for methods A-V-*F* and A-V-M. As noted in tables 2, 3, 4, and 5, method A-V-*F* maintains Type I error rates comparable to the ANOVA *F* test and equal to larger power than the ANOVA *F* test for all levels of skew and cell sample sizes. At n = 10 or larger cell sample sizes, method A-V-M maintains Type I error rates less than 0.06 levels for all combinations of significant main effects, cell sample sizes, and severity of skew. Method A-V-M also maintains competitive power to the ANOVA *F* and method A-V-*F* for all levels of skew severity.

Table 4.2

Type I Error and Power Results for Skew-normal with two nonnull main effects and a skew parameter of 0

Cell Size:	5	10	30	50	100	200	500
Method:	Null interaction effect						
ANOVA	0.0494	0.0488	0.0436	0.0522	0.0494	0.051	0.052
V-*F*	0.0206	0.0418	0.042	0.0514	0.0498	0.0498	0.0514
A-V-*F*	0.0494	0.0504	0.0472	0.0534	0.0488	0.0494	0.052
A-V-M	0.0602	0.0548	0.0486	0.054	0.0494	0.0494	0.0522
	Nonnull interaction effect						
ANOVA	0.7946	0.791	0.7918	0.7982	0.8088	0.791	0.8076
V-*F*	0.6278	0.756	0.7844	0.7938	0.8058	0.7912	0.8066
A-V-*F*	0.7806	0.784	0.787	0.7948	0.8074	0.7912	0.8068
A-V-M	0.8064	0.798	0.7924	0.7986	0.8084	0.7914	0.807

Table 4.3

Type I Error and Power Results for Skew-normal with two nonnull main effects and a skew parameter of 3

Cell Size:	5	10	30	50	100	200	500
Method:	Null interaction effect						
ANOVA	0.0484	0.0516	0.0502	0.0542	0.052	0.0528	0.052
V-*F*	0.0028	0.0268	0.0502	0.0592	0.0538	0.0548	0.0528
A-V-*F*	0.0512	0.0514	0.052	0.055	0.0512	0.0534	0.051
A-V-M	0.0596	0.0564	0.0532	0.0556	0.0512	0.0538	0.0512
	Nonnull interaction effect						
ANOVA	0.9868	0.9902	0.9892	0.9886	0.9886	0.991	0.989
V-*F*	0.7396	0.9516	0.9828	0.9868	0.9886	0.9934	0.9944
A-V-*F*	0.983	0.9896	0.9928	0.9938	0.9952	0.9962	0.9954
A-V-M	0.9868	0.9914	0.9928	0.994	0.9952	0.9962	0.9954

Table 4.4

Type I Error and Power Results for Skew-normal with one null and one nonnull main effect and a skew parameter of 3

Cell Size:	5	10	30	50	100	200	500
Method:			Null interaction effect				
ANOVA	0.0524	0.0498	0.048	0.054	0.0574	0.0496	0.0496
V-*F*	0.0516	0.0488	0.0492	0.0522	0.0578	0.0504	0.0492
A-V-*F*	0.052	0.0498	0.048	0.0514	0.0584	0.0506	0.0488
A-V-M	0.0614	0.0546	0.049	0.0528	0.0586	0.0506	0.0488
			Nonnull interaction effect				
ANOVA	0.985	0.9878	0.9906	0.9888	0.9898	0.9908	0.9892
V-*F*	0.9818	0.987	0.9956	0.993	0.9958	0.9952	0.995
A-V-*F*	0.984	0.988	0.9958	0.9942	0.9956	0.9956	0.9948
A-V-M	0.9884	0.9898	0.996	0.9942	0.9958	0.9956	0.9948

Table 4.5

Type I Error and Power Results for Skew-normal with two null main effects and a skew parameter of 3

Cell Size:	5	10	30	50	100	200	500
Method:			Null interaction effect				
ANOVA	0.0446	0.047	0.0498	0.051	0.0528	0.0464	0.0484
V-*F*	0.0476	0.049	0.0492	0.0542	0.052	0.0468	0.0474
A-V-*F*	0.0464	0.0496	0.0502	0.053	0.0528	0.0462	0.0468
A-V-M	0.0558	0.0542	0.0522	0.0536	0.0534	0.0462	0.0468
			Nonnull interaction effect				
ANOVA	0.992	0.9872	0.9888	0.9906	0.9912	0.9926	0.9894
V-*F*	0.989	0.988	0.9936	0.9946	0.995	0.9974	0.996
A-V-*F*	0.9858	0.987	0.993	0.9942	0.9944	0.9972	0.996
A-V-M	0.9898	0.9874	0.9936	0.9942	0.9946	0.9972	0.9962

Skew Logistic Distribution

There is no test rejected for the skewed logistic design for Type I error rates being too liberal. There is no combination of skew severity, significant main effects, or cell sample sizes for which Type I error rate for any method exceeds 0.06. As shown in tables 6 and 7, method V-*F* performs conservatively in the presence of two nonnull main effects for the logistic distribution as well leading to less power than other methods including the

ANOVA *F* test in the presence of small cell sample sizes. This result further supports the conclusion that method V-*F* should be rejected as a viable test. As shown in tables 6, 7, 8, and 9, methods A-V-*F* and A-V-M result in larger power than the ANOVA *F* test for all combinations of significant main effects, cell sample sizes, and skew severities. The tabled Type I error rate results also indicate that methods A-V-*F* and A-V-M are robust for the skew logistic distribution.

Table 4.6

Type I Error and Power Results for Skew Logistic with two nonnull main effects and a skew parameter of 1

Cell Size:	5	10	30	50	100	200	500
Method:				Null interaction effect			
ANOVA	0.0502	0.056	0.0506	0.0476	0.0448	0.0486	0.0506
V-*F*	0.044	0.056	0.0512	0.0468	0.0464	0.0488	0.0496
A-V-*F*	0.0494	0.053	0.0494	0.0478	0.046	0.0498	0.0496
A-V-M	0.0598	0.0588	0.0506	0.0486	0.046	0.05	0.0498
				Nonnull interaction effect			
ANOVA	0.3476	0.3394	0.3384	0.3338	0.3488	0.3286	0.341
V-*F*	0.329	0.3442	0.3516	0.3474	0.3634	0.3462	0.3526
A-V-*F*	0.3512	0.346	0.3562	0.3482	0.3662	0.3442	0.353
A-V-M	0.3812	0.3616	0.3606	0.3524	0.3672	0.345	0.3532

Table 4.7

Type I Error and Power Results for Skew Logistic with two nonnull main effects and a skew parameter of 4

Cell Size:	5	10	30	50	100	200	500
Method:	Null interaction effect						
ANOVA	0.0464	0.0518	0.0516	0.053	0.0516	0.0464	0.0462
V-*F*	0.0326	0.0502	0.0496	0.0548	0.0534	0.0474	0.0462
A-V-*F*	0.0446	0.0524	0.051	0.053	0.0514	0.0484	0.0472
A-V-M	0.055	0.0568	0.052	0.0532	0.0522	0.0486	0.0472
	Nonnull interaction effect						
ANOVA	0.5354	0.5272	0.5242	0.5308	0.5268	0.5208	0.5304
V-*F*	0.446	0.5116	0.5458	0.564	0.571	0.5716	0.5862
A-V-*F*	0.5412	0.5612	0.5786	0.5898	0.5868	0.5864	0.5946
A-V-M	0.5756	0.577	0.5838	0.5912	0.589	0.5878	0.5948

Table 4.8

Type I Error and Power Results for Skew Logistic with one null and one nonnull main effect and a skew parameter of 4

Cell Size:	5	10	30	50	100	200	500
Method:	Null interaction effect						
ANOVA	0.0448	0.0518	0.0484	0.0442	0.0504	0.0564	0.047
V-*F*	0.0482	0.049	0.0484	0.0426	0.0488	0.0566	0.0462
A-V-*F*	0.0434	0.0492	0.0476	0.0438	0.0476	0.0562	0.0466
A-V-M	0.0534	0.0548	0.049	0.0448	0.048	0.0562	0.0466
	Nonnull interaction effect						
ANOVA	0.546	0.5384	0.5224	0.5422	0.5126	0.5232	0.5326
V-*F*	0.5478	0.5696	0.5842	0.6026	0.5804	0.5882	0.5924
A-V-*F*	0.555	0.5648	0.5838	0.5966	0.5812	0.589	0.5926
A-V-M	0.5924	0.5808	0.5892	0.5994	0.5824	0.5894	0.5932

Table 4.9

Type I Error and Power Results for Skew Logistic with two null main effects and a skew parameter of 4

Cell Size:	5	10	30	50	100	200	500
Method:				Null interaction effect			
ANOVA	0.049	0.0488	0.0536	0.0488	0.053	0.0446	0.049
V-*F*	0.0524	0.0504	0.0526	0.049	0.0528	0.0444	0.0486
A-V-*F*	0.0504	0.0516	0.0528	0.0506	0.053	0.0454	0.0488
A-V-M	0.0598	0.0556	0.0536	0.0512	0.0534	0.0456	0.0488
				Nonnull interaction effect			
ANOVA	0.5336	0.5366	0.5348	0.5372	0.5276	0.531	0.5278
V-*F*	0.5554	0.5818	0.588	0.5962	0.5908	0.5922	0.59
A-V-*F*	0.54	0.5708	0.583	0.5954	0.5896	0.5922	0.5884
A-V-M	0.5794	0.5858	0.587	0.598	0.5902	0.5928	0.5884

Asymmetric Laplace Distribution

Once again, no method is rejected for having Type I error rates outside the liberal error band of 0.07. Methods A-V-*F* and A-V-M have strong Type I error rate and power results across all asymmetry severities, cell sample size, and significant main effect combinations for the asymmetric Laplace distribution. As shown in tables 10, 11, 12, 13, and 14, the Type I error rates of all methods remain consistently below 0.06. Furthermore, methods A-V-*F* and A-V-M consistently result in considerably larger power than the ANOVA *F* test. The tabled Type I error rate and power results indicate that methods A-V-*F* and A-V-M are more robust for the asymmetric Laplace distribution than the ANOVA *F* test. Method V-*F* shows similar conservatism for small cell sample sizes in the presence of two nonnull main effects but both Type I error and power results converge to A-V-*F* and A-V-M method results as cell sample size increases.

Table 4.10

Type I Error and Power Results for Asymmetric Laplace with two null main effects and an asymmetry parameter of 1

Cell Size:	5	10	30	50	100	200	500
Method:	Null interaction effect						
ANOVA	0.0456	0.0484	0.0524	0.0474	0.0506	0.046	0.0508
V-*F*	0.0512	0.0478	0.054	0.0488	0.0486	0.046	0.0512
A-V-*F*	0.0494	0.0506	0.0522	0.0488	0.0492	0.0456	0.0516
A-V-M	0.0574	0.0538	0.0534	0.0492	0.0494	0.0462	0.0516
	Nonnull interaction effect						
ANOVA	0.8024	0.7924	0.805	0.8016	0.8076	0.8122	0.8028
V-*F*	0.8184	0.8336	0.875	0.8784	0.8848	0.8892	0.8926
A-V-*F*	0.8028	0.8278	0.8712	0.8738	0.8828	0.8884	0.8922
A-V-M	0.8308	0.8374	0.8736	0.8754	0.8828	0.8886	0.8922

Table 4.11

Type I Error and Power Results for Asymmetric Laplace with two nonnull main effects and an asymmetry parameter of 1

Cell Size:	5	10	30	50	100	200	500
Method:	Null interaction effect						
ANOVA	0.0454	0.041	0.0534	0.0486	0.049	0.0476	0.05
V-*F*	0.0184	0.0362	0.0542	0.0478	0.0496	0.0484	0.051
A-V-*F*	0.046	0.043	0.0544	0.0472	0.0498	0.0468	0.0522
A-V-M	0.0562	0.0476	0.0564	0.0484	0.05	0.047	0.0522
	Nonnull interaction effect						
ANOVA	0.8066	0.7964	0.8014	0.8078	0.8002	0.797	0.8038
V-*F*	0.619	0.7606	0.8412	0.8586	0.8656	0.8684	0.8808
A-V-*F*	0.8102	0.8308	0.8644	0.8786	0.8724	0.8764	0.8834
A-V-M	0.8332	0.843	0.8674	0.8796	0.8728	0.8768	0.884

Table 4.12

Type I Error and Power Results for Asymmetric Laplace with two null main effects and an asymmetry parameter of 1/4

