THE APPLICATION OF CONFIGURAL FREQUENCY ANALYSIS USING STIRLING'S FORMULA TO SINGLE CASE DESIGNS

A Dissertation

by

MARC AARON PATIENCE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Daniel F. Brossart
Committee Members,	Kimberly J. Vannest
	William A. Rae
	Steve Balsis
Head of Department,	Victor L. Willson

August 2015

Major Subject: Counseling Psychology

ABSTRACT

Single case experimental designs (SCEDs) have found their place among a range of fields including psychology, education, and medicine. SCEDs provide rigorous experimental evaluation of treatment effects. Currently, SCED evaluation methods, such as visual analysis and effect size estimation statistics, provide evidence for determining treatment effects. Although useful, best practices for SCED data analysis are debated.

Configural frequency analysis (CFA) is introduced as a statistical method for the analysis of data from SCEDs. This research compared CFA statistical significance results to non-overlap SCED analysis methods that do and do not provide for statistical significance, as well as visual analysis methods. As there are currently no agreed upon best methods for the analysis of data obtained by SCEDs, it was important to explore additional statistical methods to aid researchers choosing to utilize SCEDs.

CFA was compared to 5 non-overlap treatment effect evaluation methods that provide statistical significance and effect size values, 5 methods that only provide for effect size values, and visual analysis performed by 6 doctoral (PhD) students trained in SCEDs. A review of 23 years of *The Journal of Behavior Therapy* and *The Journal of Behavior Modification* resulted in 168 SCED data sets for comparison methods. Graphs were analyzed using each non-overlap effect size procedure as well as CFA.

Results suggest that CFA aligned well with existing statistical significance calculations. Visual analysis appeared to align with simple non-overlap effect size methods rather than with CFA calculations, hinting at the importance of including

statistical significance when evaluating treatment effects. Overall, this research found that CFA performed well when compared to other SCED data analysis techniques.

ACKNOWLEDGMENTS

I would like to thank my committee, Dr. Brossart, Dr. Vannest, Dr. Rae, and Dr. Balsis, for their valued support and dedication to this process. I have been humbled by the magnitude of a dissertation, and made better by the support and attention by each of my committee members. I would also like to thank my resident training director at Joint Base San Antonio Lackland, Dr. Hryshko-Mullen. Her ever-present expectation of excellence, wealth of support, and investment in my success was paramount to the conclusion of my doctoral studies.

Most importantly, I would like to thank my wife Brenda Gámez-Patience. I have never felt such love in my life. Her confidence in my ability to succeed far surpasses my own, and continues to serve as a motivational fire that simply will not die. She has shown me that there is truly no evil in this world, only places where love has yet to touch, as she strives every day to spread her love and positive energy to all corners of our world. I hope my efforts may reflect the type of person she has encouraged me to be.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

Page

CHAPTER I

INTRODUCTION

Single case experimental designs (SCEDs) are enjoying a renewed interest given their flexible utility and potential for cost effective research projects in many areas of applied psychology. As SCEDs gain recognition and become commonplace, questions must be answered regarding statistical analysis tools used to explore treatment effects. Presently, although statistical analyses are abundant for SCEDs, it is unclear as to which methods are best to use in what data scenarios (Kratochwill et al., 2010).

A few suggestions for moving forward are proposed here. One is to continue to evaluate existing statistical methods and determine their strengths and weaknesses in given situations. This may help to determine which methods perform best across different scenarios, and help researchers understand why methods may underperform in others. Fortunately, many have taken this challenge to heart. Article after article have been published to explore the interactions between treatment effect evaluation tools and unique SCED data presentations (e.g., Parker, Vannest, Davis, & Sauber, 2010; Parker, et al., 2005; Smith, Vannest, & Davis, 2011, etc.) Although efforts continue, the question remains as to which methods are best or most appropriate across various data presentations.

In a similar manner one may examine other statistical methods that may show promise or develop new methods for analyzing single-case data. This current study

examines an exploratory method commonly applied to randomized control trials, and attempts to adapt it to SCED data analysis.

Configural frequency analysis (CFA; Von Eye, 1990), using Stirling's Formula, may be able to analyze a variety of SCED data presentations, and in the process provide statistical significance values, which assists one in evaluating treatment effects. The potential benefits of using this tool include the ability to calculate multiple statistical significance values per data set and the ability to adapt and be applied using a variety of non-overlap effect size techniques. Non-overlap methods may characterize data in different ways depending on the method used, and CFA may have the potential to be applied to each of these different data configurations such as changing 2x2 graphical representations across AB phase transitions. Much has been done with CFA applied to exploratory research and large group designs, but a literature review suggested that CFA had yet to be extended to SCEDs.

To start the process, this research applied CFA to over 150 SCED data sets, and used 10 different treatment effect evaluation methods, including visual analysis for comparison. By using numerous data sets and comparison methods, it is possible to determine how CFA can be applied to SCED data. The first study compared CFA to non-overlap effect size estimation tools that estimate statistical significance, the second study compared CFA to non-overlap effect size tools that do not estimate statistical significance, and the third study compared CFA to visual analysis techniques. By comparing CFA to each of these groups of methods, the goal was to present a comprehensive introduction of CFA to SCEDs.

CHAPTER II

MANUSCRIPT #1: THE APPLICATION OF CONFIGURAL FREQUENCY ANALYSIS USING STIRLING'S FORMULA TO SINGLE-CASE DESIGNS: A COMPARISON TO NON-OVERLAP INDICES THAT PROVIDE STATISTICAL SIGNIFICANCE

Introduction

SCEDs provide a rigorous method for the statistical evaluation of treatment effects in settings where resources (i.e. time, money, participants) may be limited. Researchers have used visual analysis, regression-based techniques, and both parametric and nonparametric methods of effect size estimation to aid in the evaluation of treatment effects from SCEDs.

Although SCEDs have found their place in a wide range of settings (psychology, education, medicine), presently there are no agreed upon best methods for the analysis of SCED data (Kratochwill et al., 2010). Numerous researchers have worked to clarify the most appropriate tools to use in specific scenarios (e.g., Parker, Vannest, Davis, & Sauber, 2010; Parker, et al., 2005; Smith, Vannest, & Davis, 2011), however, the panel from What Works Clearinghouse (Kratochwill et al., 2010) indicated that all current statistical methods for SCED data analysis have limitations, especially when used as stand alone indicators of treatment effect rather than paired with other methods.

The current study set out to evaluate a method used primarily in large group research and randomized control trials. Configural frequency analysis (CFA; von Eye, 1990) is a method used to determine the statistical significance by comparing observed

data frequencies to expected data frequencies across cells of a frequency data matrix. In a simple AB design, CFA may be applied to SCEDs by treating graphically represented data as a 2x2 matrix, which can be created in various ways. One way of creating these 2x2 matrices is to use a non-overlap method and use the graphic representation of this method for the 2x2 matrix. Since there are many non-overlap methods, it is important to first determine which non-overlap methods will be used to create cells for the CFA.

Determining a Family of Methods for Comparison Purposes

A number of regression, parametric, and non-parametric techniques have been developed for and applied to SCEDs. When developing measures, researchers must recognize that SCEDs present unique constraints due to their typical short time series data. To account for these constraints, Solanas, Rumen, and Patrick (2011) proposed the following guidelines:

> A 'useful' effect size index needs to meet the following criteria: (a) to represent correctly the true data characteristics, (b) to offer valuable and easily interpretable information, and (c) to be easily applicable by researchers with scarce statistical expertise. (pg. 200)

Regression based methods may appear ideal when selecting a method that would provide the most evidence of a treatment effect. These methods allow the modeling of patterns such as linear and quadratic trend, level change, and differences in slope from phase A to phase B (Huitema & McKean, 2000). For example, by analyzing changes in

trend, a quantification of treatment effect can be calculated by the difference in trend values of phase A data and trend values of phase B data. Although regression based methods have many strengths, researchers have presented significant criticisms of parametric methods when applied to the short time series data typical of single case designs.

First, regression-based techniques have been found to perform unsatisfactorily with respect to criterion A proposed by Solanas, Rumen, and Patrick (2010) (Beretvas & Chung, 2008; Brossart, Parker, Olson, & Mahadevan, 2006; Parker & Brossart, 2003). Regression methods tend to yield large effect size estimations when applied to SCED data (Campbell, 2004). Parker, Vannest, and Brown (2009) found evidence that the large effect size estimations provided by regression methods may not accurately represent the data, and may be artificial inflations rather than accurate depictions of the data.

> A recent examination and re-analysis of over 150 published single-case research data series found less than half with large effects (according to both visual and statistical analyses), and more than one quarter showed small or debatable results. (pg. 135)

This indicates that the potentially larger effect size estimations obtained when using regression methods likely do not accurately represent the data. Obtaining large or inflated effect size estimates suggest that regression techniques may lack the ability to

sufficiently differentiate between subtle yet important changes in singe-case data. This makes the interpretation of results difficult when using regression methods, causing regression methods to fall short of criterion B.

Parametric methods such as ordinary least squares regression have a long history of use in large N studies, and show great flexibility and power when applied to SCED data (Allison & Gorman, 1993; Busk & Serlin, 1992; Parker & Brossart, 2003). As such, it is no wonder why researchers continue trying to find ways to effectively apply parametric methods to SCEDs. However, parametric methods as they currently stand may not be appropriate for SCED data. Parametric methods require a normal score distribution, constant variance, interval level measurement (Vickers, 2005), and assume statistical independence between the dependent and independent variable. SCED data *rarely* meet the assumptions of constant variance, normality, independence between variables, and the scaling assumption of at least interval-level data (Kutner, Nachtsheim & Neter, 2004).

For the purpose of this study, regression and parametric methods were determined to have too many limitations to be used for initial comparison. As such, this study focused on comparisons using established nonparametric overlap methods that provide statistical significance values. Not only were these methods appropriate based on guidelines from Solanas, Rumen, and Patrick (2010), but as will be seen, non-parametric overlap methods naturally arrange the data in a manner appropriate for CFA. This study used the extended celeration line (ECL; White & Haring, 1980), also known as split middle line, non-overlap of all pairs (NAP), Tau-A, Tau-B, and Tau-U. These

methods are not influenced by the same shortcomings of parametric or regression based methods, and as such will serve as appropriate comparison methods to begin determining whether or not CFA has a place in SCED data analysis. Existing papers present the techniques and one paper reviews them in more depth (Parker, Vannest, & Davis, 2010) so we will not attempt to explain each method in depth here. Rather, we will demonstrate how each forms a four quadrant or 2x2 matrix that naturally applies to CFA.

Extended Celeration Line (ECL)

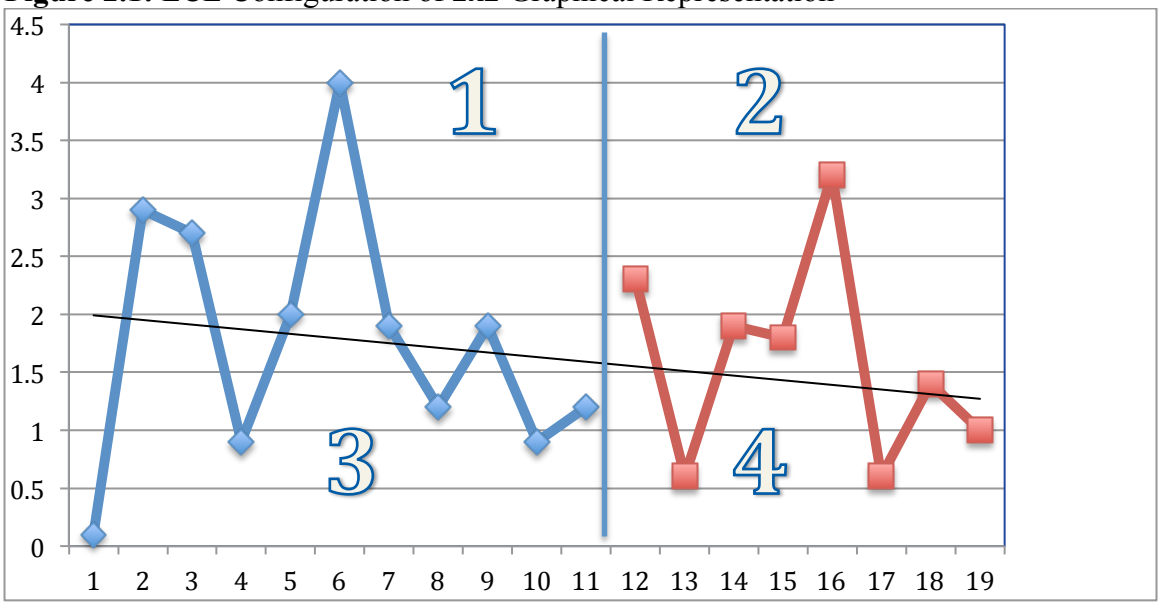
Also known as "split middle line", ECL is completed by extending a median slope, or other trend line such as a linear trend fitted using a computer program, calculated from Phase A data, and extending this line into Phase B (White & Haring, 1980; Parker, Vannest, & Davis, 2010). Once extended, a non-overlap effect size is calculated by taking the number of data points in phase B over the line (or under, depending on the desired/expected direction of a treatment effect) and dividing that by the total number of data points in phase B. Due to the expectation that the line splits the Phase A data in half, with close to 50% above and 50% below the line in phase A, a "no treatment effect" result would be indicated by obtaining close to .50 effect size from the calculation.

Based on these calculations and expectations, we can then compare the obtained phase B ratio with the expected ratio (.50) using a statistical test of one proportion (Parker, Vannest, & Davis, 2010). For example, if we were to use ECL on Figure 2.1 displayed below, utilizing a median trend line on the phase A data, we obtain an ECL ratio from phase B data of 5/8 or .625. From the test of one proportion, we obtain a *p*

value of .4621, indicating a statistically non-significant result. For readers who are familiar with SCEDs and visual analysis, this result should resonate with some face validity, as a quick visual analysis of these data suggests a minimal treatment effect from phase A to phase B.

Per CFA, two quadrants are created in phase A (one above the ECL, one below) and two quadrants in phase B. It will be made clear in the section covering CFA why this 2x2 system is ideal. For now, it is important to recognize the bivariate system that is created, either having the data point above the line or below the line.

Figure 2.1: ECL Configuration of 2x2 Graphical Representation



Non-Overlap of All Pairs (NAP)

Introduced in 2009 by Parker & Vannest (2009), NAP is a nonparametric technique which compares each phase A data point with each phase B data point. By

counting the number of non-overlapping pairs (either by hand or through a statistical tool as used in this study), and subtracting this from the total possible pairs, we obtain a "non-overlap count" (Parker & Vannest, 2009). Total possible pairs is calculated by multiplying the number of data points in phase A by the number of data points in phase B. To obtain an effect size, NAP takes 1 minus the number of non-overlapping pairs (X) subtracted from the total possible pairs (Y) divided by the total possible pairs (Y). Represented mathematically, the equation is as follows.: 1 – (Y-X/Y). For our example graph, Figure 2.1, total possible pairs equals 88 (11x8), non-overlapping pairs equals 39, and NAP equals 45%. Statistical significance was built into the online effect size calculation tool utilized for the purposes of this study (Vannest, Parker, & Gonen, 2011). As NAP compares each A phase data point to each B phase data point, NAP does not create a 2x2 configuration that can be used with CFA. Later, we will introduce a measure of central tendency used to create a 2x2 configuration and supplement for instances such as this when the non-overlap method does not create a 2x2 configuration.

Kendall's Tau

A non-parametric correlation coefficient that measures "the percent of data that improve over time," (Parker, et al., 2011) Kendall's Tau is considered a rank order correlation. Kendall's Tau, similar to NAP, does not naturally create a 2x2 quadrant system across the A and B phase. Rather, a decision matrix such as that which is published in Brossart, Vannest, Davis & Patience, (2014), is the method for hand calculation of Tau. Although Tau doesn't naturally align with CFA by providing a 2x2 matrix configuration, it is a frequently utilized method of SCED data evaluation,

provides a statistical significance test using the S distribution (Parker, Vannest, Davis, & Sauber, 2011), and deserves to be compared to the performance of new methods for analyzing single-case data. As the Tau analysis is unable to be directly converted into a 2x2 table, again a robust mean based on the A phase data and extended horizontally through phase B (similar to ECL) was used as a comparison method. Further discussion on the selection of the robust mean is presented later on.

Variations of Tau used in this study include Tau-A, Tau-B, and Tau-U. Tau-A accounts for simple non-overlap comparisons in a time-forward direction (Parker, Vannest, & Davis, 2011). While Tau-A does not account for ties between phases, Tau-B is adjusted to account for ties in data rankings. Tau-U is a combination of Kendall's Tau and the Mann-Whittney U statistical analysis (Parker, Vannest, Davis, & Sauber, 2011) and controls for phase A trend. It is designed to compensate for regression based analyses, which frequently violate the data assumptions previously discussed, and to compensate for small statistical power regularly found within non-overlap methods (Parker & Vannest, 2009). Readers interested in the mathematical calculations are encouraged to read articles and books published on these topics, as these go beyond the scope of the comparison purposes of this study.

Configural Frequency Analysis (CFA)

CFA may have the ability to address many data characteristics (i.e. trend, variability, autocorrelation, non-constant variance, non-normality, non-linearity of relationship, failing to meet the scaling assumption of at least an interval-level scale) that are inherent with real behavioral data acquired through SCEDs. Prior to determining if

CFA can address all of these data characteristics, however, CFA must be introduced to SCED data to explore basic application considerations. What follows is a conceptual overview of CFA.

CFA is a method for the analysis of bi- or multivariate cross-classifications of categorical variables (Von Eye, 2002). CFA looks for statistically significant effects at the level of individual cells, or groups of cells, in a table. Single-case designs can be fitted with contingency tables using an appropriate method for generating the matrix (e.g., 2x2, 2x2x2, 3x3x3, etc.), and can be designed in many ways to account for features of single-case data such as phase (i.e. mean, median, etc), or trend (i.e. linear, curvilinear, quadratic, etc.). Unlike the effect size estimators described before, CFA is concerned with identifying unlikely configurations of data at the cell level that statistically contradict the null hypothesis of no treatment effect. This produces independent statistical significance results for each B phase quadrant. CFA does not produce an effect size, and should not be considered a replacement for effect size estimation techniques, but rather as a complimentary statistical tool that estimates statistical significance.

When using CFA applied to SCED data, a null hypothesis would constitute a configuration of cells that statistically contains as many data points as expected if the treatment produced no effect. When a configuration contradicts the null hypothesis because they contain more cases than expected, this is called a *type-constituting* configuration. If CFA identifies a configuration that contains fewer cases than expected, this is termed an *anti-type*. The researcher determines what configuration is expected by

establishing a base model. The base model takes into account those effects that are NOT of interest to the researcher, such as an A phase that is visually the same as a B phase, indicating that the treatment had no effect.

In other words, a null hypothesis when applying CFA to SCEDs would read as follows: "The upper A phase quadrant will contain the same percentage of data as the upper B phase quadrant." Equivalently, "the lower A phase quadrant will contain the same percentage of data points as the lower B phase quadrant." The flexibility of CFA applied to SCEDs is achieved through the quadrant system, as the method for producing the quadrants can be determined by the researcher to best fit the design of the study or the characteristics of the data.

There are five decision-making steps researchers take when applying CFA (Von Eye, 2007). First, a researcher selects a base model and estimation of expected frequencies. This is a chance model that indicates the probability a particular configuration of data is expected to occur. The base model represents configurations that are NOT of interest to the researcher. Within SCEDs, the base model would reflect no treatment effect, which would look like an equivalent, or mirrored, data presentation between the baseline (A) phase and the treatment (B) phase. This would reflect the treatment having no effect, and analysis would likely show statistically equivalent (i.e. non-significant difference) A and B phase data presentations.

The second step of CFA involves selecting a concept of deviation from independence (Von Eye, 2007). For example, if the treatment is independent from the measured behavior, a researcher would expect to see no statistical difference between

upper baseline and upper treatment quadrants, as well as no statistical difference between the lower baseline and treatment quadrants. Deviation from independence would suggest a treatment phase that statistically differs from the baseline phase, suggesting that the treatment is correlated with the dependent variable. This is akin to setting an alpha (α) level, and can be measured through many statistical measures. For example, if the base model proposes variable independence, that is, completely unrelated variables, than deviation from independence can be measured using a marginal-dependent measure (Goodman, 1991) such as Pearson's X^2's Φ -coefficient which measures the strength of association between two dichotomous variables. Larger data sets over more than two phases in future research may find the Φ coefficient an appropriate measure of statistical independence.

Third, the researcher selects a significance test to determine the probability (statistical significance) that types or antitypes do not exist. Many tests have been proposed (von Eye, 2002), but simulation studies have shown that none of these tests outperforms the other tests under all of the examined conditions (Indurkhya & von Eye, 2000). Simulations suggest that statistical analyses such as Pearson's X^2, the z test, and Fisher's exact binomial test perform equally well under many conditions (von Eye, 2002). For the purposes of this study, we will use Stirling's approximation of the binomial (Bergman, & von Eye, 1987). This approximation is practically as powerful as Fisher's exact test, but is simple to calculate, rendering it useful to researchers.

Comparably, Fisher's exact binomial test is considered difficult to calculate in large and small sample sizes such as those observed in most single-case studies. Some

alternative tests require each cell within a configuration to have greater than five data points (i.e. z test), which can lead to distortions in the search for antitypes and/or require extensive tail probability calculations for extreme alpha levels (e.g., the F approximation of binomial tests; Von Eye, 1990). Stirling's formula meets the ideal criteria noted for SCED data analysis by maintaining simplicity in calculation, maintaining utility in small data sets, and not requiring constant variance, normality, linearity of relationship, independence between variables, and the scaling assumption of at least interval-level data. Stirling's formula is presented in formula 2.1, and placed in terms of single-case data, where *n* equals the total number of data points within A and B phases, *o* = the number of observed data points in the quadrant of interest, *p* is the expected frequency of data points divided by n (*p*=e/n); and *q*=1-*p*.

Formula 2.1: Approximation of the binomial test using Stirling's formula

$$\hat{B}'^{(o)} = \left(\frac{n}{2\pi o(n-o)}\right)^{\frac{1}{2}} \left(\frac{np}{o}\right)^{o} \left(\frac{nq}{n-o}\right)^{n-o}$$

Fourth, a Bonferroni correction is performed to protect the significance level, alpha (α). This is frequently done when a large number of tests are conducted, as in exploratory analysis and randomized control trials. As testing SCED phases will require two comparisons (one for the upper A and B quadrants, and one for the lower), this procedure was not necessary in this study. Typically correcting for multiple tests is important because CFA is frequently used in exploratory analysis which often uses many tests.

Finally, an interpretation of types and antitypes is performed (von Eye, Mair, & Mun, 2010). This step is only performed if CFA results in a statistically significant finding in one cell. If this is the case, the researcher then determines if the statistically significant result is a type or an antitype. First, some basic information about the study needs to be available. For example, one intervention may result in a student responding more frequently in a desirable manner during a token economy implementation (i.e. treatment phase) than when positive behavior is not reinforced (i.e. baseline). If the purpose of the study was to increase behavior, and the increase was statistically significant, this would constitute a type. However, if the purpose was to increase behavior, and the intervention resulted in a statistically significant decrease, this would be considered an anti-type.

In SCEDs, types and anti-types have different meanings when an individual may be responding more often to the selected criteria after intervention (type) or less often (anti-type). As we will be focused on individual, published behavioral data, we will be focused on the detection of statistically significant changes and less concerned with qualitative interpretations (i.e. types or antitypes). Determining types and anti-types would require an understanding of each SCED published article utilized for analysis, which is beyond the scope of this exploratory study to determine if CFA can be applied to SCEDs.

Robust Means for Selecting the Intercept Line for 2x2 Data

For exploratory purposes, and to have a comparison method for non-overlap effect size estimation methods that do not naturally create a 2x2 table (i.e. Tau and

NAP), an all-purpose method was developed. This study used a horizontal line calculated using a measure of central tendency from data in phase A, extended across phase B to create the 2x2 contingency table necessary to perform CFA.

A common measure of central tendency is the mean. This measure would add together all data within a phase (i.e, phase A) and divide the sum by the total number of data points in that phase. Although simple to calculate, the mean is not a robust measure, and slight changes in a distribution can have a large effect (Wilcox, 2012). It may be beneficial when using behavioral data to utilize a robust measure to determine an estimate of central tendency, as Brossart, Parker, & Castillo (2011) found 21.3% data sets of a series of single-case data had four or more outliers.

When choosing between robust measures, the breakdown point can help determine how much influence outliers may have (Donoho & Gasko, 1992). The breakdown point is the proportion of extreme values introduced into a sample to cause an arbitrarily large result. The mean has a breakdown point of 0%. A single large datum point can throw the mean off. The median, however, has a breakdown point of 50%. 50% is the highest breakdown point possible. If more than half of the observations are contaminated, it is not possible to distinguish between the underlying distribution and the contaminating distribution (He, Simpson, & Portnoy, 1990). As the median is simple to calculate by hand in small data sets, it lends itself quite well to single-case data, as used in existing non-overlap methods (i.e, Percentage Exceeding the Median, PEM; Ma, 2006).

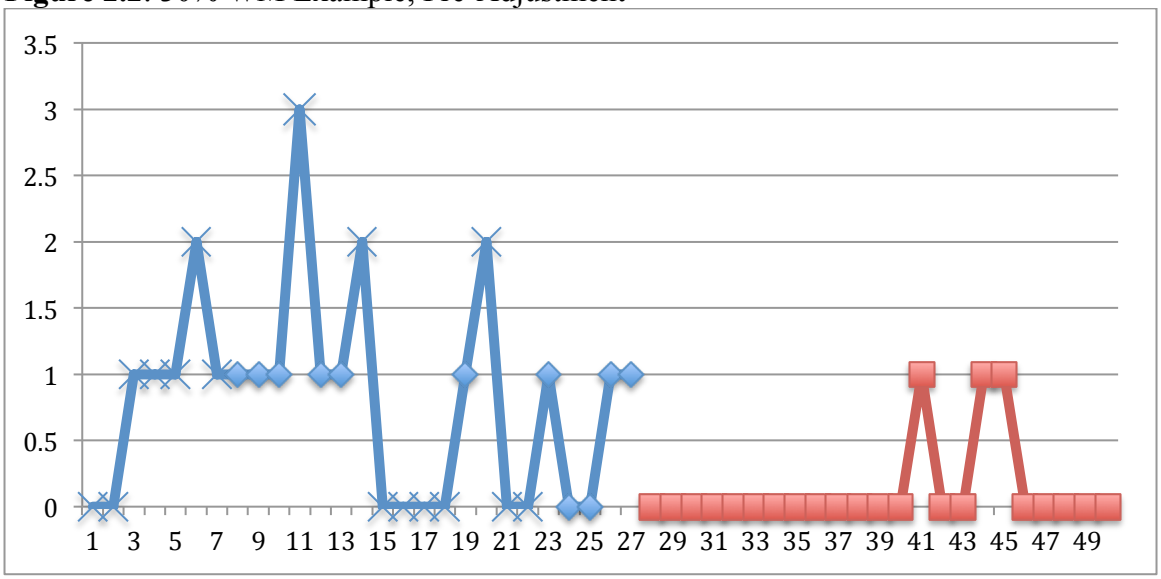
Because the tails of a distribution can overly influence the value of the mean, one strategy for addressing this is to give less weight to values at the tails and, consequently, more to those central to the distribution. This can be done by *Winsorizing* the distribution, and obtaining a Winsorized Mean (WM; Staudte, & Sheather, 1990; Wilcox, & Keselman, 2003). The WM gives more weight to the central portion of a distribution by transforming the data at the tails of the distribution. This is done by modifying one or more data points at the end of the tails of the distribution to the next highest/lowest value within the distribution. To determine the number of points to modify, the WM breakdown point is chosen by the researcher, and consequently may be placed at the maximum percentage, 50% (i.e. modifying 25% of the data on each end of the distribution).

This has costs and benefits. Although confident that a robust measure has been applied, a 50% Winsorization takes 50% of the data points from the data set, and shifts them toward the nearest central distribution point. With behavioral data, a 50% Winsorization may imply that the researcher considers 50% of the measured behaviors to be exaggerated, i.e., outliers. This is a difficult claim to make. More commonly, the decision is based on the distribution or visual representation of the data. For single-case data, a researcher may identify one or two data points that clearly extend beyond a reasonable indication of central tendency, and apply a Winzoriation to adjust. Winsorization may be useful in single-case behavioral data, and is relatively easy to calculate by hand.

Presented here is an example of a breakdown point applied to SCED data. Figure 2.2 represents an SCED graph with one A and one B phase. In phase A, there are 27 data points, with a mean value of .814. If using a 30% WM to create 2x2 configurations, 30% of 27 data points require adjusting to find the 30% WM. As such, 8 data points from the "tails" or extremes of the A phase distribution will be moved to the next closest data point. As such, 8 data points from each "tail" or the extreme values of the A phase distribution will be moved to the next closest data point. This results in the lowest 8 and highest 8 from the number set being moved to the next closest entry. Data set 2.1 represents the data in phase A ordered from smallest to largest values with the bold numbers representing the 8 lowest and 8 highest data points. Figure 2.2 graphically represents this data set, pre 30% WM adjustment, where data points marked with an "X" represent bolded data from data set 2.1.

Data Set 2.1: Figure 2.2 A Phase Data
(**0,0,0,0,0,0,0,0**,0,0,1,1,1,1,1,1,1,1,1,**1,1,1,1,2,2,2,3)**

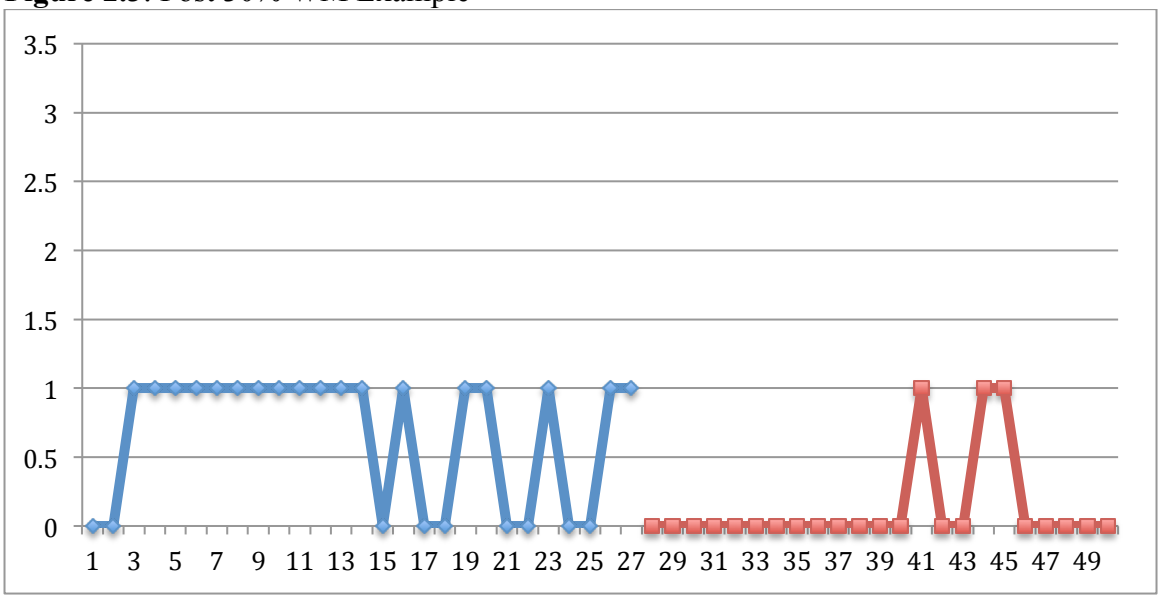
Figure 2.2: 30% WM Example, Pre-Adjustment



To account for the two-tailed distribution, data points are adjust from both the upper and lower extremes. The X on the graph above indicates these data points. Each of these data points is moved to the next closest value. For the upper extreme, one data point is moved from 3 to 1, while the remaining 3 upper extreme points are moved from 2 to 1. The resulting 30% WM A phase data with effected numbers bolded is presented in data set 2.2, and graphical representation of the 30% WM adjusted data is presented in Figure 2.3.

Data Set 2.2: Figure 2.3 A Phase Data Post 30% WM
(**0,0,0,0,0,0,0,0,**0,0,1,1,1,1,1,1,1,1,1,**1,1,1,1,1,1,1,1,1**)

Figure 2.3: Post 30% WM Example



As seen in Figure 2.3, the A phase data has moved to a more central location, without removing the data point completely. The resulting mean of the Post 30% WM data is .630, compared to the simple mean of .814. This represents a more appropriate central measure of the data, as it is less effected by outliers, but continues to consider the data relevant as it does not completely eliminate the data point.

Based on these considerations, a 30% WM was selected to create 2x2 configurations for NAP, Tau-A, B, and U comparison purposes, as well as comparisons across ECL. This allows for a reduction in the influence of outliers, without completely removing the behavioral validity of variable data. In addition, the calculations for a 30% WM are quick, simple, can be performed either by hand or by computer, and can be utilized by future researchers interested in CFA as a statistical analysis tool.

Illustrative Articles

This study was completed using published SCED data. A review of articles from *The Journal of Behavior Therapy* and *The Journal of Behavior Modification* published between 1990 and 2013 was performed. Graphs were selected based on the following criteria. First, clear A and B phases were necessary. Only one AB phase was necessary, as this study only compared the first A phase to the first B phase. To explore the utilization of CFA outside of "best case scenarios", this study stretched the minimal standards set by What Works Clearinghouse (Kratochwill et al, 2010). As such, a minimum of three data points in phase A and 3 data points in phase B were necessary to be considered, rather than 5 as recommended. The rationale was to see if CFA would function in instances where minimal data was available. When multiple data sets were present on one graph, all were separately analyzed and used for this study.

147 graphs were pulled for the purposes of this study. This resulted in 168 data sets, as some graphs had multiple participants/data sets. The average number of data points in phase A, B, and in total was 7.24, 17.81, and 25.05 respectfully. The median number of data points in phase A, B, and in total was 4, 9.5, and 15 respectfully.

Data Extraction and Analysis

Graphs were extracted, digitized, and analyzed across multiple steps. First, graphs from published studies were saved in PDF form. Each graph was copied from the published articles using the "Snipping Tool" from Microsoft Office version 2010. This digital snapshot tool captured just the image of the graphic data, which was then uploaded into the digitizing program, GraphClick for Apple OS X. Each data set was

uploaded individually, and the scale for the X and the Y-axis were set to match the axis values within the graph. Finally, each data point was specified by "point and click" to assure concordance with the original study data.