Cell Size:	5	10	30	50	100	200	500
Method:	Null interaction effect						
ANOVA	0.0444	0.0498	0.0476	0.047	0.054	0.0506	0.0516
V-*F*	0.052	0.049	0.046	0.0484	0.0524	0.0522	0.0516
A-V-*F*	0.045	0.0466	0.0426	0.046	0.0488	0.0476	0.0508
A-V-M	0.0536	0.049	0.043	0.0458	0.0484	0.0478	0.0506
	Nonnull interaction effect						
ANOVA	0.1998	0.1874	0.1758	0.1708	0.165	0.1762	0.1734
V-*F*	0.3058	0.3422	0.4132	0.447	0.4564	0.48	0.5086
A-V-*F*	0.2498	0.2866	0.3564	0.3966	0.42	0.455	0.4914
A-V-M	0.2732	0.2932	0.3572	0.3976	0.4198	0.4546	0.4914

Table 4.13

Type I Error and Power Results for Asymmetric Laplace with one null and one nonnull main effect and an asymmetry parameter of 1/4

Cell Size:	5	10	30	50	100	200	500
Method:	Null interaction effect						
ANOVA	0.0478	0.0478	0.0514	0.0484	0.05	0.0488	0.049
V-*F*	0.0586	0.0516	0.0524	0.0482	0.0508	0.0488	0.0488
A-V-*F*	0.0514	0.0444	0.0462	0.0474	0.0472	0.0454	0.0494
A-V-M	0.0592	0.046	0.0462	0.0476	0.0472	0.0452	0.0492
	Nonnull interaction effect						
ANOVA	0.2014	0.1786	0.1702	0.1704	0.1692	0.164	0.163
V-*F*	0.2728	0.3046	0.3882	0.404	0.4534	0.4662	0.4822
A-V-*F*	0.2432	0.276	0.3598	0.3838	0.4336	0.4554	0.4764
A-V-M	0.2684	0.285	0.3604	0.3836	0.4334	0.4552	0.4762

Table 4.14

Type I Error and Power Results for Asymmetric Laplace with two nonnull main effects and an asymmetry parameter of 1/4

Cell Size:	5	10	30	50	100	200	500
Method:				Null interaction effect			
ANOVA	0.0524	0.0548	0.0506	0.046	0.048	0.0496	0.0486
V-*F*	0.0548	0.0546	0.051	0.0528	0.0502	0.0522	0.0506
A-V-*F*	0.0514	0.0512	0.0452	0.0464	0.0468	0.0498	0.0514
A-V-M	0.0588	0.0528	0.0452	0.0464	0.047	0.05	0.0514
				Nonnull interaction effect			
ANOVA	0.1998	0.187	0.1804	0.1696	0.1628	0.1724	0.1674
V-*F*	0.2026	0.2406	0.3134	0.3406	0.3732	0.4058	0.4484
A-V-*F*	0.2406	0.2834	0.3662	0.3902	0.4274	0.4562	0.483
A-V-M	0.2628	0.2926	0.368	0.3912	0.4272	0.4556	0.4828

Discussion

The purpose of this study is to explore the usage of Van der Waerden normal scores data transformation in combination with the rank transform, aligned rank transform, and McSweeney nonparametric tests. Methods A-V-*F* and A-V-M maintained Type I error rate for all simulated designs indicating these tests are robust to Type I error rate even when used in combinations with small sample sizes. Method A-V-*F* uses an alignment procedure in combination with a Van der Waerden normal scores data transformation prior to referring the transformed data to the ANOVA *F* statistic. Method A-V-M uses an alignment procedure in combination with a Van der Waerden normal scores data transformation prior to referring the transformed data to the McSweeney statistic. Both methods A-V-*F* and A-V-M occasionally demonstrated mildly liberal Type I error rates for small (n = 5 and n = 10) cell sample sizes. However, using the Type I error rate cutoff bands outlined by Bradley (1978), both methods remain viable tests for all cell sample sizes. Comparing these results to those of Hornsby Brown (2023), replacing the rank transformation with a Van der Waerden normal scores transformation does improve small sample size Type I error rates for the McSweeney test.

Power results likewise remained consistent or improve when using a Van der Waerden normal scores transformation in combination with the McSweeney statistic. Methods A-V-*F* and A-V-M are consistently as or more powerful than the ANOVA *F* test for all simulated distributions. Notably, for the asymmetric Laplace distribution, using a Van der Waerden normal scores transformation with the McSweeney test provided for larger power increases relative to the ANOVA *F* test than the base McSweeney test does as cell sample size increases based on the results of Hornsby Brown (2023). Based on these results, it is safe to conclude that using an alignment procedure in combination with a Van der Waerden normal scores data transformation provides for a more powerful test with both the ANOVA *F* statistic and McSweeney H_y statistic compared to the ANOVA *F* test. Both of these testing methods provide for a more powerful and more robust test than the ANOVA *F* test.

The tabled results demonstrate the power loss consequence of incorrectly assuming the normality assumption is satisfied when analyzing data. Effect size for each cell sample size is chosen to yield 80% power under standard normal theory. It is evident from the tabled results that the ANOVA *F* test suffers a more significant power reduction due to lack of robustness than the two viable nonparametric testing methods. Results demonstrate using the nonparametric methods can mitigate power loss by as much as 6 percentage points for the logistic distribution and 31 percentage points for the Laplace distribution with larger power mitigation occurring as skew parameters increase. For researchers in the behavioral and social sciences, fields where Micceri (1989) notes virtually no measures are normally distributed, the recommended nonparametric tests provide increased power in testing for both main and interaction effects than when using the parametric analysis of variance *F* test while making no assumption regarding residual distribution.

The results from this study demonstrate that using a Van der Warden normal scores data transformation in combination with the McSweeney test provides for a more powerful and more robust test than the ANOVA *F* test. The provided alternative nonparametric tests appear to be risk-free alternatives to the parametric ANOVA *F* test that are just as powerful when normality assumptions are satisfied and more powerful when normality assumptions are not satisfied for the simulated distributions. This broad result is beneficial for researchers situated in the behavioral and social sciences in need of powerful and robust statistical inference methods for the factorial two- or more- way design. Further research is needed to ensure these favorable results remain consistent in the presence of other common distributions, heteroskedasticity, unequal group sample sizes, and for more complex designs.

This research employs simulations of residuals following only normal, logistic, and Laplace distributions with varying severities of skew. I studied the nonparametric tests based on ranks using four levels of skew severity for each of the three mentioned distributions. This study includes only a two-factor design with two levels of each main effect. This line of research naturally reveals several subsequent lines of inquiry. As Micceri (1989) acknowledges, there should also be acknowledgement in the field of psychometric measures of the prevalence of multimodality, lumpiness, and other frequently occurring deviations from normality in population distributions apart from skew. Additionally, further study is needed regarding the suggested nonparametric tests in the presence of designs having more than two levels of each main effect, having unbalanced cell sample sizes, and in designs where cell residuals have different distributions including but not limited to different severities of skew. Given the results demonstrated in this study and the prevalence of nonparametric techniques introduced, a proportionate increase in efficacy comparisons for these methods is needed to ensure that robust testing methods which require fewer distributional assumptions than traditional parametric tests do not fall through the cracks and their potential left untapped.

References

Bell, C. B., & Doksum, K. A. (1965). Some new distribution-free statistics. *Annals of Mathematical Statistics*, *36*(1), 203-214.

Blair, R. C. (1980). *A comparison of the power of the two independent means t test to that of the wilcoxon's rank-sum test for samples of various populations* [Doctoral dissertation, University of South Florida).

Bradley, J. V. (1968). *Distribution-free statistical tests*. Prentice-Hall.

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician*, *31*(4), 147–150. https://doi.org/10.2307/2683535

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

Bradley, J. V. (1979). A nonparametric test for interactions of any order. *Journal of Quality Technology, 11*(4), 177–184. https://doi.org/10.1080/00224065.1979.11980909

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

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

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

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

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). John Wiley.

Curtis, D. A., & Marascuilo, L. A. (1992). Point estimates and confidence intervals for the parameters of the two-sample and matched-pair combined tests for ranks and normal scores. *The Journal of Experimental Education*, *60*(3), 243–269.

DeWees, T. A., Mazza, G. L., Golafshar, M. A., & Dueck, A. C. (2020). Investigation into the effects of using normal distribution theory methodology for Likert scale patient-reported outcome data from varying underlying distributions including floor/ceiling effects. *Value in Health*, *23*(5), 625–631.

Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores, and Kruskal-Wallis test under violations of assumptions. *Educational and Psychological Measurement*, *34*, 789–799.

Feng, Y., Hancock, G. R., & Harring, J. R. (2019). Latent growth models with floor, ceilings, and random knots. *Multivariate Behavioral Research*, *54*(5), 751–770.

Fisher, R. A., & Yates, F. (1949). *Statistical tables for biological, agricultural, and medical research* (3rd ed.). Hafner.

Gaito, J. (1959). Non-parametric methods in psychological research. *Psychological Reports*, *5*, 115-125.

Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research*, *45*(1), 43-57.

Gibbons, J. D. (1985). *Nonparametric Statistical Inference* (2nd ed.). Marcel Dekker.

Harter, H. L. (1961). Expected values of normal order statistics. *Biometrika*, *48*, 151–165.

Harwell, M. R. (1991). Completely randomized factorial analysis of variance using ranks. *British Journal of Mathematical and Statistical Psychology*, *44*(2), 383–401. https://doi.org/10.1111/j.2044-8317.1991.tb00970.x

Hellstrom, A. (1993). The normal distribution in scaling subjective stimulus difference: Less “normal” than we think? *Perception & Psychophysics*, *54*(1), 82–92.

Hodges, J. L., & Lehmann, E. L. (1961). Comparison of the Normal Scores and Wilcoxon Tests. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1). Berkeley; University of California Press.

Hodges, J. L., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics*, *33*(2), 482–497. https://doi.org/10.1214/aoms/1177704575

Hoeffding, W. (1951). Berkeley Symposium on Mathematics, Statistics, and Probability. In *Optimum nonparametric statistics* (pp. 83–92). Berkeley, CA; University of California Press.

Hornsby Brown, M. E. (2023). *Nonparametric Tests of Interaction for the Two-Way Design with Skewed Distributions*. [Manuscript submitted for publication].

Kelley, D. L., Sawilowsky, S. S., & Blair, R. C. (1994). Midwestern Educational Research Association. In *Comparison of ANOVA, McSweeney, Bradley, Harwell-Serlin, and Blair-Sawilowsky tests in the balanced 2x2x2 layout.* Chicago, IL.

Keselman, H. J., & Toothaker, L. E. (1973). 81st Annual Convention. In *An Empirical Comparison of the Marascuilo and Normal Scores Nonparametric Tests and the Scheffe and Tukey Parametric Tests for Pairwise Comparisons* (pp. 15–16). APA.

Klotz, J. H. (1964). On the normal scores two-sample rank test. *Journal of the American Statistical Association*, *59*, 652–664.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Liu, Q., & Wang, L. (2021). T-test and ANOVA for data with ceiling and/or floor effects. *Behavior Research Methods*, *53*(1), 264–277.

Lu, H. T., & Smith, P. J. (1979). Distribution of the normal scores statistic for nonparametric one-way analysis of variance. *Journal of the American Statistical Association*, *74*(367), 715–722. https://doi.org/10.1080/01621459.1979.10481676

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit model. *Gifted Child Quarterly, 54*(4), 314–320.

McSweeney, M. T. (1967). An empirical study of two proposed nonparametric tests [Doctoral dissertation, University of California, Berkeley].

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

Noortgate, W. V. den, Boeck, P. D., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational & Behavioral Statistics*, *28*(4), 369–386.

Owen, D. B. (1962). *Handbook of statistical tables*. Addison-Wesley.

Padmanabhan, A. R. (1977). A comparison of the efficiencies of the c-sample normal scores and the Kruskal-Wallis tests in the case of grouped data. *The British Journal of Mathematical & Statistical Psychology*, *30*(2), 222–226.

Penfield, D. A. (1994). Choosing a two-sample location test. *The Journal of Experimental Education*, *62*(4), 343–360. https://doi.org/10.1080/00220973.1994.9944139

Penfield, D. A., & McSweeney, M. T. (1968). The normal scores test for the two-sample problem. *Psychological Bulletin*, *69*(3), 183–191.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Savalei, V. (2006). Logistic approximation to the normal: the KL rationale. *Psychometrika*, *71*(4), 763–767.

Sawilowsky, S. S. (1985). *Robust and power analysis of the 2 x 2 x 2 ANOVA, rank transformation, random normal scores, and expected normal scores transformation tests*. [Doctoral dissertation, University of South Florida].