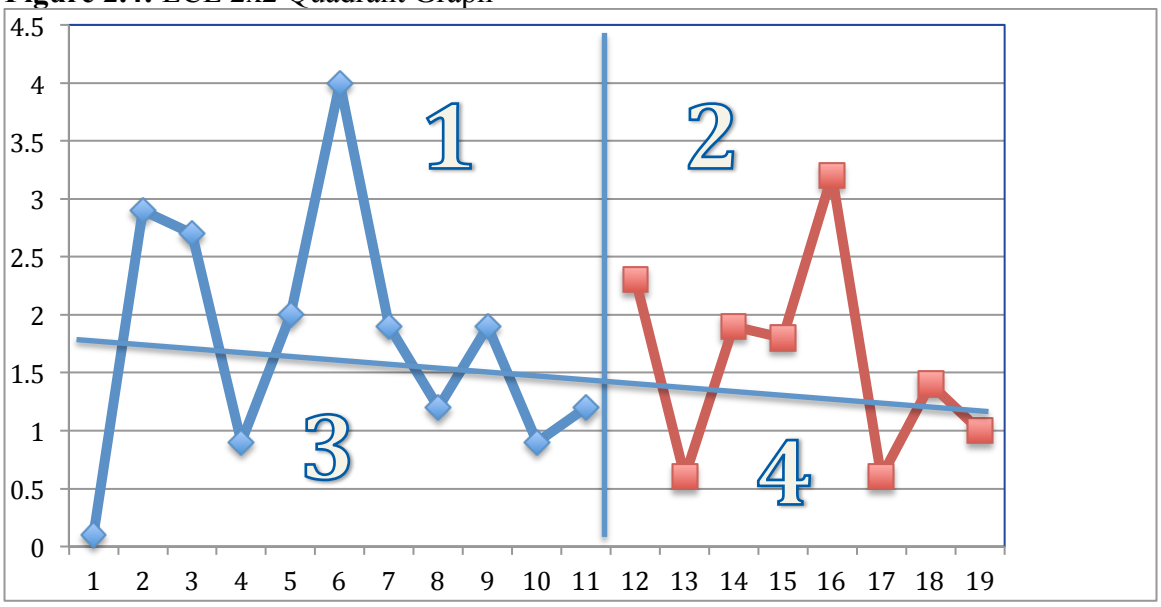
Using these data, effect size estimations and statistical significance values were calculated for ECL, NAP, and Tau-A, Tau-B, and Tau-U. The 2x2 matrices CFA analyzes were made using two methods, ECL and a 30% WM procedure. ECL was conducted using "pen and paper" techniques described by White and Haring (1980) applied to graphical representations of the data on a computer. By placing the data in Excel, and extending a linear line of best fit (calculated by splitting A phase data into two phases, calculating the median value of each phase, and drawing a straight line through the two median points extended into Phase B), this researcher calculated the percentage of data points above or below the line (depending on anticipated direction of treatment effect). Using a test for one proportion with SPSS, an effect size and *p value* were obtained. Effect sizes and *p* values for NAP were obtained using a web based calculator for SCED analysis provided by Vannest, Parker. and Gonen (2011). Similarly, effect sizes and p-values for Tau-A, Tau-B, and Tau-U were calculated using the statistical package R, with a script developed by Tarlow (2014).

For CFA statistical significance analysis, Stirling's formula was used. First, data sets were set in four quadrant configurations using ECL and a 30% WM configurations. Expected and observed frequencies were determined, as described in the next paragraph. Then, Stirling's formula (Formula 2.1) was used to calculate statistical significance. Stirling's formula was placed into an Excel spreadsheet, where variables including total

number of data points, observed data points in the quadrant, and expected frequency of data points could be entered.

The expected frequency was calculated by first extending the phase A line based on the type of analysis (i.e. ECL or the 30% Winzorized mean used for Tau and NAP) into phase B, and using the phase line to set 4 quadrants, illustrated below in Figure 2.4.

Figure 2.4: ECL 2x2 Quadrant Graph



Expected frequencies for quadrants 2 and 4 were calculated from the observed frequencies in quadrants 1 and 3. For example, in Figure 2.4 there are 6 observed data points in quadrant 1, out of a total 11 data points in phase A. 55% of the phase A data is in quadrant 1. As such, 55% of the data in phase B could be expected in quadrant 2. As there are 8 data points in phase B, 4.4 data points (55%) are expected to be in quadrant 2. As observed in Figure 2.4, 5 data points are observed in quadrant 2.

Using this methodology, and Stirling's formula, statistical significance was calculated for quadrant 2 and quadrant 4 for each extracted single case graph. Two *p* values were obtained for each graph, one above the split line (quadrant 2) and one below (quadrant 4). Since two values were created for each graph, the most conservative, or largest, significance value was used to compare CFA to the values obtained from non-overlap techniques. This allowed for simple and convenient comparisons to methods that provide on value rather than two. ECL statistical significance and CFA based on an ECL configuration were compared using a *t* test. CFA statistical significance values obtained by a 30% WM configuration were compared to NAP and Tau-A, B, and U statistical significance values using a one-way analysis of variance (ANOVA). By using an ANOVA, CFA statistical significance values obtained from the 30% WM could be compared to each group of *p* values obtained from NAP, Tau-A, B, and U without the added threat of type-1 error seen when using multiple *t* tests. In addition, as two statistical significance values were calculated for each graph, the percentage of agreement between the two quadrants was calculated. Agreement was defined as when both upper and lower B phase quadrants were either statistically significant ($p < .05$) or bother not statistically significant ($p > .05$). Disagreement was defined as one B phase quadrant being statistically significant, while the other was not. Appropriate descriptive statistics were also compiled and presented in table 2.1.

Results

ECL, NAP, TAU-A, TAU-B, and Tau-U effect sizes and *p* values were computed for each data set using an AB contrast. ECL effect size values were calculated

manually using a median trend line plotted from Phase A data and extended through Phase B. Statistical significance may be estimated by many statistical packages using the feature that calculates the difference between two proportions, under "proportion statistics" or "risk analysis" (Parker et al., 2009).

Table 2.1: Description of Graphs Pulled for Analysis

Total Graphs	168
Graphs from Journal of Behavior Modification	61
Graphs from Journal of Behavior Therapy	107
Ave number of data points in phase A	7.2
Ave number of data points in phase B	17.8
Average total number of data points	25.1
Median number of data points in phase A	4
Median number of data points in phase B	9.5
Median total number of data points	15
Min A Phase A	3
Min B Phase N	3
Min Total N	7
Max A Phase N	47
Max B Phase N	141
Max Total N	151

Descriptive statistics presented in table 2.2 indicated a large range of effect size values were obtained from the sample of SCED graphs. The majority of effect sizes produced from ECL and NAP were between .7 and 1, indicating a potential skew towards large effect sizes, however, Tau-A, Tau-B, and Tau-U effect sizes were more

evenly distributed across the range of results. NAP has a minimum value of .5, and as such the distribution above .5 was expected. 85.1% of the NAP effect size values were above .7.

Table 2.2: Effect Size Descriptive Statistics

	ECL	NAP	Tau-A	Tau-B	Tau-U
Max ES	1	1	.96	.98	.97
Min ES	0	.50	<.01	<.01	<.01
ES Range	1	.50	.96	.98	.96
Mean ES	.75	.86	.54	.56	.61
SD ES	.37	.14	.24	.24	.23
N ES (0-.5)	38 (22.6%)	0 (0%)	67 (39.9%)	61 (36.3%)	46 (27.4%)
N ES (.5-.7)	7 (4.2%)	25 (14.9%)	50 (29.8%)	48 (28.6%)	54 (32.1%)
N ES (.7-1)	123 (73.2%)	143 (85.1%)	51 (30.3%)	59 (35.1%)	68 (40.5%)

CFA *p* values were calculated by applying four quadrants to each graph for each method (ECL configuration for comparison with ECL, and a 30% Winzorized Mean (WM) for comparison with NAP, Tau-A, Tau-B & Tau-U) and calculating a *p* value for both upper and lower phase B quadrants. The larger *p* value between the two quadrants was used for comparison purposes to methods that produce one *p* value. Methodologically, this allowed for simplistic comparison to overlap methods that produce one *p* value rather than two. Conceptually, researchers would likely not ignore one of the two quadrants when making treatment effect interpretations. This multiple quadrant system may be of benefit to researchers choosing to use CFA. By providing two statistical significance values, one for each quadrant, more information is provided. In comparison, all other statistical significance measures used in this study provide one *p*

value. Additionally, when one quadrant results in statistical significance and the other does not, this broadens the information the researcher has to evaluate the treatment effect. Instead of a decision made by one *p* value, multiple values may lead to a more comprehensive view of the data. Researchers could have the most confidence that a treatment had a significant effect if both *p* values were statistically significant, less if only one were statistically significant, and if neither resulted in statistical significance then the researcher would have the least evidence that the treatment had an effect on the dependent variable.

The range of *p* values obtained was much greater for the non-overlap methods than for CFA by these non-overlap methods, as CFA had a max obtained *p* value of .3614. CFA by ECL configuration was the only method to average statistically significant results ($p < .05$). Both CFA by ECL and 30% WM configurations obtained the smallest standard deviation for *p* values. This may be an artifact of a limited range based on Stirling's formula. These results are presented below in table 2.3.

Table 2.3: *P* Value Descriptive Statistics

	ECL	**CFA by ECL**	**NAP**	**Tau-A**	**Tau-B**	**Tau-U**	**CFA 30% WM**
Max *p*	1	.3614	.9803	1.00	1.00	.96	.353
Min *p*	.0001	<.001	<.001	0	0	0	<.001
***p* Range**	.9999	.3614	.980	1.00	1.00	.96	.353
Mean *p*	.0569	.0447	.11	.09	.08	.07	.0597
SD *p*	.168	.0705	.20	.22	.21	.190	.0783

When applying CFA using non-overlap configurations there was a possibility that a B phase quadrant would have 0 data points, resulting in a divide by zero calculation for that quadrant when using Stirling's approximation of the binomial formula. These occurrences were tracked, and presented in table 2.4. This did not limit the analysis of the overall data set, as one quadrant would have all remaining data, and could still be analyzed for statistical significance using CFA. Over half of the CFA by ECL and 30 %WM configurations resulted in one B phase quadrant or the either having all of the data points for this sample of graphs. Said another way, over half of the ECL and 30% WM configurations resulted in all of the data being in either quadrant 2 or quadrant 4 in a graph presented such as Figure 2.5. This may be a product of the sample of data sets obtained, or the process by which the 2x2 quadrants are created. The large number of data sets utilized may indicate that this is not a unique pattern to applying CFA, and can be an expectation rather than an exception. Either way, this is not a concern overall. By simply analyzing the B phase quadrant that has data, and ignoring the quadrant that does not, we calculated a statistical significance value for the data set, which is used for evaluating the treatment effect.

Table 2.4: Divide By Zero Percentage of Occurrence

	ECL	**30% WM**
% DIV/0 Configuration	66.7%	53.0%

When two significance values were calculated for each data set, agreement between upper and lower quadrant statistical significance was analyzed. Agreement was

defined as when both upper and lower B phase quadrants either were statistically significant (p <.05), or both were not statistically significant (p >.05). Both ECL and 30% WM CFA configurations resulted in complete agreement between upper and lower quadrants, as presented in table 2.5. Said another way, either both quadrants constituted a type or anti-type, or both quadrants were not statistically significant. As will be seen in later chapters, this is not always the case with CFA. There are instances when CFA applied to SCEDs results in a statistically significant result in one B phase quadrant, while the other quadrant is not statistically significant. However, this event was not observed when using CFA based on an ECL or 30% WM configuration.

Table 2.5: CFA Statistical Significance Agreement Between Quadrants

	ECL	**30% WM**
% Agreement	100	100

To examine differences between non-overlap method calculations of p values and CFA calculated p values, an ANOVA was used for comparing CFA by 30% WM and NAP, Tau-A, Tau-B, and Tau-U. This is presented below in table 2.6 and table 2.7. This was utilized to protect against type-1 error, as seen in using multiple t-tests. A p value of .175 indicated that the 6 group p value means were not statistically different. Overall, this indicated that CFA by 30% WM performed similarly to established measures that provide p values.

Table 2.6: ANOVA Comparing Statistical Significance of CFA 30% WM and NAP, Tau-A, Tau-B, and Tau-U.

	CFA 30% WM	**NAP**	**Tau A**	**Tau B**	**Tau U**	**Total**
N	168	168	168	168	168	840
Mean	0.060	0.11	0.090	0.084	0.075	0.083
s^2	0.0061	0.041	0.050	0.045	0.036	0.036
SD	0.078	0.20	0.22	0.21	0.19	0.19
SE	0.0060	0.016	0.017	0.016	0.015	0.0065

Table 2.7: ANOVA Summary Table

Source	**SS**	**df**	**MS**	***F***	***p***
Treatment (Between Groups)	0.23	4	0.056	1.59	0.18
Error	29.68	835	0.036		
Total	29.90	839			

CFA by ECL configuration used the same quadrant configuration created by a median trend line in ECL. As such, one unpaired t-test, presented in Table 2.8, was used to compare means of CFA *p* values and ECL *p* values obtained using a test of one proportion. Results were similar to that of the ANOVA presented in table 2.6. The t-test between ECL and CFA by ECL indicated no statistically significant difference between the two *p* value group means.

Table 2.8: ECL & CFA by ECL Configuration Unpaired *t* Test

		Confidence Interval
p	.385	.0154 - .0398
t	.871	
df	334	
SE	.014	

Discussion

The impetus for this study was the lack of agreed upon best methods for SCED data analysis. Research examining the more common statistical methods has been undertaken by many researchers (Parker, Vannest, Davis, 2010; Parker, et al., 2005; Smith, Vannest, & Davis, 2011). These efforts have resulted in clarification, elaboration, and direction in the use of multiple types of analysis. However, the question remains as to which methods are best for use in what scenarios, and as many methods remain understudied, these questions will likely go unanswered for quite some time. In an effort to continue this exploration, this study compared CFA, which is a method frequently utilized in large group randomized control trials (von Eye, 2007) and applied it to SCEDs to answer a few basic questions. First, can CFA be used to evaluate SCEDs? Without a history of application, it was unknown as to whether data from SCEDs would be appropriate for CFA, what complications may occur, and if CFA would obtain results that make sense in the context of SCED data. Second, how does CFA compare to overlap methods that provide statistical significance? Would the data be similar to established methods? Or would there exist discernable differences between CFA and overlap methods? Would CFA provide something new?

With few complications, CFA was effectively applied to SCED data through the utilization of Stirling's approximation of the binomial formula. The basic application of graphical analysis took no longer than a hand calculation of comparable methods such as ECL once a formula was created in Microsoft Excel. The formula was straightforward in regards to placing it into an Excel spreadsheet, and only 3 data values (n, the total

number of data points across A and B phases; o, the observed data points in the quadrant being analyzed for significance; and e, the expected frequency of data points in the quadrant being analyzed) were necessary to obtain statistical significance values for a B phase quadrant. The naturally occurring 4-quadrant system when using non-overlap methods of analysis was convenient for CFA analysis. When these quadrants were not simple to establish, as was the case when using NAP and Tau-A, B, and U, a 30% WM application created an appropriate 4 quadrant configuration.

Frequency expectations were also simple to calculate by utilizing percentages of data frequency in phase A and transferring those percentages into phase B. Overall, the basic application to SCED data was a success, and was no more difficult than existing hand-calculation methods for SCED data analysis. This is an important if CFA applied to SCEDs are to remain intuitive and accessible to researchers of limited statistical experience.

A unique complication arose when using Stirling's formula. If one of the B phase quadrants had no data points when split using ECL or the 30% WM, then the calculation would result in a divide by zero situation. This occurred because the denominator of the formula would equal 0, and the calculation could not be completed. Specific to SCED non-overlap designs, this occurred when there was no overlap in the data between the A and B phases, based on the non-overlap method used for analysis. Within this study, 66.7% of ECL configurations, and 53% of WM data configurations resulted in this complication.

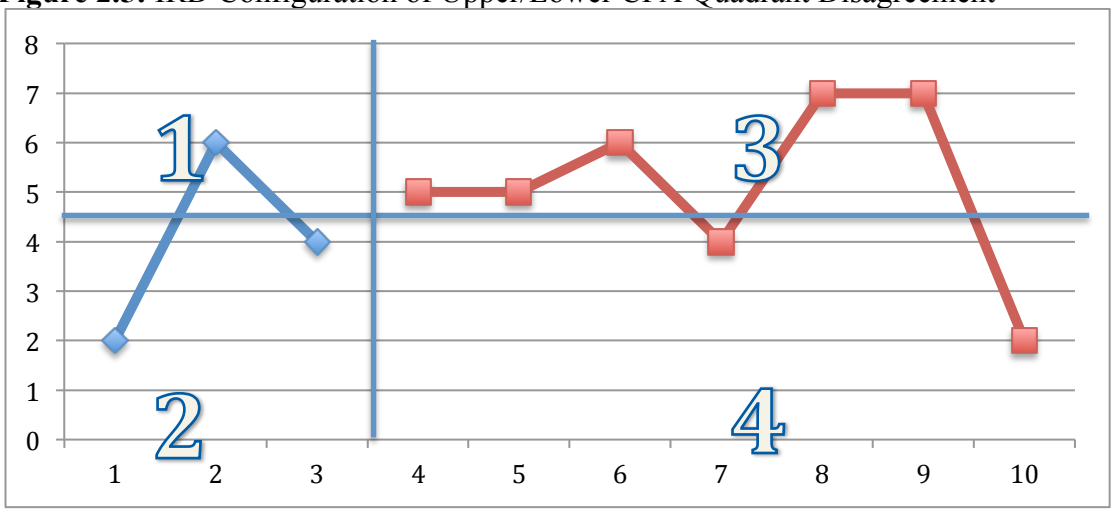
Although this would limit the graph to having one B phase quadrant for analysis, it did not prevent the graph from being analyzed for statistical significance, as the quadrant remaining could still be analyzed. Put another way, by ignoring the B phase quadrant with no data, the graph could still be analyzed for one statistical significance value using the other quadrant. This divide by zero complication may be specific to Stirling's formula. Other formulas such as the chi-square approximation to the z test are able to account for quadrants with no data. However, as Stirling's formula allowed for simple, quick calculations and could be used when quadrants had fewer than 5 data points, this formula was selected for the introductory nature of this study.