Sawilowsky, S. S. (1989). *Rank transform: The bridge is falling down*. American Educational Research Association. San Francisco.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60*(1), 91–126. https://doi.org/10.3102/00346543060001091

Still, A. W., & White, A. P. (1981). The approximate randomization test as an alternative to the F test in analysis of variance. *British Journal of Mathematical and Statistical Psychology*, *34*(2), 243–252. https://doi.org/10.1111/j.2044-8317.1981.tb00634.x

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

Terry, M. E. (1952). Some rank order tests which are most powerful against specific parametric alternatives. *The Annals of Mathematical Statistics, 23*(3), 346–366. https://doi.org/10.1214/aoms/1177729381

Thompson, R., Doksum, K. A., & Govindarajulu, Z. (1966). One-sample normal scores test, distribution and power. *Review of the International Statistics Institute*, *34*, 24–26.

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2017). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review*, *41*(5), 472-505.

Toothaker, L. E., & Newman, D. (1994). Nonparametric competitors to the two-way ANOVA. *Journal of Educational and Behavioral Statistics*, *19*(3), 237–273. https://doi.org/10.2307/1165296

Van der Laan, P. (1964). Exact power of some rank tests. *Publications de l'Instituit de Statistique de l'Universite de Paris*, *13*, 211–234.

Van der Laan, P., & Oosterhoff, J. (1965). Monte Carlo estimation of the powers of the distribution-free two sample tests of Wilcoxon, Van der Waerden and Terry and Comparison of these powers. *Statistica Neerlandica*, *19*, 265–275.

Van der Laan, P., & Oosterhoff, J. (1967). Experimental determination of the power functions of the two sample rank tests of Wilcoxon, Van der Waerden, and Terry by Monte Carlo techniques – I. normal parent distributions. *Statistica Neerlandica, 21*, 55-68.

Van der Waerden, B. L. (1952). Order tests for the two-sample problem and their power. *Indagationes Mathematicae*, *14*, 453-458.

Van der Waerden, B. L. (1953). Order tests for the two-sample problem. *Indagationes Mathematicae*, *15*, 303–316.

Van der Waerden, B. L. (1956). The computation of the X-distribution. *Proceedings of the third Berkeley symposium on mathematical statistics and probability*. Berkeley; University of California Press.

Walberg, J. H., Strykowski, B. F., Rovai, E., & Hurg, S. S. (1984). Exceptional performance. *Review of Educational Research*, *54*, 87–112.

Weber, J. M. (1972). The heuristic explication of a large-sample normal scores test for interaction. *British Journal of Mathematical and Statistical Psychology*, *25*, 246–256.

Wiedermann, W. T., & Alexandrowicz, R. W. (2011). A modified normal scores test for paired data. *Methodology*, *7*(1), 25–38.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–82. https://doi.org/10.2307/3001968

Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics*, *3*, 119–122.

Zimmerman, D. W. (1996). A note on homogeneity of variance of scores and ranks. *The Journal of Experimental Education*, *64*(4), 351–362. https://doi.org/10.1080/00220973.1996.10806603

CHAPTER 5

DISCUSSION

Researchers in the behavioral and social sciences are often faced with data which does not satisfy the traditional assumptions of common parametric methods of analysis. Micceri's (1989) meta-analysis analyzed 440 large sample educational and psychological measures and found all to be significantly nonnormal via some class of contamination. Of the 440 measures analyzed by Micceri, only 15.2% had both tail weights at or about normal, 50.2% had at least one tail heavier than normal, and 14.8% had both tail weights less than normal. Of the remainder, 3.2% were concluded as uniform and 16.6% were concluded as Laplace. Additionally, of the 440 measures, 40.7% were concluded as moderately asymmetric, 19.5% as extremely asymmetric, and 11.4% as having exponential asymmetry. Of the categories for the measures analyzed by Micceri, nonnormal tail weights and asymmetry is demonstrated most prevalent in measures of criterion mastery, where 60% of measures were found to be distributed as Laplace, 37.1% of measures had extreme asymmetry, and 57.1% of measures had exponential asymmetry. Thus, Micceri suggests the existence of normally distributed data in educational and psychological measures is as improbable as the existence of a unicorn. Micceri's findings not only directly conclude the prevalence of the Laplace distribution in education and social science measures but also provide evidence of the prevalence of the logistic distribution by highlighting the frequency of mixed-normal measures and criterion mastery measures. The logistic distribution is often used to approximate mixed-normal distributions, model distributions of subjective quantitative measures, and is prevalent in item response theory analysis (Hellstrom, 1993; Noortgate et al., 2003; Savalei, 2006). This would imply that researchers in these fields who limit their methods of analysis to those dependent upon the assumption of normality will rarely be using methods as robust and powerful as available alternative nonparametric methods of analysis. Given the prevalence of measures distributed as Laplace, measures distributed as logistic, and measures having varying degrees of asymmetry, methods of analysis capable of handling these variables are crucial for researchers in education and social science fields.

In this dissertation, I examine select rank-based and normal score-based nonparametric tests of interaction to determine their power and robustness in the presence of deviations from normality. My study focuses on the performance of seven nonparametric tests of interaction effects and compares these to the commonly used parametric test for the 2 x 2 design using Monte Carlo simulation techniques. Power and Type I error rates are recorded and tabled for each test using data from normal, logistic, and Laplace distributions with various levels of a skew parameter. I explore these distributions with all combinations of null and nonnull main and interaction effects and with a range of cellsample sizes so as to study both small- and large-sample robustness. I use a nominal α of 0.05 for all simulations.

Nonparametric Tests Based on Ranks

In this study I compared four nonparametric tests of interaction along with the parametric ANOVA *F* test. These methods are: (1) the rank transform test, (2) the aligned rank transform test, (3) the McSweeney test, and (4) the rank transform Thompson *T* test.

The rank transform test procedure is to pool and rank all data prior to referring the transformed data to the ANOVA *F* statistic. The aligned rank test procedure is to align the data to remove both main effects estimated from the observed data prior to pooling, ranking, and referring the data to the ANOVA *F* statistic. The McSweeney test procedure is also to align the data prior to pooling and ranking. However, the McSweeney statistic is calculated as:

$$H_Y = (N-1)\left(\frac{SS_{Effect}}{SS_{Total}}\right) \tag{1}$$

The null hypothesis of no effect for the McSweeney test is rejected when:

$$H_Y > \chi^2_{\alpha,df_{Effect}} \tag{2}$$

The rank transform Thompson *T* test procedure is to pool and rank the data, the ANOVA *F* statistic is calculated, and significance is also determined using a chi-square test statistic as in the McSweeney test.

For large cell sample sizes, the rank transform test provides power that is similar to the aligned rank transform test and the McSweeney test while also maintaining Type I error rates at the nominal level. However, in the presence of two nonnull interaction effects, the test has lower power and Type I error rates when testing for an interaction. This power reduction ranges from 0-24 percentage points for the skew-normal, 0-4 percentage points for the skew logistic, and 0-24 percentage points for the asymmetric Laplace compared to the aligned rank transform test and the McSweeney test, with larger power reductions occurring as cell sample size decreases and severity of the skew parameter increases. Furthermore, when testing for interactions in the presence of two nonnull main effects, the rank transform test yields less power than the ANOVA *F* test for small cell sample sizes (n = 5) as well.

Power and Type I error rates converge to those of the aligned rank transform test and McSweeney test as cell sample sizes increase, reaching competitive levels for all distributions and skew severities once n = 50. Given the increasing feasibility of using the aligned rank transform test for large sample sizes with the modern computing power that is available, coupled with the robustness of this test for all combination of null and nonnull main effects, I recommend the rank transform test be dismissed as a viable test.

The aligned rank transform test has power that compares favorably to that of the ANOVA *F* test for all distributions, severities of skew, null and nonnull main and interaction effect combinations, and cell sample sizes. The power of the aligned rank transform test is near that of the ANOVA *F* test when the underlying population follows a skew-normal distribution, becomes 5 percentage points higher for the logistic distribution, and as much as 15 percentage points higher for the Laplace distribution. The test has mildly liberal Type I error rates for small cell sample sizes before converging to rates comparable to the ANOVA *F* test but never exceeds a liberal error band cutoff of 0.07. Thus, the aligned rank transform test offers a robust and powerful nonparametric alternative to the ANOVA *F* test that has little risk, at least for the designs used in this research. The robust Type I error and power properties of the aligned rank transform test observed in this study align with previous research supporting the robustness of the test in the presence of nonnormal data (Blair & Sawilowsky, 1990; Fawcett & Salter, 1984; Groggel,1987; Mansouri & Chang, 1995; Salter & Fawcett, 1985, 1993). I recommend the aligned rank transform test for use instead of the ANOVA *F* test under empirical conditions similar to the designs included in this study. Furthermore, given it is robust in the presence of two nonnull main effects unlike the rank transform test, never yields Type

I error rates greater than 0.07 for any cell sample size, and has competitive power for all cell sample sizes compared to other nonparametric tests, I also recommend using this test instead of the rank transform test, McSweeney test, and Thompson *T* test.

The McSweeney test has power similar to the aligned rank transform test but has liberal Type I error rates for small (n = 5) cell sample sizes. These Type I error rates often exceed the liberal error band cutoff of 0.07. The liberal Type I error rates for small cell sample sizes observed in this study align with results in previous research (Harwell, 1991; Kelley et al., 1994; McSweeney, 1967; Toothaker & Newman, 1994). While the Type I error rate converges to that of the ANOVA *F* test for larger cell sample sizes, the test is not a competitive, robust, and risk-free alternative to the parametric *F* test as is the case with the aligned rank transform test for all cell sample sizes. Therefore, I do recommend selecting the McSweeney test instead of the ANOVA *F* test once cell sample sizes become n = 10 or larger, though the aligned rank transform test continues to be a good choice for these cell sizes, as well as for smaller sample sizes.

The Thompson *T* test is conservative when it is used to test for an interaction effect in the presence of two nonnull main effects, as is the case with the rank transform test. Also akin to the rank transform test, power and Type I error rates converge to those of the aligned rank transform test and McSweeney test as cell sample size increases. However, unlike the rank transform test, the Thompson *T* test has Type I error rates that exceed the liberal error band cutoff of 0.07 for varying designs when cell sample sizes are small (n = 5). Due to these properties, I recommend choosing the aligned rank transform test instead of the Thompson *T* test regardless of cell sample size.

Nonparametric Tests Based on Normal Scores

Although no prior research has explored power and robustness properties of the McSweeney test to the presence of increasing severities of skew for various distributions, the liberal Type I error rates of the McSweeney test for small cell sample sizes found in this study is consistent with results from previous research using different distributions (Harwell, 1991; Kelley et al., 1994; McSweeney, 1967; Toothaker & Newman, 1994). McSweeney (1967) suggested using a normal scores data transformation as a method of potentially mitigating the small sample liberal Type I error rates. Weber (1972) reported favorable results when replacing the rank transformation data step in the McSweeney test with a Terry-Hoeffding normal scores data transformation. No prior research has explored the viability of replacing the rank transformation with a Van der Waerden normal scores data transformation when using the McSweeney statistic. The three novel nonparametric tests for interaction that I proposed are the equivalent of using the rank transform test, the aligned rank transform test, and the McSweeney test, yet with a Van der Waerden normal scores data transformation in place of a rank transformation. These methods can be viewed as: (1) a Van der Waerden normal scores data transformation in combination with an *F* statistic (V-*F*), (2) an alignment data transformation prior to a Van der Waerden normal scores data transformation in combination with an *F* statistic (A-V-*F*), and (3) an alignment data transformation prior to a Van der Waerden normal scores data transformation in combination with the McSweeney statistic (A-V-M).

Power and Type I error properties vary across the proposed testing methods. I recommended excluding the rank transform test as a viable method due to conservative Type I error results and smaller power than the ANOVA *F*, aligned rank transform, and

McSweeney tests when testing for an interaction effect in the presence of two nonnull main effects. Further, replacing the rank transformation with a Van der Waerden normal scores transformation (V-*F*) does not mitigate this effect to a degree that would warrant recommendation of this procedure. However, power levels of method V-*F* in the presence of two nonnull main effects converge to those of method A-V-*F* at around n = 30 cell sample sizes for the standard normal and skew-normal distributions and at around n = 10 for the skew logistic distribution with small skew severities, but do not converge up to n = 500 cell sample sizes for the skew logistic distribution with large skew severities and for the asymmetric Laplace distribution. In contrast, the test yields favorable results when testing for an interaction effect in the presence of two null main effects. In the presence of two null main effects, method V-*F* has competitive power to the ANOVA *F* test for the skew-normal distribution, larger power (1-6 percentage points) for the skew logistic distribution for all cell sample sizes, and larger power (1-32 percentage points) for the asymmetric Laplace, increasing as cell sample sizes and skew severity increases. Method V-*F* has larger power than the ANOVA *F* test for all designs studied once cell sample sizes become n = 30 or larger. Despite these comparisons, I cannot recommend method V-*F* because the Type I error properties of the method are dependent on the presence of null or nonnull main effects for small cell sample sizes.

Methods A-V-*F* and A-V-M are strong competitors to the parametric ANOVA *F* test. A-V-*F* and A-V-M occasionally result in mildly liberal Type I error rates for small (n = 5 and n = 10) cell sample sizes. However, using the Type I error rate cutoff bands outlined by Bradley (1978), both methods remain viable tests for all cell sample sizes. Comparing these results to those of the base McSweeney test, replacing the rank transformation with a Van der Waerden normal scores transformation does improve small sample size Type I error rates for the McSweeney test. Power results likewise remain consistent or improve when using a Van der Waerden normal scores transformation in combination with the McSweeney statistic. Methods A-V-*F* and A-V-M are consistently as or more powerful than the ANOVA *F* test for all simulated distributions and cell sample sizes. Method A-V-*F* has power similar to the ANOVA *F* test for the standard normal and all skew-normal designs and has power advantages over the ANOVA *F* test as large as 6 percentage points for the skew logistic and 31 percentage points for the asymmetric Laplace, with the power advantage increasing as skew severity and cell sample size increases. Method A-V-M also has power similar to that of the ANOVA *F* test and method A-V-*F* for the standard normal and all skew-normal designs. Method A-V-M demonstrates as high, or higher power, than method A-V-*F* for the skew logistic and asymmetric Laplace. This power advantage is as large as 3 percentage points for the skew logistic and 2 percentage points for the asymmetric Laplace, increasing in power as cell sample size decreases. Power of method A-V-M is larger than method A-V-*F* for small cell sample sizes with the skew logistic and asymmetric Laplace before the two methods converge to similar power as cell sample sizes increase. These results suggest that using an alignment procedure in combination with a Van der Waerden normal scores data transformation provides for strong testing potential using either the *F* statistic or McSweeney H_y statistic.

Both methods A-V-*F* and A-V-M provide for a more powerful and more robust test than the ANOVA *F* test. Furthermore, both methods A-V-*F* and A-V-M demonstrate power competitive to the aligned rank transform test for all distributions and cell sample sizes. For small cell sample sizes (n = 5 and n = 10), the aligned rank transform test demonstrates power advantages of up to 2 percentage points for the skew logistic distribution and up to 3 percentage points for the asymmetric Laplace compared to method A-V-*F*. However, method A-V-*F* converges to power levels up to 3 percentage points larger for the skew logistic and up to 7 percentage points larger for the asymmetric Laplace once cell sample sizes become n = 50. Method A-V-M has competitive power to the aligned rank transform test for small cell samples sizes for the skew logistic and asymmetric Laplace distributions. Once cell sample sizes become n = 50, method A-V-M converges to power levels up to 1 percentage point larger for the skew logistic and up to 6 percentage points larger for the asymmetric Laplace compared to the aligned rank transform test. In summary, the choice between the aligned rank transform test and method A-V-*F* is dependent on cell sample size but I recommend method A-V-M over the aligned rank transform test for all cell sample sizes.

In this study, I found that three nonparametric tests of interactions are powerful, robust, and risk-free alternatives to the ANOVA *F* test. These three tests are: (1) the aligned rank transform test, (2) the aligned rank transform test with a Van der Waerden normal scores data transformation in place of the rank transformation (method A-V-*F*), and (3) the McSweeney test with a Van der Waerden normal scores data transformation in place of the rank transformation (method A-V-M). I can recommend all three of these tests as alternatives to the ANOVA *F* test for the studied distributions. I recommend method A-V-M as the best of these three for the test of interaction with the studied and simulated distributions because of its robust Type I error properties and strong small- and large-sample power properties compared to the other two nonparametric tests. The

McSweeney statistic used in combination with an alignment data transformation and a Van der Waerden normal scores data transformation yields the most favorable power and Type I error properties of all methods compared. There is strong potential in this testing method that requires further study to ensure strong power and robust Type I error rates when residuals have nonnormal distributions other than the ones included in this study. The following example demonstrates the benefit of using the recommended nonparametric tests when the distribution of the residuals indicates skewed or nonnormal data.

Example

Consider a hypothetical study in which a researcher will test for the presence of nonnull effects for two dichotomous categorical treatment variables on a criterion mastery response variable for a balanced sample of 60 students. (The data for this study are Appendix C.) These factors in this study might be, for example, the presence or absence of supplemental tutoring (factor A) and a traditional and alternative method of instruction (factor B). Though the effects of both tutoring and alternative instruction are of interest, the first step is to determine if these two factors interact so that the effect of one treatment differs for the different levels of the other treatment. The common course of action is for a researcher to assume normal sampling distributions for the effects of interest. Typically, this is a reasonable assumption because of the sample size of 60.

Thus, the researcher performs a parametric analysis of variance setting the maximum tolerance for errors at 0.05 for each of the three hypothesis tests in this study (these being factor A, factor B, and the interaction of factor A with factor B). Table 3 is the ANOVA table using the example data.

Table 5.1
Example ANOVA Table

	Df	SS	MSE	F-value	p-value
A	1	0.1	0.057	0.01	0.922
B	1	5.9	5.87	0.987	0.325
A*B	1	14.3	14.278	2.4	0.127
Residuals	56	333.2	5.95		

From these results, we conclude that there is a small interaction effect present in the sample but we do not have substantial evidence to infer an interaction effect in the larger, target population (p = 0.13). The researcher would not conclude that application of either or both treatment effects has an influence on the response variable.

Rather than assuming normal sampling distributions, we will first look at the distribution of residuals using our model estimates. From Figure 3, a histogram of these residuals, we note a skew effect and kurtosis that could potentially be better modelled with a Laplace distribution instead of a normal distribution. Micceri (1989) noted that 60% of observed criterion mastery measures he studied were Laplace distributed.

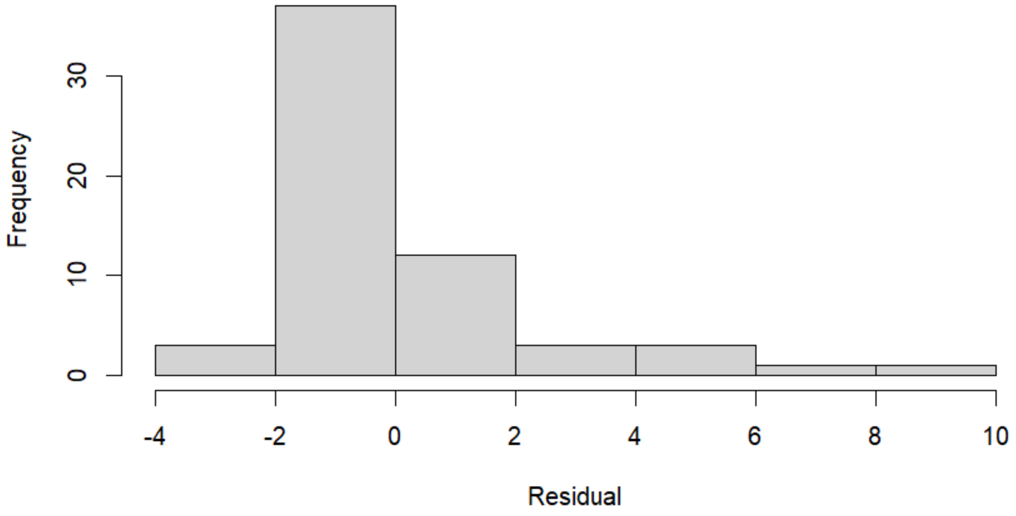

Figure 5.1
Histogram of Residuals

This preliminary analysis prompts us to the McSweeney test with a Van der Waerden normal scores data transformation in place of the rank transformation (Method A-V-M). Table 4 is the results using the McSweeney test with a Van der Waerden normal score data transformation.

Table 5.2
Example data results using Method A-V-M

	Effect SS	H_Y	p-value
A	0.0869	0.0964	0.7562
B	6.3761	7.0725	0.0078
A*B	7.858	8.7163	0.0032

From this table, the researcher would not only conclude and infer the presence of a significant interaction effect between the treatments, but also a significant effect for treatment B as well. The estimated effect size for treatment B is 0.3128 and the estimated effect size for the interaction effect is 0.48782. The researcher would conclude that treatment B has a positive impact on the response variable and that the two treatment effects used in combination also have a positive impact on the response variable.

The original data was simulated using a 0 effect size for the A effect, a 0.2 effect size for the B effect, and a 0.55 effect size for the interaction effect. A researcher using the nonparametric method recommended in this dissertation, the McSweeney test in combination with a Van der Waerden normal scores data transformation, would have concluded both the presence of the B and interaction effects and generalized the effect sizes to the larger, target population using the predetermined alpha level.

Limitations and Future Research

I used simulations of residuals following only normal, logistic, and Laplace distributions with varying severities of skew. I studied the nonparametric tests based on ranks using seven levels of skew severity and the nonparametric tests based on normal scores using four levels of skew severity. This study includes only a two-factor design with two levels of each main effect.

Results tabled in this dissertation provide clear evidence that robust nonparametric methods are needed for the behavioral and social sciences. The parametric ANOVA *F* test suffers severe power reduction when residuals are not normally distributed. Further research into nonparametric methods capable of testing for interaction effects is needed. The novel nonparametric test proposed and simulated of aligning the observed data, performing a Van der Waerden normal scores data transformation, and calculating the McSweeney H_y statistic requires further research both in its asymptotic properties and its performance in the presence of research design aspects not represented in this study. These include designs with residuals from distributions other than those simulated in this dissertation, designs with more than two main effects, designs with more than two levels of each main effect, and more complicated designs than the two- or more-groups factorial design. Of particular interest is the uniform distribution, which is the only distribution not simulated in the designs included in this dissertation. This may not be relevant, as Micceri (1989) found the uniform distribution to comprise 3.2% of achievement and psychometric measures.

Using nonparametric tests based on ranks, including those tests based on normal scores, to make inference regarding an interaction effect requires discussion of the interpretation of results when the test result is a significant p-value. A significant p-value resulting from a nonparametric test of interaction based on ranks implies the existence of a nonzero interaction effect between levels of the treatment effects. For the nonparametric tests included in this study, there is no method of quantifying the effect size of the significant interaction effect other than resorting to the group mean differences a researcher would typically calculate when using the parametric ANOVA *F* test. Research is needed to determine if there are more accurate methods of measuring a significant interaction effect size when using nonparametric tests of interaction based on ranks.

Conclusion

The results of this dissertation provide evidence of the consequences of incorrectly assuming normally distributed data. Results demonstrate an incorrect normality assumption in combination with use of the ANOVA *F* test can result in over 45 percentage points in less power than expected when the distribution is logistic rather than normal and 60 percentage points in less power than expected when the distribution is a Laplace distribution, the actual power loss depending upon sample size and skew severity. While the recommended nonparametric tests for interaction do not always ensure the 80% power based on the normal theory effect size used in this study, the tests mitigated the power loss by as much as 6 percentage points for the logistic distribution and 31 percentage points for the Laplace distribution depending upon sample size and skew severity.

A statistically significant p-value resulting from a nonparametric test of interaction based on ranks or normal scores implies the existence of a nonzero interaction effect between levels of the treatment effects. This effect can be estimated using group mean differences just as when using the common parametric ANOVA *F* test. Statistically significant results of a nonparametric test of interaction imply there is an interaction effect, just as when using the parametric *F* test, and the same measures of the size of this effect remain valid.

Results presented in this dissertation provide evidence of a method for which researchers situated in the behavioral and social sciences can use the novel, proposed, and recommended McSweeney test with a Van der Waerden normal scores transformation to test hypotheses of significant main and interaction effects in the 2 x 2 design. The proposed test is adaptable to all factorial designs with categorical predictor variables, although further research is needed to ensure power and Type I error properties remain favorable for higher level designs. The McSweeney test with a Van der Waerden normal scores data transformation provides for a nonparametric test of interaction that is robust to Type I error, competitive to the ANOVA *F* test when normality can be safely assumed, and more powerful than the ANOVA *F* test for distributions concluded as those occurring in applied measures in the behavioral and social sciences by Micceri (1989). Restated, these measures are predominantly those which are asymmetric, have larger tail weights than the normal distribution, or are Laplace. The results in this dissertation provide evidence that the McSweeney test with a Van der Waerden normal scores data transformation is more powerful than the ANOVA *F* test for all of the distributions found by Micceri (1989) as occurring in achievement and psychometric measures except for the uniform distribution, which Micceri found to comprise only 3.2% of observed measures. Therefore, it is strongly recommended that the McSweeney test with a Van der Waerden normal scores data transformation be used for analysis with achievement, psychometric, and other applied measures over the *F* test. It is hoped this dissertation contributes to the field of research on powerful and robust nonparametric tests of interaction, as well as the use of nonparametric methods in practice when such methods can help researchers have more power to detect effects.