The differences between statistical significance values for upper and lower B phase quadrants within data sets may add a new level of data evaluation. As seen with comparison methods in this study, it is routine to obtain one statistical significance value for each data set. CFA applied to SCEDs calculate two separate values, one for upper and one for lower B phase quadrants. This leads to the possibility for statistical significance "disagreement" within one data set. Disagreement meaning that one quadrant indicates a statistically significant result (e.g. $p < .05$) while the other indicates no statistically significant result (e.g. $p > .05$). Said in another way, one B phase cell indicates that the treatment had a significant effect on the dependent variable, while the other cell indicates that the treatment had no significant effect. Although this study terms this disagreement, this may also be conceptualized as evidence on a sliding scale. Researchers could be more confident in concluding that treatment effects were significant when both quadrants indicate statistical significance, less when only one

quadrant is statistically significant, and when both are not statistically significant this could be interpreted as evidence the treatment had no statistically significant effect. As always, researchers should make these decisions with additional evidence to support conclusions, such as effect size values and visual analysis.

There were no instances of disagreement between quadrants in this study using CFA by ECL or 30% WM configurations. However, to further elaborate on this issue, we will work through an instance of disagreement with a different CFA configuration method here. An example of an Improvement Rate Difference (IRD; Sackett et al., 1997) configuration with CFA disagreement between upper and lower quadrants is presented below in Figure 2.5.

Figure 2.5: IRD Configuration of Upper/Lower CFA Quadrant Disagreement



For the purposes of this graph, the horizontal line represents the split line necessary to remove data points from each phase equally to remove all overlap. In this

instance, 1 data point from phase A (quadrant 1), and 2 data points from phase B (quadrant 4) are removed to eliminate all overlap. This results in four quadrants and data present in each.

Visually, there are certain data characteristics that stick out in Figure 2.5 that may lead to a better understanding of disagreement between quadrant 3 and 4. First, there are few data points in phase A. This results in 66% of B phase data expected in quadrant 4, as 66% of A phase data is in quadrant 2. As there are 7 data points in phase B, 4.67 data points (66% of 7) are expected in quadrant 4. Two data points are observed in quadrant 4. As such, CFA for quadrant 4 results in a *p* value of .067. This is not considered statistically significant.

Conversely, 33% of phase B data is expected in quadrant 3, as 33% of the phase A data is in quadrant 1. 33% of the 7 data points results in an expected quadrant 4 data frequency of 2.33. 5 data points are observed in quadrant 4. As such, CFA for quadrant 3 results in a *p* value of .0473. This is considered statistically significant. Remember that without knowing the context of the study (i.e. whether or not the intervention was intended to reduce or increase the independent variable) we do not know if this statistical significance constitutes a type or an anti-type.

This example presented in Figure 2.5 was not the only data presentation across this series of studies that resulted in a disagreement in statistical significance classification between upper and lower quadrants. When exploring alternative methods such as Percentage of Non-Overlapping Data (PND; Scruggs, Mastropieri, & Casto, 1987) in the next article in this series, we will elaborate further as to potential causes of

these differences. For now, the bottom line is that when disagreement between upper and lower quadrants is present, it may allow for the researcher or clinician to evaluate their data beyond a simple qualitative "statistically significant" or "not statistically significant" label. This disagreement may encourage the researcher further by suggesting the need for greater treatment evaluation. This appears to create a middle ground where although the treatment may be effective at some level, it appears to warrant additional evaluation and extra thought regarding if the treatment should be considered effective. Although less straightforward than a simple "significant" or "not significant" determination, this type of evidence may be helpful if researchers are open to integrating these results into treatment effect and future research decisions.

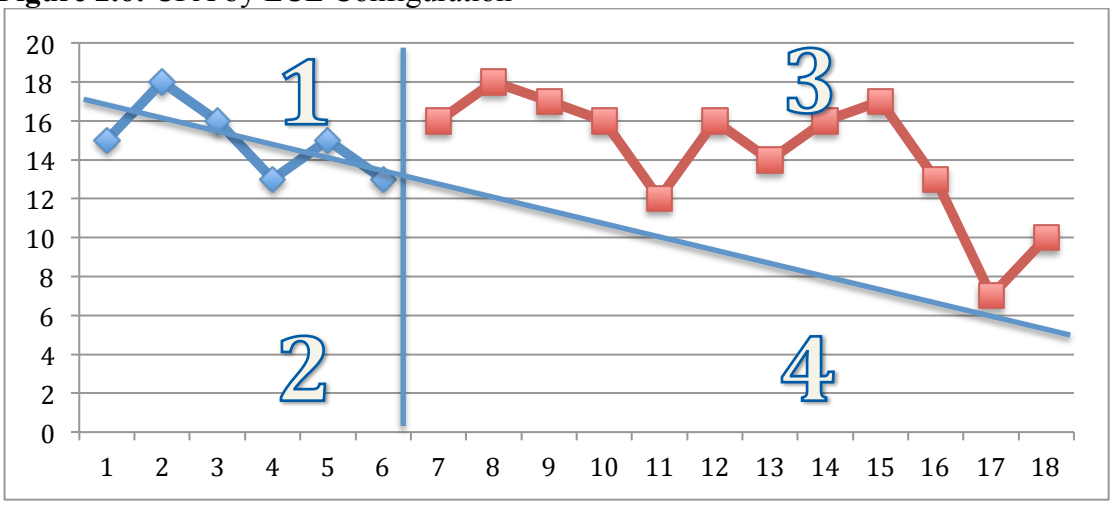
At face value, Stirling's formula appeared to respond appropriately to changes in SCED data presentations. When the observed frequency and expected frequency difference within a quadrant increased or decreased, the *p* value responded as expected. Differences needed to be larger for statistical significance when there was less overall data in the series, compared to when the total number of data points across A and B phases were larger. As a test, 6 total data points were entered into Stirling's formula, and statistical significance could still be reached with this limited, indicating potential sensitivity to data sets with few data points.

Descriptive statistics pointed to a smaller range of potential statistical significance values, as CFA had a *p* value range of .361, while ECL had a range of .999, NAP a range of .980, Tau-A and B a range of approximately 1, and Tau-U a range of .96. This limited range may have resulted in a lower average *p* value when compared to

alternative measures, but did not result in more statistically significant treatment effects. For this data set, CFA by ECL and CFA by 30% WM configurations found fewer data sets to be statistically significant than ECL, Tau-A, Tau-B, and Tau-U. Only NAP resulted in less statistically significant findings when compared to CFA. Based on CFA appearing more stringent, researchers who choose to utilize Stirling's formula could be as or more confident that they were avoiding a type-1 error, and that CFA was not resulting in artificial statistical significance. This is more of a speculation and would need more specific research in the future to determine if this was the case.

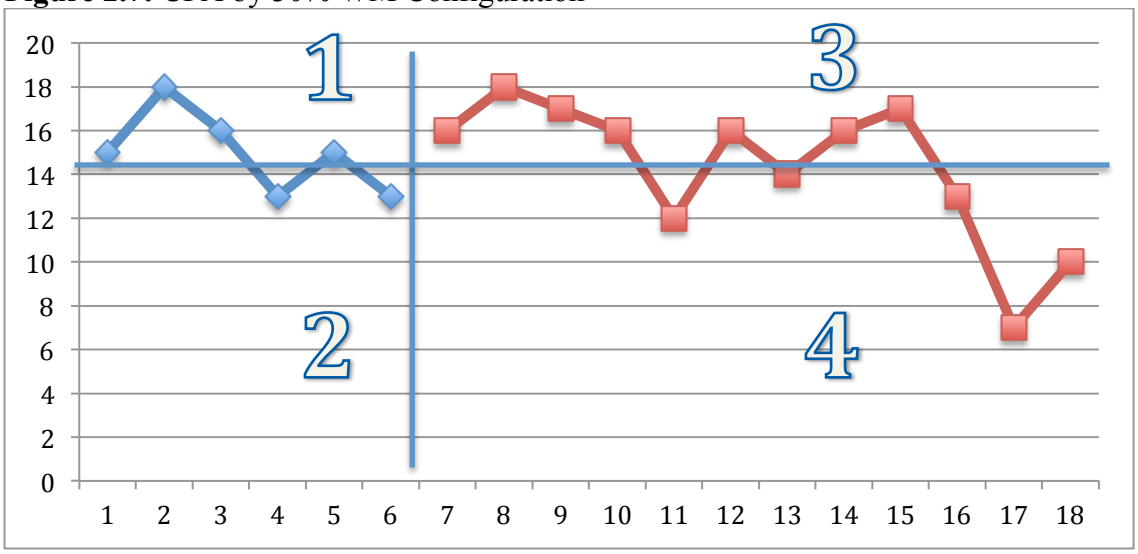
As expected, statistical significance values for one graph were different when CFA was run on the same graph using 2 different configurations. For example, Figure 2.7 below represents one data set pulled from the Journal of Behavioral Modification. When CFA was run using an ECL configuration, the result was a statistically significant finding in quadrant 3 ($p = .003$). Since there was 50% of the phase A data in quadrant 1, 50% of the B phase data is expected in quadrant 3. For Stirling's formula in quadrant 3, expected equals 6 (50% of 12), and observed equals 12. No statistical significance calculation was done for quadrant 4, as quadrant 4 has no data, which does not compute in Stirling's formula. Overall, this results in a statistically significant finding. This is matched by the test of two proportions run for ECL, which resulted in a statistically significant finding as well ($p < .001$).

Figure 2.6: CFA by ECL Configuration



Now, let's compare Figure 2.6 to Figure 2.7, a CFA by 30% WM configuration. The horizontal 30% WM line is set to 14.6 with 4 data points in quadrant 1, 2 in quadrant 2, 7 in quadrant 3, and 5 in quadrant 4. As there is data available in both quadrant 3 and 4, Stirling's formula was able to calculate statistical significance values for each. Phase A results in 66% of the data in quadrant 1, and 33% in quadrant 2. As such, 66% of Phase B data is expected in quadrant 3 (8 data points) compared to the 7 observed, while 33% is expected in quadrant 4 (4 data points) compared to the 5 observed. Stirling's formula results in statistical significance values of .1722 and .1803 for quadrants 3 and 4, respectively.

Figure 2.7: CFA by 30% WM Configuration



The visual difference in graphical representations should be clear. Using the same data, we obtain quite different data conceptualizations when choosing to use either a measure of phase A data trend (i.e. ECL), or use a measure of phase A central tendency (i.e. 30% WM) to characterize the data. For statistical significance, we also obtain quite different results, as CFA by ECL is statistically significant, while CFA by 30% WM is not.

This example shows a few important considerations. First, when using CFA to analyze SCED data, data conceptualization is key to the process. This is a positive piece of evidence for the introduction of CFA to SCED data. Instead of removing the researcher's overlap conceptualization of the graphical data, CFA requires that the conceptualization used to select the non-overlap effect size estimation be integrated into the calculations. Said another way, CFA needs the quadrant system formed by these non-overlap measures, and values produced by CFA will change across different

methods for generating the 2x2 matrix. The decisions one makes for constructing the quadrants or 2x2 matrix has a considerable impact on the results CFA produces and should be made with care. If a method such as Tau-A, B, or U is chosen, a general measure of central tendency may be an appropriate addition. Second, CFA is able to discriminate between data conceptualizations for the same set of data, providing statistically significant results for one, while providing not statistically significant results for the other. The ability to discriminate between data conceptualizations adds to CFA's overall utility, as it appropriately adjusts for the method that creates the quadrants.

When compared to non-overlap methods that provide statistical significance values using an ANOVA and a two-tailed unpaired t Test, CFA using Stirling's formula was found to perform similar to established methods. This should be interpreted that CFA maintains the statistical rigor that is expected by researchers using SCEDs, and as such gains value through its flexibility of application across methods and its potential application to analytic procedures that do not provide for statistical significance. The flexibility in changing the formation of the four quadrants to match the overlap methodology without losing statistical rigor is likely where CFA can lay claim as a valuable addition to SCED data evaluation. This goes beyond adding a statistical significance value from one method to an effect size value from another. Rather, it connects the statistical significance value to the effect size estimation method through the initial conceptualization used to select the non-overlap method. This, added to the potential benefit of analyzing both upper and lower quadrants rather than the B phase as a whole, may help distinguish CFA as a valuable statistical tool. Finally, by having the

ability to be applied across many non-overlap methods, CFA potentially creates a common statistical significance language to use across non-overlap analysis techniques. Rather than using a different statistical significance technique paired with different non-overlap methods, this study provides support that CFA could be applied to multiple non-overlap methods used to generate the 2x2 matrices, rather than having to switch to a different statistical tool for each new method.

Limitations

This paper focused on introducing a statistical method that is typically used in RCTs and applying it to SCED data for treatment effect analysis. Accordingly, this study had multiple limitations, as the scope of this research was on the application of a technique to a new research discipline, and few, if any, questions about CFA applied to SCED data had been answered prior to this study. First, this research focused on the application to published data. Although this provides more assurance that the method works with real world data, the limits of CFA applied to SCED data could be more fully explored in a simulation or monte carlo study. Nevertheless, this was a good start at evaluating CFA, but had the sample been larger, with a wider variety of single-case data sets, we may have more confidence in how CFA performs with SCED data.

Stirling's formula applied to SCED data also remains largely unexamined. This includes the potential mathematical restricted range of Stirling's formula applied to SCEDs. It may be important to know when, if ever, Stirling's formula reacts in ways that contraindicate its application to SCED data. This may be in the form of the size of data sets, or non-overlap configurations that CFA and/or Stirling's formula is ill

equipped to handle such as the divide by zero scenarios. As such, Stirling's formula may not be the most appropriate formula for SCED data configurations. However, this remains unknown until further research tackles these specific questions, which could be answered using simulated data to test the limitations of CFA.

Due to the scope of this study, the specific limitation regarding divide by zero scenarios was not examined. Rather, the occurrence was simply tracked and reported. Non-overlap methods applied to SCED data appear to regularly result in a B phase quadrant with no data, and future research may benefit from determining a solution to this problem. This may be done by comparing how Stirling's formula performs versus the chi-square approximation to the z test, a formula that reportedly has the potential for overcoming these configurations (von Eye, 2010). Potential solutions also include utilizing partial recordings of data points such as dividing a data point into two for the purpose of giving the denominator in Stirling's formula a value, or by simply ignoring the B phase quadrant with no data and consider the CFA *p* value for one B phase quadrant representative of the whole data set.

Finally, this research is limited to comparing CFA to non-overlap effect size estimation methods, and does not consider the many other available statistical methods for calculating effect sizes and statistical significance values. Additionally, this method only applies CFA across phase A and B, and does not address other design structures. Each of these factors may deserve additional research for further exploration of CFA applied to SCEDs.

Conclusions

The purpose of this research was to apply CFA as a statistical significance analysis tool for use in SCEDs to determine if CFA had potential as a method for analyzing single-case data. The utility of CFA in analyzing single-case data was briefly explored, and potential complications when using CFA with SCED data was discussed. Overall, this research showed that CFA could be applied to SCED data using Stirling's formula. This preliminary application of CFA to SCED data suggests it may be useful in the analysis of single-case data..

It appears that the multiple quadrant evaluation of data, paired with results comparable to statistical significance results from established methods, showed that CFA could be a valuable clinical tool in the evaluation of treatment effects. Additionally, as this study used published data, it appears that CFA can be used with data characteristic of clinical research such as short data series. As this study used a relatively large set of published data, researchers and clinicians who choose to use CFA moving forward should be confident that it could perform across many data presentations.

Per flexibility, CFA was able to be effectively applied to multiple data configurations through the utilization of the 2x2 quadrant system. Depending on how the researcher chooses to analyze their data, this study showed that CFA may be applied to multiple different non-overlap or 2x2 configurations. When not directly applicable, a 30% WM showed comparable results to Tau-A, Tau-B, and Tau-U statistical significance results. This wide array of applicability may provide some clarity moving forward in the choice of which techniques to use in what scenarios, as CFA appears

applicable and adaptable across a wide array of data presentations and effect size calculation methods.

The take home message from this introductory research is that CFA can be applied to SCED data presentations, it can be applied across multiple different non-overlap effect size 2x2 data configurations, and may provide valuable information in the process of determining treatment effects. As such, it is recommended that when researchers choose to use ECL for effect size estimations, researchers also take the time to apply CFA using the ECL quadrant configuration to provide for additional information regarding the treatment effect on the independent variable. When researchers choose to use Tau-A, B, U or NAP, CFA using a 30% WM was statistically equivalent with values provided by Tau and NAP. However, it is clear that a general measure of central tendency is not an equivalent comparison to the manner in which NAP and Tau are calculated. As such, it is recommended that CFA not replace existing ECL, NAP, or Tau statistical significance calculations at this point. Rather, CFA with a 30% WM quadrant configuration could be an addition to these analytic methods because it provides additional information regarding treatment effect evaluation. Additional research will be necessary to fully understand the role CFA should play in the analysis of single-case data.

CFA was not without limitations. Additional research will be necessary to establish the technique as an appropriate statistical significance measure for SCED data. The current study presents an initial exploration of CFA with single-case data. Overall, CFA appeared to perform as expected in the evaluation of SCED data. It is important to

remember that one statistical analysis should not be the sole determining factor of an overall treatment effect. Factors such as the design, context, visual analysis of graphed data, and multiple effect size calculations based on different techniques should be considered in order to accurately evaluate a treatment effect.

CHAPTER III

MANUSCRIPT #2: THE APPLICATION OF CONFIGURAL FREQUENCY ANALYSIS USING STIRLING'S FORMULA TO SINGLE-CASE DESIGNS: A COMPARISON TO NON-OVERLAP INDICES THAT DO NOT PROVIDE STATISTICAL SIGNIFICANCE

Introduction

The first in this series of articles provided evidence that CFA may be applied to data from SCEDs. It showed that CFA produced statistical significance values that were statistically no different than statistical significance values from established methods. The results also suggested that CFA could be applied to a 2x2 matrix produced with a non-overlap method. However, as is well documented, there are many different statistical treatment effect evaluation methods and which ones perform the best is still debated (Kratochwill et al., 2010). As such, questions still exist as to which effect size estimation methods may create an appropriate 2x2 configuration that CFA can be applied to, how well they will compare to alternative methods, and if CFA is a value added tool for SCED researchers making treatment effect determinations?

To expand on the initial application of CFA, this article takes CFA and applies it to four non-overlap effect size estimation methods that do not provide statistical significance values. The aim was to further explore applications of CFA to additional effect size estimation methods used for generating the required 2x2 matrix. Specifically, how would CFA, which only provides a statistical significance value, compare to

methods that only provide for an effect size value? As these non-overlap methods are frequently used in SCED research, this was a logical next step in CFA's introduction to SCED data.

Non-Overlap Methods Without Statistical Significance

This study applied CFA, which produces a *p* value, to non-overlap methods that do not provide *p* values, and compared those results. Methods for comparison were selected by looking at research comparing non-overlap methods to one another (Parker, Vannest, & Davis, 2011), and selecting methods that did not have built in statistical significance calculations. Each of these methods was a non-parametric analysis tool, as non-parametric tools have been determined appropriate for use with SCED data (Parker, Vannest, & Brown, 2009).

Applying CFA to non-parametric, non-overlap effect size 2x2 configurations versus non-overlap methods that do not provider statistical significance values may answer several questions. First, does CFA adapt well to data configurations made by these additional methods? The first study utilized the extended celebration line method (ECL), and a 30% Winzorized mean to create 2x2 configurations and applied CFA to these configurations accordingly. Then, *p* values calculated using CFA with these configurations were compared to *p* values calculated using established methods, as well as *p* values from Kendall's Tau-A, Tau-B, and Tau-U. It was determined that CFA could be used with ECL to generate a 2x2 matrix. Also, by using a 30% WM for Tau-A, B, U and NAP which do not provide 2x2 matrices, the results of CFA with a 30% WM could be compared to the results of Tau-A, B, U and NAP.

Second, how do CFA statistical significance values compare with effect size values obtained using non-overlap effect size estimation methods? Said another way, if CFA uses non-overlap methods to create a 2x2 configuration, how do the statistical significance values compare to the effect sizes obtained from the method used to create the configuration? For example, when a large effect size is present, with an abundant amount of data points, one may reasonably expect that CFA would provide a statistical significance value that suggests statistical significance. However, what if a large effect size was present, with minimal data points? This is frequently seen in methods such as percentage of non-overlapping data (PND). Using these measures, whether there are 3 data points or 30 data points, if there is no overlap between the SCED baseline and treatment phases, then the effect size will reach it's ceiling of 1. CFA could potentially fill this void by providing statistical significance based on data frequency expectations when the sensitivity of non-overlap techniques reaches a ceiling.

To answer these questions, CFA was calculated using 2x2 matrices from PND (Scruggs, Mastropieri, & Casto, 1987), the percentage of all non-overlapping data (PAND; Parker, Hagan-Burke, & Vannest, 2007), the percentage exceeding the median (PEM; Ma, 2006), the improvement rate difference (IRD; Cochrane Collaboration, 2006; Sackett et al., 1997), as well as a 30% Winsorized Mean (WM; Staudte, & Sheather, 1990; Wilcox, & Keselman, 2003). Each of these methods (except the 30% WM which was used solely to provide a 2x2 matrix) was also run on all data sets, and results from the CFA calculations and non-overlap methods were compared. Each method provides unique strengths and weaknesses when it comes to SCED data analysis. However,

extensive review of each method is beyond the scope of this study. Readers interested in learning more are encouraged to explore the wealth of articles that elaborate on these methods, including the article used for the selection of comparison methods (Parker, Vannest, & Davis, 2011). For the purposes of this study, we move forward in reminding readers how to use Stirling's formula with CFA. After this, we briefly describe the non-overlap effect size methods used in this study and illustrate the 2x2 matrices made by each method. These illustrations will help show how CFA is applied using each 2x2 matrix.

Stirling's Formula for the Approximation of the Binomial Test

As we move through the different types of effect size estimation methods used for this study, we will present examples of CFA using Stirling's formula to illustrate how statistical significance can be calculated. Stirling's formula (Formula 3.1) is presented below. For the purposes of this formula, three values need to be determined to calculate a statistical significance value for a given phase B quadrant. First we determine *n*, which equals the total number of data points within A and B phases. Second, we determine *o*, which is the number of observed data points in the quadrant of interest, such as the upper quadrant in phase B. Next, we determine the expected frequency of data points in the quadrant of interest, e, which is used to calculate *p* (the expected frequency of data points divided by n; (p=e/n)). Lastly, for Stirling's formula, we calculate *q*, which equals 1-*p*.

Formula 3.1: Approximation of the binomial test using Stirling's formula

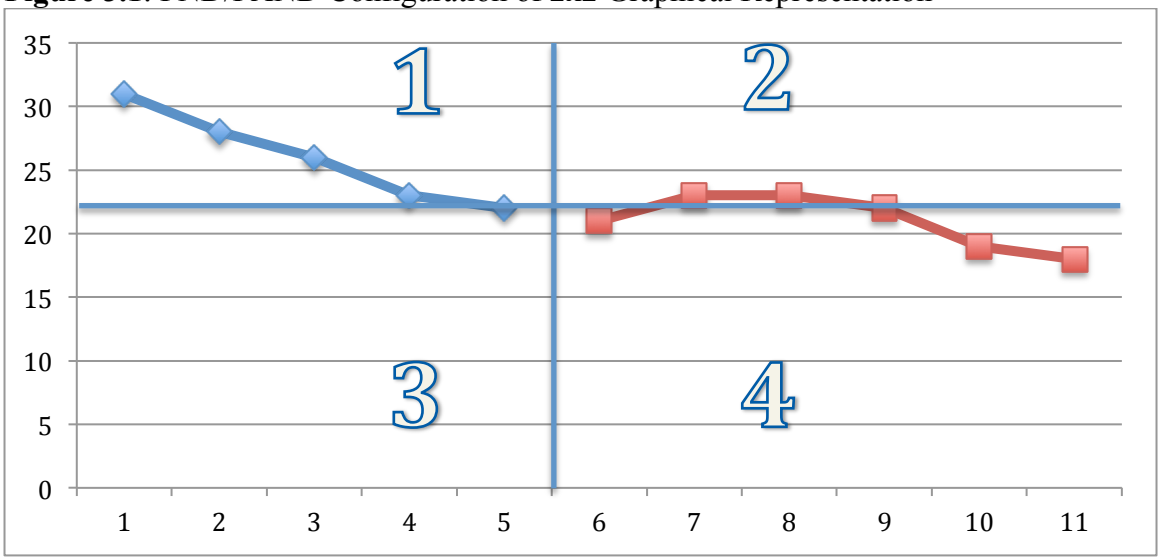
$$\hat{B}'^{(o)} = \left(\frac{n}{2\pi o(n-o)}\right)^{\frac{1}{2}} \left(\frac{np}{o}\right)^{o} \left(\frac{nq}{n-o}\right)^{n-o}$$

As p and q are calculated using e and n, we need only determine the values of o, e, and n to calculate statistical significance using Stirling's formula. As we move forward, we will present examples of each so that readers may have a greater understanding of how CFA using Stirling's formula can be applied to various non-overlap configurations. For the purposes of this study, we used a Microsoft Excel file to calculate Stirling's formula. This allowed for quick changes of o, e, and n, and was convenient as this study used many data sets to apply CFA across various methods.

Percentage of Non-overlapping Data (PND)

As the only non-overlap method that emphasizes one data point in phase A (Parker, Vannest, and Davis, 2011) PND has clear, documented, and emphasized limitations. Calculated as the percentage of data in phase B that exceeds the highest data point in phase A (Scruggs, Mastropieri, & Casto, 1987), PND is arguably the simplest effect size estimation to calculate, and correlates well with visual analyses (Parker et al., 2007). Although a complete desertion of PND has been proposed (Kratochwill et al., 2010; Parker & Vannest, 2009), it remains a frequently published method, and therefore will be used as a comparison method for the purposes of this study.

Figure 3.1: PND/PAND Configuration of 2x2 Graphical Representation



Here, we walk through an example of CFA by Stirling's formula with a PND/PAND configuration. First, for PND/PAND, we determine the "highest" data point in phase A, or as in the case with Figure 3.1, the lowest data point. This is based on the expected direction of treatment effect. Here we assumed an expected reduction in the dependent variable upon introduction of the treatment. Accordingly, in Figure 3.1 we used the lowest data point. To set this line in Figure 3.1 ahorizontal line was placed through the lowest data point (5,22) and extended through both phases. As seen in Figure 3.1, this creates the necessary 2x2 quadrant system to use CFA with Stirling's formula.

Next, we determine the total number of data points across all phases (n) and the observed (*o*) and expected (*e*) values for the quadrant of interest. Here, we calculate statistical significance for quadrant 4. First, frequency of observed data points is determined by counting all data points across both phases. For Figure 3.1, $o = 11$.

Second, we determine the observed data points for quadrant 4. Here, we present a unique consideration for CFA. As seen in Figure 3.1, the horizontal line runs through, or "splits" the fourth data point in phase B (9,22). For the purposes of this study, we split the value of the data point equally across quadrants. As such, half of the data point (.5) was counted for quadrant 2, and half for quadrant 4. This results in an observed frequency in quadrant 4 of 3.5 (o = 3.5).

Finally, we calculate the expected frequency (e) of data in quadrant 4. To do so, we first determine the percentage of phase A data present in quadrant 3. Again, a datum point (5,22) is split by the horizontal line. We count half of the data point for quadrant 1, and half for quadrant 3. This results in .5 data points present in quadrant 3, which represents 10% of the phase A data (.5 divided by 5 data points equals .1 or 10%). As such, we would expect 10% of phase B data to be in quadrant 4 if the treatment had no effect. 10% of 6 data points equals .6 data points expected in quadrant 4 (e = .6). Using Stirling's formula, CFA for quadrant 4 results in a statistical significance p value of .006. This would indicate a statistically significant treatment effect.

Percentage of All Non-overlapping Data (PAND)

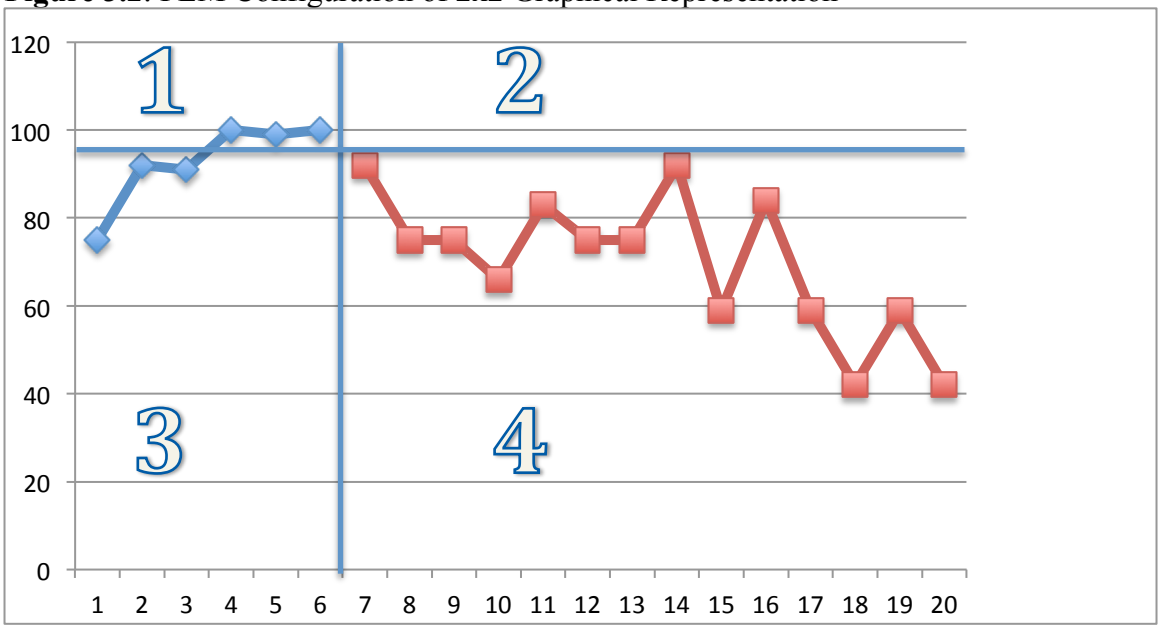
PAND takes the concept of non-overlap presented in PND, and removes the emphasis on one data point by considering all overlap present between phase A and phase B. PAND is calculated by setting a horizontal line through the highest/lowest data point, removing the minimum number of data points that eliminates all overlap, and dividing the remaining number of data points by the total number of data points across both phases (Parker, Hagen-Burke, & Vannest, 2007). Data may be removed from phase

A, phase B, or equally from both phase A and phase B. As seen in Figure 3.1, the 2x2 configuration created when using PAND is equivalent to that of PND. Although effect size estimation will change between the two measures, CFA statistical significance calculations will be the same when the non-overlap configuration is established in the same way. As CFA uses data frequency in cells to calculate statistical significance, and frequencies are equivalent across PND and PAND, CFA with Stirling's formula will produce the same statistical significance values.

Percentage Exceeding the Median (PEM)

Simple to calculate, PEM takes the number of B phase data points that exceed an A phase median trend line extended from phase A, and divides this by the total number of B phase data points (Ma, 2006). This technique can also be supplemented with alternative robust means of central tendency other than the median. Used frequently in applied research (Parker & Hagan-Burke, 2007) and utilized for meta-analysis of single case designs (Ma, 2009; Preston, & Carter, 2009), PEM has established its place in SCED research. Below is an example of the 2x2 configuration created when using PEM.

Figure 3.2: PEM Configuration of 2x2 Graphical Representation



When using PEM, we first determine the median value of the phase A data. For Figure 3.2, the median value of phase A data (75, 91, 92, 99, 100, 100) is 95.5. Next, a horizontal line is set at 95.5, and extended through all phases. Consequently, a four quadrant presentation is created, as seen in Figure 3.2. This data presentation creates a unique consideration for Stirling's formula. As there are no data points in quadrant 2, Stirling's formula cannot be used to calculate statistical significance for quadrant 2.

We suggest that when this happens, researchers who choose to use Stirling's formula ignore the quadrant with no observed data points, and simply calculate statistical significance for the alternative quadrant. This will result in one statistical significance value. Here, we will calculate statistical significance for quadrant 4. First, we determine the total number of data points across both phases (n) by counting the number of observed data points. For Figure 3.2, $n = 20$. Next, we determine the total number of

observed data points (*o*) in the quadrant of interest. In this case, *o* = 16, as all of phase B data is in quadrant 4.

Finally, we calculate the expected frequency (*e*) of data in quadrant 4. We do this by first determining the percentage of phase A data in quadrant 3 (50%) and use this to calculate the expected data frequency in quadrant 4 (50% of 16 data points equals 8 data points). As such, *e* = 8. When entered into Stirling's formula, the resulting statistical significance value $p < .001$. This would be considered statistically significant, and provide evidence that the treatment had a statistically significant effect on the dependent variable.