BIBLIOGRAPHY

Agresti, A., & Pendergast, J. (1986). Comparing mean ranks for repeated measures data. *Communications in Statistics - Theory and Methods*, *15*(5), 1417–1433. https://doi.org/10.1080/03610928608829193

Akritas, M. G. (1990). The rank transform method in some two factor designs. *Journal of the American Statistical Association*, *85*(409), 73-78.

Akritas, M. G., & Arnold, S. F. (1994). Fully nonparametric hypotheses for factorial designs: Multivariate repeated measures designs. *Journal of the American Statistical Association*, *89*(425), 336–343. https://doi.org/10.1080/01621459.1994.10476475

Akritas, M. G., & Brunner, E. (1997). A unified approach to rank tests for mixed models. *Journal of Statistical Planning and Inference*, *61*(2), 249–277. https://doi.org/10.1016/s0378-3758(96)00177-2

Andrews, F. C. (1954). Asymptotic behavior of some rank tests for analysis of variance. *Annals of Mathematical Statistics*, *25*, 724–736.

Arnholt, A. T. (2007). Resampling with R. *Teaching Statistics*, *29*(1), 21–26.

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

Beasley, T. M., & Zumbo, B. D. (2009). Aligned rank tests for interactions in split-plot designs: Distributional assumptions and stochastic heterogeneity. *Journal of Modern Applied Statistical Methods*, *8*(1), 16–50. https://doi.org/10.22237/jmasm/1241136180

Bell, C. B., & Doksum, K. A. (1965). Some new distribution-free statistics. *Annals of Mathematical Statistics*, *36*(1), 203-214.

Berry, K. J., & Mielke, P. W. (1983). Moment approximations as an alternative to the F test in analysis of variance. *British Journal of Mathematical and Statistical Psychology*, *36*(2), 202–206. https://doi.org/10.1111/j.2044-8317.1983.tb01125.x

Bhapkar, V. P., & Gore, A. P. (1974). A nonparametric test for interaction in two-way layouts. *The Indian Journal of Statistics, Series A*, *36*(3), 261–272.

Bishop, T. (1976). *Heteroscedastic ANOVA, MANOVA and multiple comparisons* [Doctoral dissertation, Ohio State University].

Blair, R. C. (1980). A comparison of the power of the two independent means t test to that of the wilcoxon's rank-sum test for samples of various populations [Doctoral dissertation, University of South Florida].

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. (1980a). A comparison of the power of the t test and the Wilcoxon statistic when samples are drawn from a certain mixed normal distribution. *Evaluation Review*, *4*, 645–656.

Blair, R. C., & Higgins, J. J. (1980b). A comparison of the power of the Wilcoxon’s rank-sum statistic for that of student’s t statistic under carious non-normal distributions. *Journal of Educational Statistics*, *5*(4), 309–335.

Blair, R. C., & Higgins, J. J. (1981). A note on the asymptotic relative efficiency of the wilcoxon rank-sum test relative to the independent means t test under mixtures of two normal distributions. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 124–128. https://doi.org/10.1111/j.2044-8317.1981.tb00623.x

Blair, R. C., & Higgins, J. J. (1985). A comparison of the power of the paired samples rank transform statistic to that of Wilcoxon’s signed rank statistic. *Journal of Educational Statistics*, *10*(4), 368–383.

Blair, R. C., & Sawilowsky, S. S. (1990). American Educational Research Association. In *A test for interaction based on the rank transform*. Boston, MA.

Blair, R. C., Sawilowsky, S. S., & Higgins, J. J. (1987). Limitations of the rank transform statistic in tests for interactions. *Communications in Statistics - Simulation and Computation*, *16*(4), 1133–1145. https://doi.org/10.1080/03610918708812642

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16. https://doi.org/10.3102/0013189x013006004

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

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

Box, G. E. P. (1953). Non-normality and tests of variances. *Biometrika*, *40*, 318–355.

Box, G. E. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, *25*(2), 290–302. https://doi.org/10.1214/aoms/1177728786

Bradbury, I. (1987). Analysis of variance versus randomization tests-a comparison. *British Journal of Mathematical and Statistical Psychology*, *40*(2), 177–187. https://doi.org/10.1111/j.2044-8317.1987.tb00877.x

Bradley, J. V. (1968). *Distribution-free statistical tests*. Prentice-Hall.

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *The American Statistician*, *31*(4), 147–150. https://doi.org/10.2307/2683535

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

Bradley, J. V. (1979). A nonparametric test for interactions of any order. *Journal of Quality Technology*, *11*(4), 177–184. https://doi.org/10.1080/00224065.1979.11980909

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

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

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

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

Brown, M. B., & Forsythe, A. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, *16*, 129–132.

Brunner, E., & Dette, H. (1992). Rank procedures for the two-factor mixed model. *Journal of the American Statistical Association, 87*(419), 884–888. https://doi.org/10.1080/01621459.1992.10475292

Brunner, E., & Neumann, N. (1984). Rank tests for the 2 x 2 split plot design. *Metrika*, *31*, 233–243.

Canay, I. A., Romano, J. P., & Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, *85*(3), 1013–1030. https://doi.org/10.3982/ecta13081

Chen, R. S., & Dunlap, W. P. (1993). SAS procedures for approximate randomization tests. *Behavior Research Methods, Instruments, & Computers*, *25*(3), 406–409. https://doi.org/10.3758/bf03204532

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

Christensen, W. F., & Zabriskie, B. N. (2021). When your permutation test is doomed to fail. *The American Statistician*, *76*(1), 53–63. https://doi.org/10.1080/00031305.2021.1902856

Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, *3*, 27–38.

Cochran, W. G., & Cox, G. M. (1950). *Experimental designs*. Wiley.

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). John Wiley.

Conover, W. J., & Iman, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs. *Communications in Statistics - Theory and Methods*, *5*(14), 1349–1368. https://doi.org/10.1080/03610927608827447

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, *35*(3), 124–129. https://doi.org/10.2307/2683975

Craig, A. R., & Fisher, W. W. (2019). Randomization tests as alternative analysis methods for behavior-analytic data. *Journal of the Experimental Analysis of Behavior*, *111*(2), 309–328. https://doi.org/10.1002/jeab.500

Curtis, D. A., & Marascuilo, L. A. (1992). Point estimates and confidence intervals for the parameters of the two-sample and matched-pair combined tests for ranks and normal scores. *The Journal of Experimental Education*, *60*(3), 243–269.

Davis, J. B., & McKean, J. W. (1993). Rank-based methods for multivariate linear models. *Journal of the American Statistical Association*, *88*(421), 245–251. https://doi.org/10.2307/2290719

De Kroon, J., and Van Der Laan, P. (1981). Distribution-free test procedures in two-way layouts: A concept of rank interaction. *Statistica Neerlandica, 35*, 189-213.

De Neve, J., & Thas, O. (2017). A Mann-Whitney type effect measure of interaction for factorial designs. *Communications in Statistics - Theory and Methods*, *46*(22), 11243–11260.

DeWees, T. A., Mazza, G. L., Golafshar, M. A., & Dueck, A. C. (2020). Investigation into the effects of using normal distribution theory methodology for Likert scale patient-reported outcome data from varying underlying distributions including floor/ceiling effects. *Value in Health*, *23*(5), 625–631.

Ditzhaus, M., Fried, R., & Pauly, M. (2021). QANOVA: quantile-based permutation methods for general factorial designs. *TEST*, *30*, 960–979.

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

Edgington, E. S. (1969). *Statistical inference: The distribution-free approach*. McGraw-Hill.

Edgington, E. S. (1980). *Randomization tests*. Marcel Dekker.

Elliffe, D., & Elliffe, M. (2019). Rank-permutation tests for behavior analysis, and a test for trend allowing unequal data numbers for each subject. *Journal of the Experimental Analysis of Behavior, 111*(2), 342–358. https://doi.org/10.1002/jeab.502

Erikson, R. S., Pinto, P. M., & Rader, K. T. (2010). Randomization tests and multi-level data in U.S. state politics. *State Politics & Policy Quarterly*, *10*(2), 180–198. https://doi.org/10.1177/153244001001000204

Fawcett, R. F., & Salter, K. C. (1984). A Monte Carlo study of the F test and three tests based on ranks of treatment effects in randomized block designs. *Communications in Statistics - Simulation and Computation, 13*(2), 213–225. https://doi.org/10.1080/03610918408812368

Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores, and Kruskal-Wallis test under violations of assumptions. *Educational and Psychological Measurement, 34*, 789–799.

Feng, Y., Hancock, G. R., & Harring, J. R. (2019). Latent growth models with floor, ceilings, and random knots. *Multivariate Behavioral Research*, *54*(5), 751–770.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, *A*(222), 309–368.

Fisher, R. A. (1935). *The design of experiments.* Oliver & Boyd.

Fisher, R. A., & Yates, F. (1949). *Statistical tables for biological, agricultural, and medical research* (3rd ed.). Hafner.

Fligner, M. A., & Policello, G. E. (1981). Robust rank procedures for the Behrens-Fisher Problem. *Biometrika*, *69*, 221-226.

Freidlin, B., & Gastwirth, J. L. (2000). Should the median test be retired from general use? *The American Statistician, 54*(3), 161–164. https://doi.org/10.2307/2685584

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675–701. https://doi.org/10.1080/01621459.1937.10503522

Gaito, J. (1959). Non-parametric methods in psychological research. *Psychological Reports, 5*, 115-125.

Gao, X., & Alvo, M. (2005a). A nonparametric test for interaction in two-way layouts. *The Canadian Journal of Statistics*, *33*(4), 529–543.

Gao, X., & Alvo, M. (2005b). A unified nonparametric approach for unbalanced factorial designs. *Journal of the American Statistical Association*, *100*(471), 926–941.

Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research*, *45*(1), 43-57.

Garthwaite, P. H. (1996). Confidence intervals from randomization tests. *Biometrics, 52*(4), 1387–1393. https://doi.org/10.2307/2532852

Gibbons, J. D. (1985). *Nonparametric Statistical Inference* (2nd ed.). Marcel Dekker.

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

Goodard, R. H., & Lindquist, E. F. (1940). An empirical study of the effects of heterogeneous within groups variance upon certain F-tests of significance in analysis of variance. *Psychometrica*, *5*, 263–274.

Goulden, N., McKie, S., Suckling, J., Williams, S. R., Anderson, I. M., Deakin, J. F., & Elliott, R. (2010). A comparison of permutation and parametric testing for between group effective connectivity differences using DCM. *NeuroImage*, *50*(2), 509–515. https://doi.org/10.1016/j.neuroimage.2009.11.059

Groggel, D. J. (1987). A Monte Carlo Study of rank tests for block designs. *Communications in Statistics - Simulation and Computation*, *16*(3), 601–620. https://doi.org/10.1080/03610918708812607

Guilford, J. P., & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). McGraw-Hill.

Habib, A. R., & Harwell, M. R. (1989). An empirical study of the Type I error rate and power for some selected normal-theory and nonparametric tests of the independence of two sets of variables. *Communications in Statistics, 18*(2), 793-826.

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

Harter, H. L. (1961). Expected values of normal order statistics. *Biometrika, 48*, 151–165.

Hartlaub, B. A., Dean, A. M., & Wolfe, D. A. (1999). Rank-based test procedures for interaction in the two-way layout with one observation per cell. *Canadian Journal of Statistics*, *27*(4), 863–874. https://doi.org/10.2307/3316137

Harwell, M. R. (1990). A general approach to hypothesis testing for nonparametric tests. *The Journal of Experimental Education, 58*(2), 143–156. https://doi.org/10.1080/00220973.1990.10806530

Harwell, M. R. (1991). Completely randomized factorial analysis of variance using ranks. *British Journal of Mathematical and Statistical Psychology, 44*(2), 383–401. https://doi.org/10.1111/j.2044-8317.1991.tb00970.x

Harwell, M. R. (2003). Summarizing Monte Carlo results in methodological research: The single-factor, fixed-effects ANCOVA case. *Journal of Educational and Behavioral Statistics, 28*(1), 45–70. https://doi.org/10.3102/10769986028001045

Harwell, M. R., & Serlin, R. C. (1988). An empirical study of a proposed test of nonparametric analysis of covariance. *Psychological Bulletin*, *104*(2), 268–281. https://doi.org/10.1037/0033-2909.104.2.268

Harwell, M. R., & Serlin, R. C. (1989a). A nonparametric test statistic for the general linear model. *Journal of Educational Statistics*, *14*(4), 351–371. https://doi.org/10.2307/1164944

Harwell, M. R., & Serlin, R. C. (1989b). Annual Meeting of the American Educational Research Association. In *An empirical study of the Friedman test under covariance heterogeneity*. San Francisco.