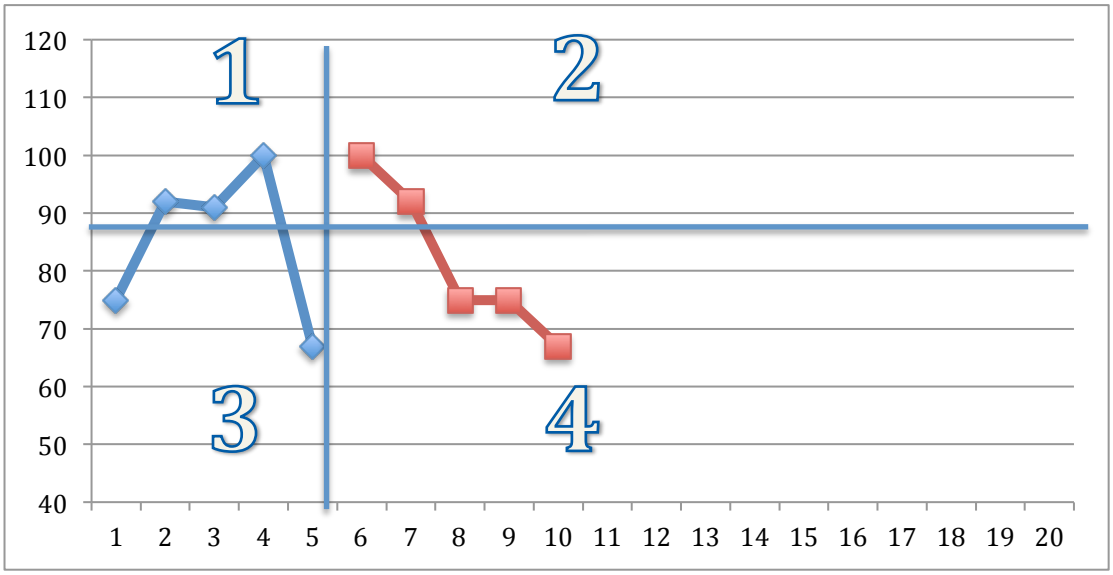
When using a 30% WM configuration to set the 2x2 quadrants for calculating CFA, researchers would use the same strategy as PEM calculations, but would set the horizontal line based on a 30% WM of phase A data. Procedures for calculating a 30% WM are presented in the first article in this series, as well as by Wilcox and Kesselman (2003). As a reminder, a 30% WM does not calculate an effect size. Rather, it is an alternative method for establishing the 2x2 quadrant configurations for CFA calculations.

Improvement Rate Difference (IRD)

Parker et al. (2009) present IRD as the difference in the A phase improvement rate and the B phase improvement rate. Similar to CFA, IRD is calculated using a 2x2 table. First, overlap is eliminated by removing the minimum number of data points from Phase A and/or Phase B. Then, two ratios are calculated, one from phase A and one from phase B. Each ratio is calculated as the remaining data within a phase divided by

the total data within that same phase. A two proportions test is completed using the two ratios. For the purposes of this study, an R script developed by Tarlow (2014) was implemented. This served two purposes. First, it allowed for simple, quick calculations of a robust IRD effect size by removing data points equally from phase A and phase B. Second, once calculated, the script placed an overlap line similar to that of PND and PAND that naturally created a 2x2 configuration, allowing for CFA calculations. An example of this line is represented in Figure 3.3.

Figure 3.3: IRD Configuration of 2x2 Graphical Representation



We will walk through an example, using Figure 3.3 to demonstrate a CFA calculation by IRD configuration. When using IRD, first we determine how many data points in phase A and phase B need to be removed. For Figure 3.3, two data points in each phase will be removed to create non-overlap between phases. Assuming that the

intended direction of this intervention was to reduce or lower the observed dependent variable, the 2 lowest data points in phase A and the two highest data points in phase B were removed. This creates a split line through both phases as shown in Figure 3.3, as this line represents where all data points below in phase A and above in phase B need to be removed to eliminate overlap. Readers should note that this line is an approximation line, and could be moved higher or lower as long as only two data points were present in quadrant 2, and two data points in quadrant 3. Moving the line within these constraints would not change the value of IRD, nor would it change the value of CFA.

As shown previously, CFA can be used to calculate two statistical significance values, one for quadrant 2 (upper phase B) and one for quadrant 4 (lower quadrant B). Here we walk through the calculation for CFA with Stirling's Formula for quadrant 2. To calculate CFA, we need to determine three values from Figure 3.3. The first, total number of data points in both phases (*n*), is determined by counting the number of data points across both phases. For Figure 3.3, N = 10. Second, we determine *o* (observed data points) for quadrant 2 by counting the number of data points in quadrant 2, which equals 2 (*o* = 2).

Finally, we determine how many data points were expected in quadrant 2. To do this, we look at the upper phase A quadrant. If we were calculating CFA for quadrant 4, we would look at the lower phase A quadrant. In quadrant 1, there are 3 data points, of the 5 data points total in phase A. This means that 60% of phase A data was in quadrant 1. If the treatment had no effect, we could hypothesize that 60% of the phase B data would also be in the upper quadrant, or quadrant 2. 60% of the phase B data points (5

data points) equal 3. The expected frequency for quadrant 2 is 3 ($e = 3$). Using Stirling's formula, we get $p = .244$. Thus, quadrant 2 data does not represent a statistically significant difference from quadrant 1, providing evidence that the treatment did not have a statistically significant effect on the dependent variable.

Illustrative Articles

This study utilized the same pool of articles and graphs used for the first article in the series. Published data was pulled from *The Journal of Behavior Therapy* and *The Journal of Behavior Modification*. The search was limited to articles published between 1990 and 2013. Selection criteria including a graphically represented A and B phase, and a minimum of three data points in each phase limited the number of graphs used. Some graphs had more than one series of data. When multiple data series were present on one graph, each was analyzed for use in the study. This resulted in a pool of 147 graphs, and 168 total data sets.

Data Extraction and Analysis

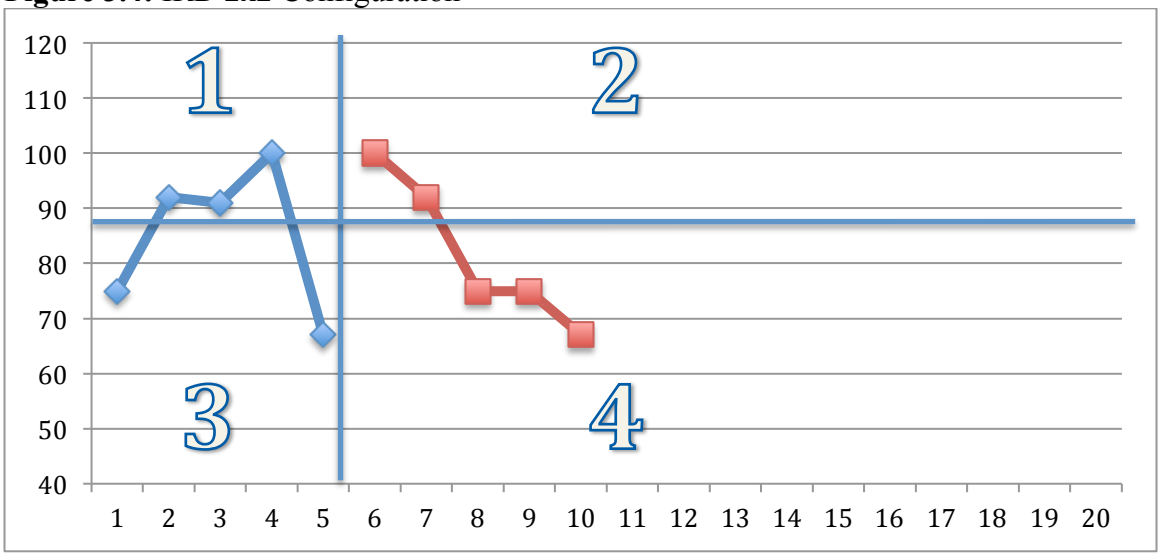
Graphs were extracted using the "Snipping Tool" from Microsoft Office version 2010, digitized using GraphClick for Apple OS X, and saved in Excel for analysis using the specified non-overlap methods and CFA. Using these data, effect size estimations were calculated for PND, PAND, PEM, and IRD. CFA was conducted using PND, PAND, PEM, IRD, and 30% WM to create the 2x2 matrices required for CFA. PAND and PEM data were analyzed using "pen and paper" techniques applied to graphical representations of the data on a computer. After placing the data in Microsoft Excel, lines were extended from phase A through phase B to represent the correct 2x2

configuration based on the method, and appropriate calculations could be performed via calculator. As noted before, a script by Tarlow, K. (2014) was used to calculate an effect size for IRD, where data points were pulled equally from phase A and phase B. This script was also used to create the IRD 2x2 configuration necessary for CFA calculations, as once the data was removed from each phase equally, a horizontal line was automatically created to represent non-overlap.

CFA with PND, PAND, PEM, IRD, and 30% WM configurations were calculated using Stirling's formula for the approximation of the binomial. A spreadsheet developed for the purposes of this study was utilized, and tested against published CFA data (von Eye, 1987) to ensure accuracy of calculations. By entering the total number of data points across A and B phases (i.e. "n"), the number of data points observed in the quadrant of interest (i.e. "o"), and the expected frequency of data points in the same quadrant (i.e. "e") the Microsoft Excel formula provided the statistical significance value for each B phase quadrant.

Expected frequency was calculated by first extending the phase A line based on the type of analysis (i.e. IRD, PEM, PAND or PND) into phase B, and using the phase line to set 4 quadrants. An additional example of this calculation method is presented below in Figure 3.5, using an IRD configuration.

Figure 3.4: IRD 2x2 Configuration



In Figure 3.4, expected frequencies (*e*) for quadrants 2 and 4 were calculated from the observed frequencies in quadrants 1 and 3. As there are 3 observed data points in quadrant 1, out of a total 5 data points in phase A, 60% of the data in the B phase could be expected to be in quadrant 2. As there are 5 data points in phase B, 3 data points (60%) are expected in quadrant 2, with an observed frequency of 2 data points.

CFA statistical significance was calculated for quadrant 2 and quadrant 4 of each single case graph. This resulted in two statistical significance values for each graph. The larger of the two values was used for comparison purposes, as this allowed for convenient comparisons with methods that provide one value rather than two. Descriptive statistics for effect size and *p* values were calculated and presented. Frequency of divide by zero presentations were compiled and presented as well, along with the percentage agreement of statistical significance between top and bottom B phase quadrants. Effect sizes and CFA with PND, PAND, PEM, IRD, and a 30% WM

configurations were compared using a simple Pearson *r* correlation, as well as a point biserial correlation between the statistical significance (i.e. above or below p = .05) and the effect size value.

Results

168 data sets that met study criteria were collected across a 23 year review of *The Journal of Behavior Therapy (BT)* and *The Journal of Behavior Modification (BM).* Although a minimum of 5 data points within A and B phases were ideal, data sets with a minimum of 3 data points were included in this study to explore the utility of CFA when applied to data sets with less then the ideal number of data points. Descriptive statistics, presented in table 3.1, indicated that most SCED graphs used had more than 4 data points in phase A, and more than 9 in phase B. A diverse range of total data points were present in this sample of graphs, suggested by the relatively large difference between minimum and maximum data points.

Table 3.1: Description of Sample of Graphs Pulled for Analysis

Total Graphs	168
Graphs from Journal of Behavior Modification	61
Graphs from Journal of Behavior Therapy	107
Average A Phase N	7.24
Average B Phase N	17.81
Average Total N	25.05
Median A Phase N	4
Median B Phase N	9.5
Median Total N	15
Min A Phase A	3
Min B Phase N	3
Min Total N	7
Max A Phase N	47
Max B Phase N	141
Max Total N	151

PND, PAND, PEM, and IRD effect sizes were calculated for each data set using an AB contrast. PND was calculated manually using Excel. A single highest data point in Phase A was identified, and a horizontal line was plotted from that point into phase B. Data points above this line were counted, and divided by total Phase B data points, resulting in PND. PAND was calculated using excel in a similar manner, identifying the minimum number of data points to remove from phase B, and calculating the ratio of number of data points not removed to the total to result in PAND. This allowed for the even removal of data points from Phase A and Phase B to eliminate all overlap between phases, and a ratio of the number of data points not removed to the total was calculated, equaling PAND.

PEM was calculated similarly to PND, by calculating the median of the A phase and extending a horizontal line through the A phase median extending into phase B. The

percentage of Phase B data above the extended line equals PEM. A robust IRD was calculated using an R script developed by Tarlow, K. (2014), allowing for the even removal of A and B phase data points to remove all overlap.

Results indicated a large range of effect size results obtained from the sample of SCED graphs, presented in table 3.2. The majority of effect size calculations for PND, PAND, PEM, and IRD were between .7 and 1, indicating a potential skew towards larger treatment effects within this sample of graphs.

Table 3.2: Effect Size Descriptive Statistics

	PND	**PAND**	**PEM**	**IRD**
Max ES	1	1	1	1
Min ES	0	.21	0	0
ES Range	1	.79	1	1
Mean ES	.62	.85	.87	.62
SD ES	.36	.15	.20	.32
N ES (0-.5)	61 (36.3%)	8 (5.8%)	12 (7.14%)	62 (36.9%)
N ES (.5-.7)	19 (11.3)	17 (10.1%)	13 (7.74%)	34 (20.2%)
N ES (.7-1)	88 (52.4%)	143 (85.1%)	143 (85.1%)	72 (42.9%)

CFA *p* values were calculated by creating four quadrants across the A and B phase for each method. Additionally, a 30% WM was utilized as a general comparison method. A *p* value was calculated for both upper and lower phase B quadrants. The larger *p* value between the two quadrants was used for comparison purposes. This provided a simple way to compare overlap methods, which produce one *p* value or effect size value to CFA which produces two *p* values per data set. Results of these calculations and descriptive values are presented below in table 3.3. CFA produced a max *p* value and *p* value range of .581 (CFA by PND/PAND). As seen in table 3.3

under "Mean *p,*" CFA by IRD configuration was the only method to average statistically significant results for this data set ($p < .05$).

Table 3.3: CFA *p* Value Descriptive Statistics

	CFA by PND	**CFA by PAND**	**CFA by PEM**	**CFA by IRD**	**CFA 30% WM**
Max *p*	.58	.58	.31	.42	.35
Min *p*	<.001	<.001	<.001	<.001	<.001
***p* Range**	.58	.58	.31	.42	.35
Mean *p*	.062	.062	.052	.041	.060
SD *p*	.11	.11	.067	.22	.078

Unique to CFA applied to SCED's, there existed a possibility that a B phase quadrant would have 0 data points, resulting in a divide by zero situation during calculation for that quadrant when using Stirling's approximation of the binomial formula. This was illustrated above in Figure 3.2. These occurrences were tracked and reported in table 3.4. This did not limit the analysis of the overall data set, as one quadrant would have all of the data points, and could still be analyzed for statistical significance. CFA by PND, PEM, and a 30%WM resulted in over half of the data points in one quadrant or the either. Additionally, the data point frequency differences between methods suggest that this is a result of which overlap method is selected to create the 2x2 matrix.

Table 3.4: Divide by Zero Percentage of Occurrence

	PND	**PAND**	**PEM**	**IRD**	**30% WM**
% DIV/0 Configuration	.667	.667	.536	.435	.530

As two significance values were calculated for each data set, agreement between upper and lower quadrant statistical significance was calculated for each CFA SCED configuration. Frequencies are presented below in table 3.5. CFA by 30% WM showed complete statistical significance agreement between upper and lower quadrants. Only 3.2% (3 Graphs) when using CFA by IRD and 1.3% (1 Graph) when using CFA by PEM resulted in upper and lower quadrant disagreement. The most disagreement resulted when using CFA by PND/PAND (13.7%, 23 graphs). The strong agreement between upper and lower phase B quadrants when using a 30% WM, PEM, and IRD may be due to how the 2x2 quadrants are created. These measures focus more on the center of the data points rather than the extreme values when establishing a line to extend from phase A data (for IRD, this is the case when taking data equally from both phase A and B) rather than selecting the highest or lowest data point to anchor the line (e.g. PND and PAND).

CFA in an RCT context may be different than CFA in a SCED context. In the context of RCT research, when some cells would be statistically significant and some would not, the exploratory nature of CFA should encourage the researcher to look deeper into what factored into the statistically significant cells. In the context of CFA applied to single-case data, disagreement or agreement falls upon a continuum of evidence for the treatment having an effect on the independent variable. If both cells in a 2x2 matrix for single-case data are statistically significant, this would provide the most evidence for a treatment effect. If one cell was statistically significant and another was not, this would be moderate evidence. If both cells were not statistically significant, this

would be no evidence of a treatment effect. As always, this continuum of evidence is not the end of treatment effect decision processes, and additional data such as effect size and visual analysis should be integrated to increase confidence in determinations made by the researcher.

Table 3.5: CFA Statistical Significance Agreement Between Quadrants

	PND	**PAND**	**PEM**	**IRD**	**30% WM**
% Agreement	87.3	87.3	98.7	96.8	100

As this study compared overlap methods that do not provide *p* values to CFA which does, statistical analyses compared how well method results correlated between effect size and statistical significance. These results are presented in table 3.6. First, a Pearson R was run between effect size by non-overlap method and CFA *p* value by each non-overlap method 2x2 configuration, and then effect size by overlap method compared to CFA using a 30% WM configuration. Results indicated that effect sizes were more highly correlated with *p* values when CFA was run with the non-overlap 2x2 configuration rather than by a 30% WM configuration. As expected, effect size and statistical significance were negatively correlated across the board. This indicated that in general, when effect size went up, *p* values got smaller. This correlation was strongest between CFA *p* values and effect size when using PND. Conversely, effect size was correlated less with *p* values when CFA was calculated using a 30% WM. This was the case across all non-overlap effect size estimation tools, as seen in table 3.6.

Table 3.6: Pearson *r* Correlations Between CFA *p* Values and Non-overlap Effect Sizes

	PND ES	PAND ES	PEM ES	IRD ES
CFA *p* value by non-overlap 2x2 configurations	-.71	-.41	-.50	-.58
CFA *p* values by 30% WM 2x2 configurations	-.36	-.34	-.41	-.34

Table 3.6. This table shows the Pearson *r* values when calculated between non-overlap effect size values and CFA *p* values when run using each non-overlap configuration. For example, Pearson *r* for effect size values for PND compared to CFA *p* values calculated using a 2x2 PND configuration is -.71.

As statistical significance can be considered a dichotomous nominal scale (significant or non-significant with alpha set at .05), a point biserial correlation was utilized to further compare effect size and statistical significance. Results are presented in table 3.7. The point biserial correlation compared effect size values with CFA statistical significance determinations (either statistically significant or not statistically significant) to further explore the relationship between effect size and statistical significance. All resulting correlation values were found to be statistically significant, calculated within the statistical package by comparing mean values of the two groups using a t test. This should not be confused as a t test comparing CFA *p* values directly to non-overlap effect size values, but rather a measure of the significance of the discrimination between effect sizes values when CFA was significant and when CFA *p* values were not significant. CFA by 30% WM statistical significance was consistently less correlated then CFA by a non-overlap 2x2 method. For example, PND & CFA by PND resulted in a .76 r_{pb}, while PND & CFA by WM resulted in a .27 r_{pb}. Both results were found to be statistically significant. Correlations between CFA statistical

significance, calculated using the 2x2 matrix of the effect size measure, and effect size values were consistently greater than when calculating CFA using a 30% WM. This echoes the results found in table 3.6, again suggesting that researchers choosing to use non-overlap effect size estimations paired with CFA should choose to do so using the non-overlap 2x2 configuration whenever possible, rather than applying a general measure of central tendency such as a 30% WM.

In addition to the echoed results from table 3.6, the point biserial correlation revealed a pattern where CFA with 30% WM always resulted in a smaller correlation than CFA based on the method used to generate the 2x2 matrix. When the effect size method is the same as the method CFA uses to calculate a *p* value, it makes sense that the correlation would be higher. This also makes interpretation easier because the quadrants are the same for both effect size and *p* value calculation. On the other hand, when this isn't possible the results suggest that although the correlation is smaller, CFA by a 30% WM may serve as an appropriate substitute.

Table 3.7: Point Biserial Correlation Between Effect Size and Statistical Significance

	PND & CFA by PND	PND & CFA by 30% WM	PAND & CFA by PAND	PAND & CFA by 30% WM	PEM & CFA by PEM	PEM & CFA by 30% WM	IRD & CFA by IRD	IRD & CFA by 30% WM
N non sig	51	58	51	58	55	58	38	58
N sig	117	110	117	110	113	110	130	110
$\mathbf{r_{pb}}$	.76	.27	.52	.26	.43	.36	.6	.28
***P* one-tailed**	<.0001	<.001	<.0001	<.001	<.0001	<.0001	<.0001	<.001
***P* two-tailed**	<.0001	<.001	<.0001	<.001	<.0001	<.0001	<.0001	<.001

Discussion

This study further verified what was found in the first of the series. CFA can be applied to SCED data using a non-overlap configuration. The calculations for observed and expected frequencies are intuitive in nature, valuable statistical significance evidence is obtained when applying CFA across multiple quadrants, and CFA can be used with different non-overlap technique configurations. Specifically, CFA was successfully applied with multiple methods to generate the 2x2 matrix for CFA (PND, PAND, IRD, PEM, and 30% WM). These findings add support to the results from the first study.

The first two articles have shown that CFA can be applied using many non-overlap methods to create 2x2 matrices. However, it was not known how CFA would compare to overlap methods that do not provide *p* values. Therefore, descriptive statistics and correlations were used to examine CFA and these non-overlap methods.

Similar to the previous study, some data sets were such that CFA could mathematically not be applied to all quadrants. When a B phase quadrant (either upper or lower) had no data points, that quadrant could not be evaluated. The denominator of Stirling's formula would be 0, resulting in a mathematical error. Within this study, 66.7% of CFA by PND/PAND, 53.6% of CFA by PEM, 43.5% of CFA by IRD, and 53% of CFA by 30% WM data configurations resulted in this complication. This limited the graph to one B phase quadrant for analysis. It was clear that this occurred more in simple non-overlap configurations (i.e. PND/PAND) than in configurations that split the data more evenly (i.e. PEM). As before, Stirling's formula has this limitation where as other more complicated formulas do not encounter this complication. None the less, each graph obtained a minimum of one CFA statistical significance value, and therefore every graph was interpretable for CFA statistical significance on some level.

Statistical significance disagreement between quadrants occurred at low frequencies (12.7% for PND and PAND; 3.8% for PEM; 3.2% for IRD; 0% for CFA by 30% WM). This may be a product of the data set, or potentially how the line to create the 2x2 configuration was drawn. Rather than having one statistical significance value for the data presentation, statistical significance could be calculated for increases and decreases in the dependent variable by using CFA on both upper and lower phase B quadrants. Small differences in observed and expected frequencies resulted in larger *p* values. The bigger the difference, the smaller the *p* value. When testing how the formula responded to changes in observed and expected frequency values by changing values in the Excel formula, Stirling's formula could obtain statistically significant

values with a minimum of 6 data points present across phase A and B, with 3 data points in each phase. This indicated that CFA may have the sensitivity required to detect treatment effects when few data points are present.

Only CFA by IRD averaged statistically significant *p* values. Standard deviations of *p* values ranged from .0673 (CFA by PEM) to .223 (CFA by IRD). The smallest standard deviation came from CFA by PEM configuration, which also had the most restricted range of *p* values. This highlights how IRD performed differently than other methods. It should be noted that the IRD results were based on Parker, Vannest, and Davis (2011) where "the preferable robust IRD is obtained by splitting the number of overlapping data points between phases." (p. 10).

Comparing CFA to established non-overlap methods that did not provide for *p* values presented some unique challenges when choosing statistical analysis methods. First, a direct comparison of the groups such as a t test was not appropriate, as CFA provided *p* values, and non-overlap methods provided effect sizes. As such, a correlation was utilized to determine if CFA *p* values and non-overlap effect sizes were appropriately correlated, both in direction and in size. When Pearson *r* correlations between CFA *p* values and non-overlap method effect sizes were run, it appears that overall they correlated moderately. At most, a .71 correlation was found between CFA *p* values by PND configuration and PND effect sizes, while CFA *p* values by PAND configuration and PAND effect size values resulted in the lowest correlation of .41. Consistently, CFA statistical significance values correlated higher with effect sizes when each was calculated using the same non-overlap effect size method. Conversely, CFA

calculated using a 30% WM correlated consistently lower with effect sizes from non-overlap methods. The stronger correlations suggest that when using CFA by Stirling's formula, it is more appropriate to use the chosen non-overlap method quadrant configuration to calculate CFA, rather than apply a general configuration such as a 30% WM. Conceptually, applying a statistical significance configuration congruent with the chosen effect size configuration could be perceived as more methodologically sound by maintaining the conceptual framework implied by the non-overlap method. The stronger correlations between CFA and effect size values when using the matching non-overlap configuration appear to support such a conclusion.

A point biserial correlation between non-overlap effect size and CFA *p* value (i.e. *p* either above or below .05) all resulted in statistically significant correlations. Again, this echoed the correlations reported earlier, indicating that stronger correlations were obtained when CFA was run by the overlap method rather than a 30% WM configuration. This adds support for running CFA by the matching 2x2 non-overlap configuration rather than a 2x2 matrix made using a 30% WM whenever possible. However, as all point biserial correlations were statistically significant, when the 2x2 matrix is not available due to the method selected a 30% WM appears to be an appropriate alternative to provide treatment effect evidence. More research will be needed to explore this potential, as this study selected methods that all provide a 2x2 matrix that CFA could use.

Despite limitations regarding Stirling's formula when there is no overlap in SCED data configurations, at a minimum CFA provides one statistical significance

value, providing evidence to determine treatment effect conclusions. This researcher suggests ignoring the quadrant that has no data, and simply analyzing the quadrant that does. This will continue to provide valuable statistical significance data without ignoring the researcher's conceptualization of the non-overlap presentation. Although a simple solution, this likely needs more attention to determine alternative solutions.

Finally, CFA may be a common statistical significance language across non-overlap analysis techniques. When presented with graphical data across many publications, non-overlap effect size methods can be performed along with CFA by the non-overlap method. For example, if presented with 10 graphs from 10 studies which all use different calculation methods, IRD and CFA by IRD could be run on the graphed data using the procedures outlined in this study. This would provide effect size and statistical significance for an overall treatment effect using a common language across all studies.

Overall, these results support the application of CFA to non-overlap SCED methods that do not provide for statistical significance. It indicates that CFA may add novel statistical significance information based on non-overlap data configurations in the determination of treatment effects. This study has shown that CFA can be effectively applied to SCED data, that this method is accessible and simple in that it requires no advanced statistical software or knowledge, and that it opens SCED data to additional analysis by evaluating both upper and lower quadrants for statistical significance.

Limitations

This study applied CFA using Stirling's formula and non-overlap method configurations, and compared it to non-overlap effect size estimation methods that do not provide for statistical significance calculations. Although this study managed to expand application of CFA to additional non-overlap measures which did not provide for statistical significance values, and therefor covered some limitations from the first article in the series, limitations of this study and of the application of CFA overall are still present. Some limitations from the first study remain, including the application to published data. This provided valuable external validity, but fell short of testing the mathematical limits of CFA, as well as the exploration of repeated scenarios across many data sets. A simulation study may be a good next step, as this would allow the exploration of mathematical limitations of Stirling's formula, as well as an opportunity to compare Stirling's formula to alternative formulas such as the chi-square test (which is based on the assumption that the expected frequencies are greater than 5).

The divide by zero scenario deserves additional attention as well. A simple solution, and potentially the only solution when using Stirling's formula, may be to ignore the quadrant that has no data and to simply analyze the quadrant that does. The effects of this decision, however, remain unknown. Would this make CFA no different than other statistical significance calculations? How much treatment effect evidence is lost when only one of the two quadrants is analyzed? Additional research is needed to explore how alternative formulas may perform with single-case data, and how these methods compare to Stirling's formula. Alternatively, the CFA non-overlap 2x2 matrix

could be drawn to ensure that data is present in all B phase quadrants. However, this may sacrifice using the non-overlap matrix drawn by the non-overlap effect size method chosen by the researcher. No matter the solution, additional time and research is necessary to determine the overall effect of this phenomena, and determine how alternative methods perform that can be applied to single-case data while addressing the divide by zero complication.

As discussed before, this research chose to focus on non-overlap methods due to the face valid comparison between CFA and non-overlap effect size estimation techniques. Further research may be warranted to compare CFA to alternative statistical methods. However, this presents complications as well, as it is unknown how CFA would be applied to SCED data without a 2x2 or similar configuration.

Furthermore, this study only applied CFA across phase A and B, and does not address other SCED design structures. SCED researchers and clinicians enjoy a wealth of design structures that help them make conclusions regarding treatment effects. Although evaluation across an AB transition is central to many of these design structures, there are certainly other design considerations that this article did not explore such as an ABAB design. Each of these factors and others deserve additional research for further exploration of CFA applied to SCEDs.

Conclusion

This research further strengthened conclusions made in the first article of the series, while expanding the utility of CFA to more non-overlap techniques. CFA was able to work with PND, PAND, IRD, and PEM, provide valid representations of data

configurations, and add valuable evidence for SCED treatment effects. Practically speaking, this evidence may be vital to researchers choosing to utilize SCEDs. For example, if choosing to use PND, simply adding a quick CFA analysis could be the difference in making weak, unfounded conclusions regarding data, or making strong, confident conclusions regarding treatment effects. As CFA performed across many non-overlap methods, at a low cost of execution requiring little time and simple calculations, it is this clinician's opinion that CFA should be applied whenever a non-overlap effect size estimation tool is used and a 2x2 or similar configuration is present. Furthermore, CFA should be applied using the non-overlap configuration chosen for effect size estimation analysis, rather than a general configuration such as a 30% WM. This further ensures that the statistical significance value is based on the data configuration that the researcher finds most appropriate for data analysis.

Per CFA's clinical utility, it appears that the multiple quadrant evaluation paired with the flexibility across non-overlap methods, and the moderate and appropriate correlations between effect sizes and *p* values show that CFA is a valuable tool providing evidence for treatment effect evaluation when using SCEDs. Due to this research using published data, the external validity of these results point to CFA as an appropriate tool in treatment effect evaluation in clinical studies. Based on the results of the first two articles in this series, researchers choosing to utilize CFA for clinical application should feel confident in the theoretical underpinnings and the initial performance of CFA applied to a wide range of SCED data presentations.

Although CFA presents as a valuable, flexible, and easy statistical significance calculation for SCEDs, as with any statistical method it was not without its limitations. However, this research sets the stage for clinicians and researchers alike to apply CFA across a wide range of applications, and to reap the benefits of an additional treatment effect evaluation tool. It is of utmost importance to remember that treatment effect evaluations, especially those using SCEDs, involves a holistic process in that no single tool should be the sole basis for evaluation. Rather, many tools such as effect size evaluation methods, statistical significance tools, and visual analysis should be incorporated in such a way that provides a broad picture of treatment effects. It is clear based on the present research that CFA is a value added tool when working to make these decisions.

CHAPTER IV

MANUSCRIPT #3: THE APPLICATION OF CONFIGURAL FREQUENCY ANALYSIS TO SINGLE-CASE DESIGNS: A COMPARISON TO VISUAL ANALYSIS TECHNIQUES

Introduction

This series of articles began by applying Configural Frequency Analysis to single-case experimental design data. With few complications, CFA was shown to perform as expected analyzing data from SCEDs. Specifically, CFA performed as expected when using non-overlap methods to produce the 2x2 data matrix for CFA. CFA was compared to other non-overlap methods for analyzing single-case data that provide effect sizes with estimates of statistical significance, as well as to non-overlap methods that only provide effect size estimates. CFA was successfully used with multiple non-overlap methods, and provided estimates of statistical significance.

To continue forward, CFA needs further evaluation against more SCED techniques, scenarios, and data presentations. In this spirit, the final article in this series of three compares CFA to a long-standing SCED evaluation technique, visual analysis. Visual analysis is a process by which the researcher looks for visual identifiers of a treatment effect on time-series data presented on a graph. This is a cornerstone of SCEDs. It allows for further analysis of data patterns and characteristics for the determination of treatment effects. It is an important adjunct to statistical analysis of

single-case data. As such, CFA was compared to visual analysis to answer a few basic questions.

Are findings between visual raters and CFA similar? For instance, when multiple raters determine a large overall treatment effect, would CFA find a statistically significant treatment effect? Additionally, would CFA line up with how visual raters see graphical data? Cognitive schemas and heuristics are alive and well in visual analysis. Raters may choose a simple visual analysis heuristics such as overlap or immediacy of treatment effect alone to make conclusions. Would CFA align with simple visual rating perspectives, or with other more complex rating schemes? This article set out to answer these questions and further explore the use of CFA in the field of SCED research.

Visual Analysis

Visual analysis has been used frequently in single-case data analysis, as it offers a quick and logical analysis of treatment effects, especially when treatment effects are large and clearly represented graphically. Data is placed on a longitudinal graph and split into baseline and treatment phases. Researchers make treatment effect judgments by examining level, trend, variability, immediacy of the effect, overlap, and consistency of data patterns across phases (Fisher, Kelly, & Lomas, 2003).

Researchers can use visual analysis to look for evidence of a relation between an independent and dependent variable (Richards, Taylor, Ramasay, & Richards, 1999), and determine the strength or magnitude of that relation (Kratochwill & Levin, 1992). Until recently, this was predominantly an informal process (Wampold & Furlong, 1981) and lacked guidelines for inferring treatment effects. Without guidelines for best

practices, researchers were likely to frequently disagree, as this was a predominantly subjective process.

Recently, guidelines were presented (Lane & Gast, 2013), but subjectivity remains inherent in the visual analysis process. The What Works Clearinghouse panel (Kratochwill et al.,2010) proposed standards as well, encouraging the use of two reviewers certified in visual or graphical analysis to verify a causal relation. Causal relationships and treatment effects are determined by visually identifying indicators within the following areas:

- Level, trend, and variability across phases.
- Immediate change upon introduction of treatment, overlap of data, and consistency of the data across phases.
- Factors such as history effects or a sudden change of level within a phase.

By visually identifying changes in level, trend, variability, immediacy of the effect, overlap, and consistency of data patterns across phases (Morgan, & Morgan, 2009), researchers can infer a significant treatment effect by detecting a minimum of three of these changes. Even with these guidelines, visual analysis remains largely unreliable, especially when performed in absence of additional treatment effect evaluation methods.

Of concern is the significant unreliability within and between raters. Multiple studies evaluating the reliability of visual analysis have found wide spread disagreement between raters (Ottenbacher, 1986; DeProspero & Cohen, 1979; Harbst, Ottenbacher, & Harris, 1991; Ottenbacher, 1990; Park, Marascuilo, & Gaylord-Ross, 1990; Brossart, Parker, Olson & Mahadevan, 2006). One study using expert raters found 27%

agreement when rating SCED graphs with statistically significant results, and 67% agreement for graphs without significant results (Park, Marascuilo, & Gaylord-Ross, 1990). Harbst, Ottenbacher, and Harris (1991) found that journal reviewers performed little better than untrained raters at graph judgment tasks. Ximenes, Manolov, Solomas, and Quera (2009) found that judges are only good at judging graphs that show no effect. If a hallmark of single-case designs is visual analysis, but trained raters cannot agree on observable evidence, additional methods for treatment effect evaluation are *necessary* to supplement the unreliable nature of visual rater conclusions.

Visual Analysis: Necessary but Not Sufficient

Based on publications analyzing visual analysis reliability, it is clear that visual analysis is too unreliable for researchers to be confident in determining treatment effects with visual analysis alone. However, visual analysis provides treatment effect evidence that cannot be obtained using effect size and statistical significance calculations alone. As such, visual analysis can be considered a necessary but not sufficient condition in the determination of a treatment effect. Necessary, because without a visual analysis of the data presented graphically, valuable evidence of treatment effects may go unnoticed. Statistical analyses of SCED data ignore the context of the treatment across time, such as non-overlap effect size estimations.