Havlicek, L. L., & Peterson, N. L. (1974). Robustness of the t-test in a guide for researchers on effect of violations of assumptions. *Psychological Reports*, *34*, 1095–1114.

Hayes, A. F. (1998). SPSS procedures for approximate randomization tests. *Behavior Research Methods, Instruments, & Computers*, *30*(3), 536–543. https://doi.org/10.3758/bf03200687

Headrick, T. C., & Vineyard, G. (2000). *An empirical investigation of four tests for interaction in the context of factorial analysis of covariance.* Southern Illinois University at Carbondale.

Hellstrom, A. (1993). The normal distribution in scaling subjective stimulus difference: Less "normal" than we think? *Perception & Psychophysics*, *54*(1), 82–92.

Helwig, N. E. (2019). Robust nonparametric tests of general linear model coefficients: A comparison of permutation methods and test statistics. *NeuroImage*, *201*. https://doi.org/10.1016/j.neuroimage.2019.116030

Hemelrijk, J. (1961). Experimental comparison of Student's and Wilcoxon's two sample test. *Quantitative Methods in Psychology*. Interscience.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. John Wiley & Sons.

Hettmansperger, T. P., McKean, J. W., & Sheather, S. J. (2000). Robust nonparametric methods. *Journal of the American Statistical Association*, *95*(452), 1308–1312.

Hildebrand, D. K. (1986). *Statistical thinking for behavioral scientists*. Duxbury Press.

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

Hodges, J. L., & Lehmann, E. L. (1961). Comparison of the Normal Scores and Wilcoxon Tests. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1). Berkeley; University of California Press.

Hodges, J. L., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics*, *33*(2), 482–497. https://doi.org/10.1214/aoms/1177704575

Hoeffding, W. (1951). Berkeley Symposium on Mathematics, Statistics, and Probability. In *Optimum nonparametric statistics* (pp. 83–92). Berkeley, CA; University of California Press.

Hoeffding, W. (1952). The large sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, *23*, 169–192.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric methods*. John Wiley.

Hornsby Brown, M. E. (2023). *Nonparametric Tests of Interaction for the Two-Way Design with Skewed Distributions*. [Manuscript submitted for publication].

Hornsnell, G. (1953). The effect of unequal group variances on the F-test for the homogeneity of group means. *Biometrika*, *40*, 128–136.

Iman, R. L. (1974). A power study of a rank transform for the two-way classification model when interaction may be present. *Canadian Journal of Statistics*, *2*(1-2), 227–239. https://doi.org/10.2307/3314695

Iman, R. L., & Conover, W. J. (1976). *A comparison of several rank tests for the two-way layout* (SAND76-0631). Albuquerque, NM: Sandia Laboratories.

Jenkins, S. J., Fuqua, D. R., & Hartman, B. W. (1984). Evaluating Criteria for selection of nonparametric statistics. *Perceptual and Motor Skills, 59,* 31-35.

Kelley, D. L., Sawilowsky, S. S., & Blair, R. C. (1994). Midwestern Educational Research Association. In *Comparison of ANOVA, McSweeney, Bradley, Harwell-Serlin, and Blair-Sawilowsky tests in the balanced 2x2x2 layout.* Chicago, IL.

Kennedy, P. E. (1995). Randomization tests in econometrics. *Journal of Business & Economic Statistics*, *13*(1), 85–94. https://doi.org/10.2307/1392523

Kepner, J. L., & Robinson, D. H. (1988). Nonparametric methods for detecting treatment effects in repeated-measures designs. *Journal of the American Statistical Association*, *83*(402), 456–461. https://doi.org/10.1080/01621459.1988.10478617

Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd ed.). Holt, Rinehart, and Winston.

Keselman, H. J., Rogan, J. C., & Feir-Walsh, B. J. (1977). An evaluation of some nonparametric tests for location equality. *British Journal of Mathematical and Statistical Psychology*, *30*, 213–221.

Keselman, H. J., & Toothaker, L. E. (1973). 81st Annual Convention. In *An Empirical Comparison of the Marascuilo and Normal Scores Nonparametric Tests and the Scheffe and Tukey Parametric Tests for Pairwise Comparisons* (pp. 15–16). APA.

Kleijnen, P. C. (1987). *Statistical tools for simulation practitioners*. Marcel Dekker.

Klotz, J. H. (1964). On the normal scores two-sample rank test. *Journal of the American Statistical Association, 59*, 652–664.

Krauth, J. (1988). *Distribution-free statistics: An application-oriented approach*. Elsevier.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *Annals of Mathematical Statistics, 23*, 525–545.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

LaFleur, B. J., & Greevy, R. A. (2009). Introduction to permutation and resampling-based hypothesis tests. *Journal of Clinical Child & Adolescent Psychology, 38*(2), 286–294. https://doi.org/10.1080/15374410902740411

Lehmann, E. L., & Stein, C. (1959). *Testing statistical hypotheses*. John Wiley.

Lemmer, H. H. (1980). Some empirical results on the two-way analysis of variance by ranks. *Communications in Statistics - Theory and Methods, 9*(14), 1427–1438. https://doi.org/10.1080/03610928008827972

Lemmer, H. H. (2001). Note on Sawilowsky's paper on the rank transform. *Perceptual and Motor Skills, 92*(2), 433–434. https://doi.org/10.2466/pms.2001.92.2.433

Leys, C., & Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology, 46*(4), 684–688. https://doi.org/10.1016/j.jesp.2010.02.007

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and Education.* Houghton.

Liu, Q., & Wang, L. (2021). T-test and ANOVA for data with ceiling and/or floor effects. *Behavior Research Methods*, *53*(1), 264–277.

Lu, H. T., & Smith, P. J. (1979). Distribution of the normal scores statistic for nonparametric one-way analysis of variance. *Journal of the American Statistical Association*, *74*(367), 715–722. https://doi.org/10.1080/01621459.1979.10481676

Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician*, *52*(2), 127–132. https://doi.org/10.2307/2685470

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement*, *7*, 263–269.

Mandeville, G. K. (1972). A new look at treatment differences. *American Educational Research Journal*, *9*, 311–321.

Manly, B. F. J. (1995). Randomization Tests to Compare Means with Unequal Variation. *The Indian Journal of Statistics*, *57*, 200–222.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Mansouri, H., & Chang, G.-H. (1995). A comparative study of some rank tests for interaction. *Computational Statistics & Data Analysis*, *19*(1), 85–96. https://doi.org/10.1016/0167-9473(93)e0045-6

Marascuilo, L. A., & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences*. Freeman.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Brooks/Cole.

Marden, J. I., & Muyot, M. E. (1995). Rank tests for main and interaction effects in analysis of variance. *Journal of the American Statistical Association*, *90*(432), 1388–1398. https://doi.org/10.1080/01621459.1995.10476644

McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit model. *Gifted Child Quarterly*, *54*(4), 314–320.

McKean, J. W., & Vidmar, T. J. (1994). A comparison of two rank-based methods for the analysis of linear models. *The American Statistician*, *48*(3), 220. https://doi.org/10.2307/2684721

McSweeney, M. T. (1967). An empirical study of two proposed nonparametric tests [Doctoral dissertation, University of California, Berkeley].

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

Mielke, P. W., & Berry, K. J. (1994). Permutation tests for common locations among samples with unequal variances. *Journal of Educational and Behavioral Statistics*, *19*(3), 217–236. https://doi.org/10.2307/1165295

Mohebbi, C., & Shoemaker, L. H. (1990). An extension of the median test to analysis of variance. *Communications in Statistics - Theory and Methods*, *19*(3), 1101–1117. https://doi.org/10.1080/03610929008830249

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

Noether, G. E. (1967). *Elements of nonparametric statistics*. Wiley.

Noether, G. E. (1981). Comment. *American Statistician, 35*(3), 129–132.

Noortgate, W. V. den, Boeck, P. D., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational & Behavioral Statistics*, *28*(4), 369–386.

Nunnally, J. C. (1975). *Introduction to statistics for psychology and education*. McGraw-Hill.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Nussbaum, E. M. (2015). *Categorical and nonparametric data analysis: Choosing the best statistical technique*. Routledge.

Olejnik, S. F., & Algina, J. (1984). Parametric ANCOVA and the rank transform ANCOVA when the data are conditionally non-normal and heteroscedastic. *Journal of Educational Statistics*, *9*(2), 129–150. https://doi.org/10.2307/1164717

Owen, D. B. (1962). *Handbook of statistical tables*. Addison-Wesley.

Padmanabhan, A. R. (1977). A comparison of the efficiencies of the c-sample normal scores and the Kruskal-Wallis tests in the case of grouped data. *The British Journal of Mathematical & Statistical Psychology*, *30*(2), 222–226.

Patel, K. M., & Hoel, D. G. (1973). A nonparametric test for interaction in factorial experiments. *Journal of the American Statistical Association*, *68*(343), 615–620. https://doi.org/10.1080/01621459.1973.10481394

Pearson, E. S. (1929). The analysis of variance in cases of nonnormal variation. *Biometrika*, *23*, 259–286.

Penfield, D. A. (1994). Choosing a two-sample location test. *The Journal of Experimental Education*, *62*(4), 343–360. https://doi.org/10.1080/00220973.1994.9944139

Penfield, D. A., & McSweeney, M. T. (1968). The normal scores test for the two-sample problem. *Psychological Bulletin*, *69*(3), 183–191.

Peres-Neto, P. R., & Olden, J. D. (2001). Assessing the robustness of randomization tests: Examples from Behavioral Studies. *Animal Behavior*, *61*(1), 79–86. https://doi.org/10.1006/anbe.2000.1576

Peterson, K. (2002). Six modifications of the aligned rank transform test for interaction. *Journal of Modern Applied Statistical Methods*, *1*(1), 100–109. https://doi.org/10.22237/jmasm/1020255240

Pillar, V. (2013). How accurate and powerful are randomization tests in multivariate analysis of variance? *Community Ecology*, *14*(2), 153–163. https://doi.org/10.1556/comec.14.2013.2.5

Pitman, E. J. G. (1948). *Non-parametric statistics* [Lecture notes]. New York: Columbia University

Potvin, C., & Roff, D. A. (1993). Distribution-free and robust statistical methods: Viable alternatives to parametric statistics. *Ecology*, *74*(6), 1617–1628. https://doi.org/10.2307/1939920

Puri, M. L., & Sen, P. K. (1969). A class of rank order tests for a general linear hypothesis. *The Annals of Mathematical Statistics*, *40*(4), 1325–1343. https://doi.org/10.1214/aoms/1177697505

Puri, M. L., & Sen, P. K. (1985). *Nonparametric methods in general linear models*. Wiley.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Randles, R. H., & Wolfe, D. A. (1979). *Introduction to the theory of nonparametric tests*. John Wiley.

Randolph, E. A., & Barcikowski, R. S. (1989). 11th Annual Meeting of the Mid-Western Educational Research Association. In *Type I error rate when real study values are used as population parameters in a Monte Carlo study*. Chicago.

Rasmussen, J. L. (1985). The Power of Student’s t and Wilcoxon W statistics. *Evaluation Review, 9*(4), 505–510.

Rasmussen, J. L. (1986). An evaluation of parametric and nonparametric tests on modified and nonmodified data. *British Journal of Mathematical and Statistical Psychology*, *39*, 213–220.

Reinach, S. G. (1965). A nonparametric analysis for a multiway classification with one element per cell. *South African Journal of Agricultural Science*, *8*, 941–960.

Reinach, S. G. (1966). Distribution-free methods in experimental design [Doctoral dissertation, University of Pretoria].

Rheinheimer, D. C., & Penfield, D. A. (2001). The effects of Type I error rate and power of the ANCOVA F test and selected alternatives under nonnormality and variance heterogeneity. *The Journal of Experimental Education*, *69*(4), 373–391. https://doi.org/10.1080/00220970109599493

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

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. *American Educational Research Journal*, *14*, 493–498.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, *17*(1), 141–159. https://doi.org/10.1214/aos/1176347007

Salter, K. C., & Fawcett, R. F. (1985). A robust and powerful rank test of treatment effects in balanced incomplete block designs. *Communications in Statistics - Simulation and Computation, 14*(4), 807–828. https://doi.org/10.1080/03610918508812475

Salter, K. C., & Fawcett, R. F. (1993). The art test of interaction: A robust and powerful rank test of interaction in factorial models. *Communications in Statistics - Simulation and Computation*, 22(1), 137–153. https://doi.org/10.1080/03610919308813085

Savalei, V. (2006). Logistic approximation to the normal: the KL rationale. *Psychometrika*, *71*(4), 763–767.

Sawilowsky, S. S. (1985). Robust and power analysis of the 2 x 2 x 2 ANOVA, rank transformation, random normal scores, and expected normal scores transformation tests. [Doctoral dissertation, University of South Florida].

Sawilowsky, S. S. (1989). *Rank transform: The bridge is falling down*. American Educational Research Association. San Francisco.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60*(1), 91–126. https://doi.org/10.3102/00346543060001091

Sawilowsky, S. S. (2000). Review of the rank transform in designed experiments. *Perceptual and Motor Skills, 90*(2), 489–497. https://doi.org/10.2466/pms.2000.90.2.489

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. https://doi.org/10.1037/0033-2909.111.2.352

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the Type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics, 14*(3), 255–267. https://doi.org/10.3102/10769986014003255

Scheffe, H. (1959). *The analysis of variance*. Wiley.

Shiraishi, T. (1991). Statistical inference based on aligned ranks for two-way manova with interaction. *Annals of the Institute of Statistical Mathematics, 43*(4), 715–734. https://doi.org/10.1007/bf00121650

Shoemaker, L. H. (1986). A nonparametric method for analysis of variance. *Communications in Statistics - Simulation and Computation*, *15*(3), 609–632. https://doi.org/10.1080/03610918608812528

Siegel. (1956). *Nonparametric statistics for the behavioral sciences*. McGraw-Hill.

Smitley, W. D. S. (1981). *A comparison of the power of the two independent means t test and the Mann-Whitney U test* [Doctoral dissertation, University of South Florida].

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

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Stephenson, W. R., & Jacobson, D. (1988). A comparison of nonparametric analysis of covariance techniques. *Communications in Statistics - Simulation and Computation*, *17*(2), 451–461. https://doi.org/10.1080/03610918808812673

Still, A. W., & White, A. P. (1981). The approximate randomization test as an alternative to the F test in analysis of variance. *British Journal of Mathematical and Statistical Psychology*, *34*(2), 243–252. https://doi.org/10.1111/j.2044-8317.1981.tb00634.x

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

Terry, M. E. (1952). Some rank order tests which are most powerful against specific parametric alternatives. *The Annals of Mathematical Statistics*, *23*(3), 346–366. https://doi.org/10.1214/aoms/1177729381

Thompson, G. L. (1991). A note on the rank transform for interactions. *Biometrika*, *78*(3), 697–701. https://doi.org/10.1093/biomet/78.3.697

Thompson, R., Doksum, K. A., & Govindarajulu, Z. (1966). One-sample normal scores test, distribution and power. *Review of the International Statistics Institute*, *34*, 24–26.

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2017). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review*, *41*(5), 472-505.

Tomarkin, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*(1), 90–99.

Toothaker, L. E., & Newman, D. (1994). Nonparametric competitors to the two-way ANOVA. *Journal of Educational and Behavioral Statistics*, *19*(3), 237–273. https://doi.org/10.2307/1165296

Umlauft, M., Konietschke, F., & Pauly, M. (2017). Rank-based permutation approaches for non-parametric factorial designs. *British Journal of Mathematical and Statistical Psychology*, *70*, 368–390.

Van der Laan, P. (1964). Exact power of some rank tests. *Publications de l'Instituit de Statistique de l'Universite de Paris*, *13*, 211–234.

Van der Laan, P., & Oosterhoff, J. (1965). Monte Carlo estimation of the powers of the distribution-free two sample tests of Wilcoxon, Van der Waerden and Terry and Comparison of these powers. *Statistica Neerlandica*, *19*, 265–275.

Van der Laan, P., & Oosterhoff, J. (1967). Experimental determination of the power functions of the two sample rank tests of Wilcoxon, Van der Waerden, and Terry by Monte Carlo techniques – I. normal parent distributions. *Statistica Neerlandica, 21*, 55-68.

Van der Waerden, B. L. (1952). Order tests for the two-sample problem and their power. *Indagationes Mathematicae*, *14*, 453-458.

Van der Waerden, B. L. (1953). Order tests for the two-sample problem. *Indagationes Mathematicae*, *15*, 303–316.

Van der Waerden, B. L. (1956). The computation of the X-distribution. *Proceedings of the third Berkeley symposium on mathematical statistics and probability*. Berkeley; University of California Press.

Van Elteren, P., & Noether, G. E. (1959). The asymptotic efficiency of the Xr2 test for a balanced incomplete block design. *Biometrika*, *46*, 465–477.

Walberg, J. H., Strykowski, B. F., Rovai, E., & Hurg, S. S. (1984). Exceptional performance. *Review of Educational Research*, *54*, 87–112.

Weber, J. M. (1972). The heuristic explication of a large-sample normal scores test for interaction. *British Journal of Mathematical and Statistical Psychology*, *25*, 246–256.

Welch, B. L. (1937). The significance of the difference between two means when the population Variances are Unequal. *Biometrika*, *29*, 350–362.

Wiedermann, W. T., & Alexandrowicz, R. W. (2011). A modified normal scores test for paired data. *Methodology*, *7*(1), 25–38.

Wilcox, R. R., Charlin, V., & Thompson, K. (1986). New Monte Carlo results on the Robustness of the ANOVA, F, W, and F* statistics. *Communications in Statistics - Simulations and Computation*, *15*, 933–944.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80–82. https://doi.org/10.2307/3001968

Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics*, *3*, 119–122.

Wilcoxon, F. (1949). *Some rapid approximate statistical procedures*. American Cyanamid.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). McGraw-Hill.

Zimmerman, D. W. (1996). A note on homogeneity of variance of scores and ranks. *The Journal of Experimental Education*, *64*(4), 351–362. https://doi.org/10.1080/00220973.1996.10806603

Zimmerman, D. W. (2011). Inheritance of properties of normal and non-normal distributions after transformation of scores to ranks. *Psicologica*, *32*, 65–85.

Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of the Wilcoxon test, the Friedman Test, and repeated-measures ANOVA on ranks. *The Journal of Experimental Education*, *62*(1), 75–86. https://doi.org/10.1080/00220973.1993.

APPENDIX A

CHAPTER 3 R CODE

```r
#Create empty vectors for storing p-values.
for (s in c(4, 3, 2, 1, 1/2, 1/3, 1/4)) {

for (a in c(0, 0.199079)) {
for (b in c(0, 0.199079)) {
for (c in c(0, 0.199079)) {

my.vector.anova.a <- vector()
my.vector.rt.a <- vector()
my.vector.art.a <- vector()
my.vector.hl.a <- vector()

my.vector.anova.b <- vector()
my.vector.rt.b <- vector()
my.vector.art.b <- vector()
my.vector.hl.b <- vector()

my.vector.anova.i <- vector()
my.vector.rt.i <- vector()
my.vector.art.i <- vector()
my.vector.hl.i <- vector()

my.vector.rt.i2 <- vector()
my.vector.rt.i3 <- vector()

#Open the loop to perform the simulation.
for (i in 1:5000) {
```

```
mu = 0
beta1 = a
beta2 = b
beta3 = c

#Select cell sample size, distribution, and skew effect.

A = c(rep(c(-1), 100), rep(c(1), 100))
B = rep(c(rep(c(-1), 50), rep(c(1), 50)), 2)
e = ralaplace(200, 0, 1, s)


y = mu + beta1*A + beta2*B + beta3*B*A + e


my.data = data.frame(cbind(A, B, y))

my.data$A <- factor(my.data$A)
my.data$B <- factor(my.data$B)

#Create output for ANOVA F test and store p-values.

aov.model = aov(y ~ A*B, data=my.data)


my.vector.anova.a <- c(my.vector.anova.a,
        summary(aov.model)[[1]][[1,"Pr(>F)"]])
my.vector.anova.b <- c(my.vector.anova.b,
        summary(aov.model)[[1]][[2,"Pr(>F)"]])
my.vector.anova.i <- c(my.vector.anova.i,
        summary(aov.model)[[1]][[3,"Pr(>F)"]])
```

```
my.data$y.ranked <- rank(my.data$y)

#Create model for rank transform test and store p-values.

aov.model.rt = aov(y.ranked ~ A*B, data = my.data)



my.vector.rt.a <- c(my.vector.rt.a,
        summary(aov.model.rt)[[1]][[1,"Pr(>F)"]])
my.vector.rt.b <- c(my.vector.rt.b,
        summary(aov.model.rt)[[1]][[2,"Pr(>F)"]])
my.vector.rt.i <- c(my.vector.rt.i,
        summary(aov.model.rt)[[1]][[3,"Pr(>F)"]])

#Create model for aligned rank transform test and store p-values.

art.model <- art(my.data$y ~ my.data$A * my.data$B)



my.vector.art.a <- c(my.vector.art.a, anova(art.model)[[1,"Pr(>F)"]])
my.vector.art.b <- c(my.vector.art.b, anova(art.model)[[2,"Pr(>F)"]])
my.vector.art.i <- c(my.vector.art.i, anova(art.model)[[3,"Pr(>F)"]])

#Create models for McSweeney test and store p-values.

align.interact <- art.model$aligned.ranks$`my.data$A:my.data$B`
align.A <- art.model$aligned.ranks$`my.data$A`
align.B <- art.model$aligned.ranks$`my.data$B`



fit.A <- aov(align.A ~ A*B, data = my.data)
summary(fit.A)

df.T = summary(fit.A)[[1]][[1,"Df"]] +
        summary(fit.A)[[1]][[2,"Df"]] +
        summary(fit.A)[[1]][[3,"Df"]] +
        summary(fit.A)[[1]][[4,"Df"]]

SS.G = summary(fit.A)[[1]][[1,"Sum Sq"]]

SS.T = summary(fit.A)[[1]][[1,"Sum Sq"]] +
        summary(fit.A)[[1]][[2,"Sum Sq"]] +
        summary(fit.A)[[1]][[3,"Sum Sq"]] +
        summary(fit.A)[[1]][[4,"Sum Sq"]]

H <- df.T*SS.G/SS.T
H

p <- 1 - pchisq(H, summary(fit.A)[[1]][[1,"Df"]])
p


my.vector.hl.a <- c(my.vector.hl.a, p)


fit.B <- aov(align.B ~ A*B, data = my.data)
summary(fit.B)

df.T = summary(fit.B)[[1]][[1,"Df"]] +
        summary(fit.B)[[1]][[2,"Df"]] +
        summary(fit.B)[[1]][[3,"Df"]] +
        summary(fit.B)[[1]][[4,"Df"]]

SS.G = summary(fit.B)[[1]][[2,"Sum Sq"]]

SS.T = summary(fit.B)[[1]][[1,"Sum Sq"]] +
        summary(fit.B)[[1]][[2,"Sum Sq"]] +
        summary(fit.B)[[1]][[3,"Sum Sq"]] +
        summary(fit.B)[[1]][[4,"Sum Sq"]]

H <- df.T*SS.G/SS.T
H

p <- 1 - pchisq(H, summary(fit.B)[[1]][[2,"Df"]])
p

my.vector.hl.b <- c(my.vector.hl.b, p)



fit.inter <- aov(align.interact ~ A*B, data = my.data)
summary(fit.inter)

df.T = summary(fit.inter)[[1]][[1,"Df"]] +
        summary(fit.inter)[[1]][[2,"Df"]] +
        summary(fit.inter)[[1]][[3,"Df"]] +
        summary(fit.inter)[[1]][[4,"Df"]]

SS.G = summary(fit.inter)[[1]][[3,"Sum Sq"]]

SS.T = summary(fit.inter)[[1]][[1,"Sum Sq"]] +
  summary(fit.inter)[[1]][[2,"Sum Sq"]] +
  summary(fit.inter)[[1]][[3,"Sum Sq"]] +
  summary(fit.inter)[[1]][[4,"Sum Sq"]]

H <- df.T*SS.G/SS.T
H

p <- 1 - pchisq(H, summary(fit.inter)[[1]][[3,"Df"]])
p
```

```
#Create model and store p-values for rank transform test
#using Thompson statistic.

my.vector.hl.i <- c(my.vector.hl.i, p)

my.vector.rt.i2 <-
  c(my.vector.rt.i2,
        1 - pchisq(summary(aov.model.rt)[[1]][[3,"F value"]],
        summary(aov.model.rt)[[1]][[3,"Df"]]))

my.vector.rt.i3 <- c(my.vector.rt.i3,
        1 - pchisq(summary(fit.inter)[[1]][[3,"F value"]],
        summary(fit.inter)[[1]][[3,"Df"]]))

}
Print model effects, Type I error rates, and power.

print("The skew effect")
print(s)

print("The A effect:")
print(a)

print("The B effect:")
print(b)

print("The Interaction effect:")
print(c)

print("A p-value:")
print(NROW(my.vector.anova.a[my.vector.anova.a < 0.05]) /
        NROW(my.vector.anova.a))
print(NROW(my.vector.rt.a[my.vector.rt.a < 0.05]) /
        NROW(my.vector.rt.a))
```

```
print(NROW(my.vector.art.a[my.vector.art.a < 0.05]) /
        NROW(my.vector.art.a))
print(NROW(my.vector.hl.a[my.vector.hl.a < 0.05]) /
        NROW(my.vector.hl.a))

print("B p-value:")
print(NROW(my.vector.anova.b[my.vector.anova.b < 0.05]) /
        NROW(my.vector.anova.b))
print(NROW(my.vector.rt.b[my.vector.rt.b < 0.05]) /
        NROW(my.vector.rt.b))
print(NROW(my.vector.art.b[my.vector.art.b < 0.05]) /
        NROW(my.vector.art.b))
print(NROW(my.vector.hl.b[my.vector.hl.b < 0.05]) /
        NROW(my.vector.hl.b))

print("Interaction p-value:")
print(NROW(my.vector.anova.i[my.vector.anova.i < 0.05]) /
        NROW(my.vector.anova.i))
print(NROW(my.vector.rt.i[my.vector.rt.i < 0.05]) /
        NROW(my.vector.rt.i))
print(NROW(my.vector.art.i[my.vector.art.i < 0.05]) /
        NROW(my.vector.art.i))
print(NROW(my.vector.hl.i[my.vector.hl.i < 0.05]) /
        NROW(my.vector.hl.i))
print(NROW(my.vector.rt.i2[my.vector.rt.i2 < 0.05]) /
        NROW(my.vector.rt.i2))
print(NROW(my.vector.rt.i3[my.vector.rt.i3 < 0.05]) /
        NROW(my.vector.rt.i3))
}
}
}
}
```

APPENDIX B

CHAPTER 4 R CODE

```r
#Create empty vectors for storing p-values.

for (s in c(1, 1/2, 1/3, 1/4)) {

for (a in c(0, 0.0626754)) {
for (b in c(0, 0.0626754)) {
for (c in c(0, 0.0626754)) {

my.vector.anova.a <- vector()
my.vector.nsanova.a <- vector()
my.vector.ansanova.a <- vector()
my.vector.ansanovam.a <- vector()

my.vector.anova.b <- vector()
my.vector.nsanova.b <- vector()
my.vector.ansanova.b <- vector()
my.vector.ansanovam.b <- vector()

my.vector.anova.i <- vector()
my.vector.nsanova.i <- vector()
my.vector.ansanova.i <- vector()
my.vector.ansanovam.i <- vector()

my.vector.ans.i <- vector()

#Open the loop to perform the simulation.

for (i in 1:5000) {
```

```
mu = 0
beta1 = a
beta2 = b
beta3 = c

#Select cell sample size, distribution, and skew effect.

A = c(rep(c(-1), 1000), rep(c(1), 1000))
B = rep(c(rep(c(-1), 500), rep(c(1), 500)), 2)
e = ralaplace(2000, 0, 1, s)


y = mu + beta1*A + beta2*B + beta3*B*A + e

my.data = data.frame(cbind(A, B, y))

my.data$A <- factor(my.data$A)
my.data$B <- factor(my.data$B)

#Create output for ANOVA F test and store p-values.

aov.model = aov(y ~ A*B, data=my.data)

my.vector.anova.a <- c(my.vector.anova.a,
        summary(aov.model)[[1]][[1,"Pr(>F)"]])
my.vector.anova.b <- c(my.vector.anova.b,
        summary(aov.model)[[1]][[2,"Pr(>F)"]])
my.vector.anova.i <- c(my.vector.anova.i,
        summary(aov.model)[[1]][[3,"Pr(>F)"]])

# Create model and record p-values for rank transform test
# with a normal scores data transformation in place of rank.
my.data$y.ns <- qnorm(rank(my.data$y)/(NROW(my.data$y)+1))
```

```
aov.model.nsanova = aov(y.ns ~ A*B, data = my.data)

my.vector.nsanova.a <-
    c(my.vector.nsanova.a,
        summary(aov.model.nsanova)[[1]][[1,"Pr(>F)"]])
my.vector.nsanova.b <-
    c(my.vector.nsanova.b,
        summary(aov.model.nsanova)[[1]][[2,"Pr(>F)"]])
my.vector.nsanova.i <-
    c(my.vector.nsanova.i,
        summary(aov.model.nsanova)[[1]][[3,"Pr(>F)"]])

# Create model and record p-values for aligned rank transform test
# with a normal scores data transformation in place of rank.


art.model <- art(my.data$y ~ my.data$A * my.data$B)

align.interact <- art.model$aligned.ranks$`my.data$A:my.data$B`
align.A <- art.model$aligned.ranks$`my.data$A`
align.B <- art.model$aligned.ranks$`my.data$B`

ans.A <- qnorm(align.A/(NROW(align.A)+1))
ans.B <- qnorm(align.B/(NROW(align.B)+1))
ans.interact <- qnorm(align.interact/(NROW(align.interact)+1))


fit.A <- aov(ans.A ~ A*B, data = my.data)
my.vector.ansanova.a <- c(my.vector.ansanova.a,
        summary(fit.A)[[1]][[1,"Pr(>F)"]])

fit.B <- aov(ans.B ~ A*B, data = my.data)
my.vector.ansanova.b <- c(my.vector.ansanova.b,
        summary(fit.B)[[1]][[2,"Pr(>F)"]])

fit.interact <- aov(ans.interact ~ A*B, data = my.data)
my.vector.ansanova.i <-
  c(my.vector.ansanova.i, summary(fit.interact)[[1]][[3,"Pr(>F)"]])

# Calculate and record p-values for McSweeney test
# with a normal scores data transformation in place of rank.

df.T = summary(fit.A)[[1]][[1,"Df"]] +
        summary(fit.A)[[1]][[2,"Df"]] +
        summary(fit.A)[[1]][[3,"Df"]] +
        summary(fit.A)[[1]][[4,"Df"]]

SS.G = summary(fit.A)[[1]][[1,"Sum Sq"]]

SS.T = summary(fit.A)[[1]][[1,"Sum Sq"]] +
        summary(fit.A)[[1]][[2,"Sum Sq"]] +
        summary(fit.A)[[1]][[3,"Sum Sq"]] +
        summary(fit.A)[[1]][[4,"Sum Sq"]]

H <- df.T*SS.G/SS.T
H

p <- 1 - pchisq(H, summary(fit.A)[[1]][[1,"Df"]])
p

my.vector.ansanovam.a <- c(my.vector.ansanovam.a, p)


df.T = summary(fit.B)[[1]][[1,"Df"]] +
        summary(fit.B)[[1]][[2,"Df"]] +
        summary(fit.B)[[1]][[3,"Df"]] +
        summary(fit.B)[[1]][[4,"Df"]]
```

```
SS.G = summary(fit.B)[[1]][[2,"Sum Sq"]]

SS.T = summary(fit.B)[[1]][[1,"Sum Sq"]] +
        summary(fit.B)[[1]][[2,"Sum Sq"]] +
        summary(fit.B)[[1]][[3,"Sum Sq"]] +
        summary(fit.B)[[1]][[4,"Sum Sq"]]

H <- df.T*SS.G/SS.T
H

p <- 1 - pchisq(H, summary(fit.B)[[1]][[2,"Df"]])
p

my.vector.ansanovam.b <- c(my.vector.ansanovam.b, p)


df.T = summary(fit.interact)[[1]][[1,"Df"]] +
        summary(fit.interact)[[1]][[2,"Df"]] +
        summary(fit.interact)[[1]][[3,"Df"]] +
        summary(fit.interact)[[1]][[4,"Df"]]

SS.G = summary(fit.interact)[[1]][[3,"Sum Sq"]]

SS.T = summary(fit.interact)[[1]][[1,"Sum Sq"]] +
        summary(fit.interact)[[1]][[2,"Sum Sq"]] +
        summary(fit.interact)[[1]][[3,"Sum Sq"]] +
        summary(fit.interact)[[1]][[4,"Sum Sq"]]

H <- df.T*SS.G/SS.T
H

p <- 1 - pchisq(H, summary(fit.interact)[[1]][[3,"Df"]])
p
```

```
my.vector.ansanovam.i <- c(my.vector.ansanovam.i, p)

ns.factor = c(rep(c(2), NROW(align.interact)/4), rep(c(1),
        NROW(align.interact)/2), rep(c(2), NROW(align.interact)/4))

# record p-values for Van der Waerden test of interaction.
# This is run only as a test as it should match the
# McSweeney test with a Van der Waerden normal scores
# transformation in place of the rank transformation.

my.data.2 <- data.frame(cbind(align.interact, ns.factor))
my.data.2$ns.factor <- factor(my.data.2$ns.factor)
nsi.test <- VanWaerdenTest(align.interact ~ ns.factor,
        data = my.data.2)

my.vector.ans.i <- c(my.vector.ans.i, nsi.test$p.value)
}

# Print model effects, Type I error rates, and power.

print("The skew effect")
print(s)

print("The A effect:")
print(a)

print("The B effect:")
print(b)

print("The Interaction effect:")
print(c)

print("A p-value:")
print(NROW(my.vector.anova.a[my.vector.anova.a < 0.05]) /
        NROW(my.vector.anova.a))
print(NROW(my.vector.nsanova.a[my.vector.nsanova.a < 0.05]) /
        NROW(my.vector.nsanova.a))
print(NROW(my.vector.ansanova.a[my.vector.ansanova.a < 0.05]) /
        NROW(my.vector.ansanova.a))
print(NROW(my.vector.ansanovam.a[my.vector.ansanovam.a < 0.05]) /
        NROW(my.vector.ansanovam.a))

print("B p-value:")
print(NROW(my.vector.anova.b[my.vector.anova.b < 0.05]) /
        NROW(my.vector.anova.b))
print(NROW(my.vector.nsanova.b[my.vector.nsanova.b < 0.05]) /
        NROW(my.vector.nsanova.b))
print(NROW(my.vector.ansanova.b[my.vector.ansanova.b < 0.05]) /
        NROW(my.vector.ansanova.b))
print(NROW(my.vector.ansanovam.b[my.vector.ansanovam.b < 0.05]) /
        NROW(my.vector.ansanovam.b))

print("Interaction p-value:")
print(NROW(my.vector.anova.i[my.vector.anova.i < 0.05]) /
        NROW(my.vector.anova.i))
print(NROW(my.vector.nsanova.i[my.vector.nsanova.i < 0.05]) /
        NROW(my.vector.nsanova.i))
print(NROW(my.vector.ansanova.i[my.vector.ansanova.i < 0.05]) /
        NROW(my.vector.ansanova.i))
print(NROW(my.vector.ansanovam.i[my.vector.ansanovam.i < 0.05]) /
        NROW(my.vector.ansanovam.i))
print(NROW(my.vector.ans.i[my.vector.ans.i < 0.05]) /
        NROW(my.vector.ans.i))
}
}
}
}
```

APPENDIX C

CHAPTER 5 EXAMPLE DATA

Factor A Level	Factor B Level	Response y
-1	-1	3.115813
-1	-1	0.664418
-1	-1	0.804796
-1	-1	1.065934
-1	-1	1.751445
-1	-1	0.643708
-1	-1	4.83077
-1	-1	11.93704
-1	-1	2.290986
-1	-1	0.556997
-1	-1	0.178869
-1	-1	0.647733
-1	-1	1.233566
-1	-1	0.200879
-1	-1	2.652717
-1	1	5.064044
-1	1	1.486174
-1	1	0.872977
-1	1	0.753693
-1	1	2.951995
-1	1	-0.38633
-1	1	2.375272
-1	1	1.79895
-1	1	0.862123
-1	1	2.399837
-1	1	0.112025
-1	1	8.202698
-1	1	0.813087
-1	1	0.479686
-1	1	-0.46117
1	-1	-0.00509
1	-1	0.108623
1	-1	7.207883
1	-1	0.115328

1	-1	-0.4595
1	-1	-0.25915
1	-1	0.790421
1	-1	-0.40838
1	-1	6.560586
1	-1	-0.82946
1	-1	-0.71608
1	-1	-0.49341
1	-1	-0.31225
1	-1	1.333793
1	-1	6.231697
1	1	2.649382
1	1	4.484691
1	1	5.423211
1	1	2.771467
1	1	3.097686
1	1	2.309422
1	1	0.959578
1	1	1.224494
1	1	3.382787
1	1	4.397402
1	1	1.185864
1	1	2.589569
1	1	2.849444
1	1	2.968131
1	1	2.59024