Visual analysis remains necessary and provides valuable evidence for the evaluation of a treatment. However, researchers such as Parker and Brossart (2003) and Campbell and Herzinger (2010) have proposed that statistical methods for SCED evaluation including effect size and statistical significance testing may improve one's

confidence to evaluate treatment effects, rather than relying on visual analysis alone. Complimentary methods may provide evidence to researchers using visual analysis so that they may be more confident in their conclusions. This study aims to evaluate CFA as a potential complimentary method to visual analysis.

Configural Frequency Analysis (CFA)

Here, we will briefly overview CFA. Then, we will dig deeper into a concept that may translate into SCED data, types and anti-types. For review, as described in the first article CFA provides statistical significance evaluations of bi- or multivariate cross-classifications of categorical variables (von Eye, 2002). Non-overlap single-case designs have been shown to provide applicable data presentations by overlaying 2x2 tables on top of graphical representations of SCED data. By doing so, expected and observed data frequencies in each quadrant can be calculated. This is done by calculating the percentage of data in phase A both above and below a non-overlap effect size line (i.e. ECL, PAND, or a measure of central tendency such as the mean, median, or winsorized mean). Then, expected values are determined by the total number of data points in phase B multiplied by upper A phase percentage, and lower A phase percentage. Then, using Stirling's formula for the approximation of the binomial, statistical significance can be calculated for both upper and lower quadrants in phase B. For a full review of this process, readers are encouraged to refer to the first article in this series. Next, we will dig deeper into a concept briefly described in the first article, types and anti-types as qualitative descriptions of CFA results.

The fifth step in the overall CFA process is an interpretation of the statistical significance value. This is done by determining if the data configuration constitutes a "type" or an "anti-type" (von Eye, Mair, & Mun, 2010). This can be considered a qualitative descriptor of the data presentation, rather than a quantitative addition to the statistical significance value obtained using Stirling's formula. Simply put, a cell would not be considered a "type" or "anti-type" if the data configuration in a quadrant was not statistically significant. For our purposes, any alpha value above .05 would not constitute a type. If not statistically significant, this is as far as this qualitative process goes. If CFA results in a statistically significant *p* value, additional steps are taken.

When statistically significant, either a "type" or an "anti-type" is determined to be present. This is based on the hypothesis of the experiment. For example, if a hypothesis predicts an increase of behavior due to a treatment effect, and an increase is present at a statistically significant level (e.g. $p < .05$), this is considered a "type". This is similar to rejecting a null hypothesis. CFA goes one step further, and presents a qualitative split for rejecting the null hypothesis. It can be rejected due to a statistically significant increase in the dependent variable, constituting a "type", or rejected due to a statistically significant decrease in the dependent variable, constituting an "anti-type."

Although this research is not specifically concerned with "types" and "anti-types", it is of interest that CFA allows one to qualitatively label the type of statistical significance depending upon the increase or decrease in the number of data points in the treatment phase of the cell being compared to the baseline cell. This further emphasizes the utility of CFA in treatment evaluation. Let us use an example of a teacher trying to

evaluate the effect of a behavioral intervention on socially appropriate behavior. The teacher hypothesizes that by implementing a token economy system, the child's frequency of socially appropriate behavior such as sharing, and using words like "please" and "thank you" will increase. However, rather than the behavior increasing, the behavior decreases, and by a statistically significant amount. This would constitute an "anti-type" data presentation. This may remind readers of a two-tailed statistical hypothesis test, and may not be quantitatively different. However, qualitatively CFA conceptualizes these presentations differently, and researchers choosing to utilize SCEDs could integrate these qualitative descriptions into their results.

Effect Size Estimation Methods for CFA Comparison to Visual Analysis

As was the case in the second article in this series, a direct comparison between visual analysis and CFA is not feasible. Visual analysis is a largely qualitative decision making process. As discussed, indicators of a treatment effect such as the immediacy of change leads visual raters to make decisions whether or not the treatment had an effect on the dependent variable. Consequently, the statistical analysis values provided by CFA are not directly comparable to visual analysis.

As non-overlap measures have been used throughout this series, and have shown to be appropriate for the application of CFA, they will again be used for this study. Readers interested in how each configuration is applied should refer to the first two articles in the series, as these go over the quadrant configurations made by the non-overlap effect size estimation tools, and the application of a robust mean when the non-

overlap method does not provide for a quadrant configuration. Here, we briefly review this process.

Non-overlap and central tendency calculation methods were used to create 2x2 configurations across graphed AB phase SCED data. Each method applied a line across both A and B phases. The make of the line (i.e. vertical location, slope, etc.) was dependent on the non-overlap or central tendency method. By doing so, 2 quadrants were created in phase A, and 2 in phase B, hence the 2x2 configuration. These 2x2 configurations are followed by CFA calculations, which utilize observed and expected data frequencies in the 2 phase B quadrants. These frequencies are then used in Stirling's Formula for the approximation of the binomial, which provides a statistical significance value. This value reflects the statistical significance of the observed versus expected frequency of data in phase B quadrants.

Methods used for comparison in this study include PND, PAND, PEM, IRD, ECL, NAP, Kendall's Tau-A, B, and U, and a 30% WM. Calculating statistical significance using CFA based on the 2x2 configuration of each of these methods, provided a wide range for comparison when using visual analysis and may provide insight as to what visual analysis schema raters are using when presented with graphically represented data.

Illustrative Articles

This study analyzed published single-case methodology data. Articles were collected from *The Journal of Behavior Therapy* and *The Journal of Behavior Modification*. Published articles between 1990 and 2013 were considered for analysis.

To test the utilization of CFA outside of what could be considered "best case scenarios", this study worked outside of the standards set by What Works Clearinghouse (Kratochwill et Al, 2010). The rationale was to see if CFA would function in instances where minimal data was available. Each graph was analyzed using only the first A and first B phase. 147 graphs were collected across the two journals. 168 data sets were pulled from these graphs, as some graphs presented multiple sets of data. From this sample of data sets, 30 were selected by two researchers to provide a wide variety of data sets for visual analysis. Aspects such as amount of data, clarity of overlap, potential A phase trend, and immediacy of effect were varied across graphs and agreed upon by the two raters prior to selection for visual analysis. For example, this researcher ensured that for the category of data volume, graphs with large amounts of data ($n > 30$), medium amounts of data ($10 < n < 30$), and small amounts of data ($n < 10$) across A and B phases were present.

Data Extraction and Analysis

Multiple steps were used for the extraction of graphs, digitization of data points, and analysis of the digitized data. To start, published graphs were saved in PDF form. This was done using the "Snipping Tool" from the 2010 version of Microsoft Office. This digital snapshot was then uploaded into GraphClick for Apple OS X, a digitizing program that provided values of data points based on locations on the PDF graph by scaling the X and Y axis to match the graph. Data points were then specified by "point and click" techniques on the computer.

Using these data, effect size estimations and statistical significance values were calculated for ECL, NAP, and Tau-U by procedures set out in Parker, Vannest, and Davis (2010). Effect size estimations were calculated for PND, PAND, PEM, and IRD set out in the same article. Statistical significance was calculated using CFA based on each 2x2 non-overlap configuration (i.e. CFA by PAND, CFA by PEM, CFA by 30% WM, etc.). Each of these non-overlap methods either set a line across A and B phases which created a 2x2 quadrant system, or if this line was not used in the non-overlap method (i.e. TAU, NAP) a 30% WM was used instead. Calculations were performed using the same practices as the previous two articles in the series, using "pen and paper" techniques, R scripts developed by Tarlow (2014), online calculation tools, tests of one proportion for *p* values, and Microsoft Excel for formulas involving expected/observed frequency calculations applied to Stirling's formula. For a full review of these specific techniques, readers are encouraged to review the first two articles in this series.

Visual Analysis Procedures

Each of the 30 graphs was analyzed using visual analysis by a class of 6 doctoral (PhD) students. All graduate students had experience in single case research methodology and design, including two or more courses, clinic or field work, and were at the end of the semester in an advanced course in SCEDs. Each graph was presented as a Microsoft Power Point slide. Students were presented one slide at a time and asked to characterize the overall treatment effect. Students were instructed to analyze the AB contrast holistically rather than by attempting effect size calculations in their heads or by use of a single dimension such as variability or overlap of data. Students were asked to

determine if each graph showed "no" treatment effect, or if there was a treatment effect if it was "small", "medium", or "large." Students were instructed to evaluate and record ratings independently from one another. After visual analysis, results were compiled and visual raters were able to anonymously review compiled results for educational purposes.

Results

Descriptive data regarding the 30 graphs that were visually analyzed and the results of the visual analysis are provided in Table 4.1. Of interest, half of the graphs presented to visual raters were determined to show either a "medium" or "large" treatment effect (i.e. visually significant), while the other half were determined to show either "no" or a "small" treatment effect (i.e. not visually significant). Variability between raters was high. For example, with 2 graphs raters were split between all four possible designations, from "no treatment effect" all the way to a "large treatment effect." 12 of the 30 graphs had ratings represented in at least 3 of the 4 categories. To further elaborate on this variability, agreement was calculated using Chronbach's alpha, using raters as "items". Chronbach's alpha was .86 as shown in table 4.1. This was similar to previous studies looking at agreement between raters (Brossart, Parker, Olson, & Mahadevan, 2006).

Table 4.1: Descriptive statistics for SCED graphs

Number of Graphs	30
Average Phase A Data Points	6.83
Average Phase B Data Points	20.0
Average Total Data Points	26.9
Median A Phase Data Points	5
Median B Phase Data Points	11.5
Median Total Data Points	17.5
Number of Graphs Rated Visually Significant	15
Number of Graphs Rated Not Visually Significant	15
Cronbach's Alpha	.86

For comparison purposes, visual analysis results were converted into "visually significant" and "visually not significant" treatment effect determinations. This allowed for comparisons with CFA, as statistical significance could be considered "significant" ($p < .05$) or "not significant" ($p > .05$). If the majority of raters determined the graph to show a "medium" and/or "large" treatment effect, the graph was classified as a "visually significant" treatment effect. Conversely, if the majority of raters determined the graph to show "no" or a "small" treatment effect, or if there was an even split between no/small and medium/large, the graph was classified as showing "no visually significant" treatment effect. For example, if 4 of the 6 raters determined that a graph showed a "medium" treatment effect, while 2 raters determined a "small" treatment effect, the graph would be classified as "visually significant." As seen in table 4.1, this led to an even split between raters, as half were determined to be "visually significant" and half were not.

Next, agreement between visual analysis and CFA by each configuration was calculated. As CFA produces two statistical significance values, the larger of the two p

values were used for comparison purposes. This was done to allow for convenient comparisons between visual analysis and CFA. In most cases this did not mean selecting the *p* value that was not statistically significant rather than the value that was, as the majority of methods showed 100% statistical significance agreement between upper and lower B phase quadrants. As a reminder, agreement was defined as the CFA *p* value in both upper and lower B phase quadrants either less than .05, or both greater than .05. Disagreement was considered when one value was above .05, and the other below.

Methods potentially affected by this larger CFA value selection were PND and PAND. For this selection of 30 graphs, PND and PAND graphs had 83% statistical agreement between B phase upper and lower quadrants. As such, 5 graphs showed one B phase quadrant with statistical significance and one without when using CFA by a PND or PAND configuration. Although this is an important feature of CFA when applied to single-case data, this study simplified this comparison for introductory purposes by selecting the larger of the *p* values for comparison methods. For these 5 graphs, this meant selecting the non-significant *p* value over the significant value. Although not ideal, this simplified comparisons to visual analysis, and aspects of CFA quadrant disagreement when using visual analysis may still be explored in later research.

After selecting the larger *p* value from the two, classifications were given to each graph. If CFA by the given non-overlap configuration (or 30% WM) resulted in a *p* value < .05, then CFA for that graph was classified as statistically significant. Based on this conceptualization, visual analysis raters and CFA statistical significance agreed little

over half of the time, as seen in table 4.2. CFA by PND and PAND configurations resulted in the highest level of agreement between visual analysis and CFA at 83.3%. This meant that of the 30 graphs presented, visual analysis and CFA agreed on the significance of the graph 25 times. For example, for any given data set, if CFA resulted in $p < .05$ (statistically significant), and the majority of raters determined either a "medium" or "large" treatment effect for that same data set, then CFA and visual analysis were determined to "agree." Results for all visual analysis and CFA statistical significance comparisons are presented in table 4.2. Percentages indicated close to chance levels of agreement between visual analysis raters and CFA by ECL, PEM, IRD, and 30% WM configurations. Alternatively, the highest level of agreement was recorded between visual analysis raters and CFA by PND and PAND.

Table 4.2: Visual Analysis Agreement

	Visual & CFA by ECL	**Visual & CFA by PND**	**Visual & CFA by PAND**	**Visual & CFA by PEM**	**Visual & CFA by IRD**	**Visual & CFA by 30% WM**
Frequency Agree	14	25	25	15	16	15
Frequency Don't agree	16	5	5	15	14	15
% agreement	46.6%	83.3%	83.3%	50%	53.3%	50%

A point biserial correlation between visual significance classifications treated as a dichotomous variable (i.e. visually significant or not visually significant) and CFA *p* values were used to further analyze the agreement between methods. Again, the larger of the two CFA *p* values was used to simplify the methods for comparison. Similar to

the basic agreement in table 4.2, PND and PAND were found to correlate the most with a $r_{pb} = .53$, significant for both one and two-tailed statistical significance as seen in table 4.3. These results indicate that visual raters may have been using a rating schema aligned with simple non-overlap measures (i.e. PND and PAND) rather than more complex analysis schemas. In contrast, although CFA by IRD and visual analysis resulted in only 53.3% agreement, shown in table 4.2, the point biserial showed a significant correlation between CFA by IRD configuration and visual analysis ($r_{pb} = .41$, $p < .05$) presented in table 4.3. The CFA by IRD and visual analysis correlations should be interpreted with caution, as the percentage of agreement shown in table 4.2 was only 3.3% higher than that of the agreement between CFA by 30% WM and visual analysis. Additionally, agreement between these methods is approximately 50%, indicating no greater than chance levels that methods would agree. Other methods (i.e. ECL, PEM, 30% WM) showed no significant correlation between the visual analysis determination and the CFA *p* value.

Table 4.3: Point Biserial Correlation between Visual Significance and CFA

	Visual & CFA by ECL	**Visual & CFA by PND**	**Visual & CFA by PAND**	**Visual & CFA by PEM**	**Visual & CFA by IRD**	**Visual & CFA by 30% WM**
N non sig	15	15	15	15	15	15
N sig	15	15	15	15	15	15
r_{pb}	.07	.53	.53	.04	.41	.04
P one-tailed	.346	.0014	.0014	.42	.013	.41
P two-tailed	.69	.0027	.0027	.84	.026	.83

To further elaborate on the relationship between visual analysis and CFA, a second point biserial correlation was calculated. Opposite of the point biserial represented in table 4.3 which used CFA *p* values correlated with visual rater "significance" determinations, this was calculated by correlating CFA statistical significance determinations (i.e. *p* either above or below .05) with visual rater value determinations (i.e. 0 = no effect, 1 = small effect, 2 = medium effect, and 3 = large effect). Results presented in table 4.4 indicate that CFA and visual analysis are not significantly correlated when CFA is run by the non-overlap methods used in this study. Although the first point biserial represented in table 4.3 indicated potential correlation between CFA by PND, PAND, and IRD configurations and visual analysis, this suggests that these two processes are not related.

Table 4.4: Point Biserial Correlation between Visual Ratings and CFA Significance

	Visual & CFA by ECL	**Visual & CFA by PND**	**Visual & CFA by PAND**	**Visual & CFA by PEM**	**Visual & CFA by IRD**	**Visual & CFA by 30% WM**
N non sig	25	18	18	18	23	20
N sig	5	12	12	12	7	10
r_{pb}	.02	.03	.03	.04	.09	.08
P one-tailed	.46	.45	.45	.41	.31	.34
P two-tailed	.92	.89	.89	.82	.63	.67

Finally, a point biserial correlation between the visual significance determinations and the effect size produced by non-overlap methods was calculated. Compared to CFA statistical significance, more significant relations were found between effect size and visual analysis than CFA and visual analysis. In total, 7 comparisons were found to have statistically significant correlations, as shown in table 4.5, compared

to 2 with visual analysis and CFA significance as seen in table 4.3. Specifically, statistically significant r_{pb} were found between visual analysis and PND, PAND, IRD, NAP, Tau-A, Tau-B, and Tau-U. This may indicate that visual raters tried to estimate an effect size to determine visual significance, as the statistical significance and visual analysis correlations were less frequent than the effect size and visual analysis correlations. Although some correlations were statistically significant, these correlations were not particularly large.

Table 4.5: Point Biserial Correlation Between Visual Significance and Effect Size

	VA & ECL	**VA& PND**	**VA& PAND**	**VA& PEM**	**VA& IRD**	**VA& NAP**	**VA& Tau A**	**VA& Tau B**	**VA& Tau U**
N non sig	15	15	15	15	15	15	15	15	15
N sig	15	15	15	15	15	15	15	15	15
r_{pb}	.09	.7	.53	.24	.45	.52	.56	.53	.49
***P* one-tailed**	.31	<.001	.0014	.096	.0065	.0017	<.001	.0011	.0032
***P* two-tailed**	.62	<.001	.0027	.19	.013	.0034	.0014	.0021	.0064

Table 4.5. VA = Visual Analysis

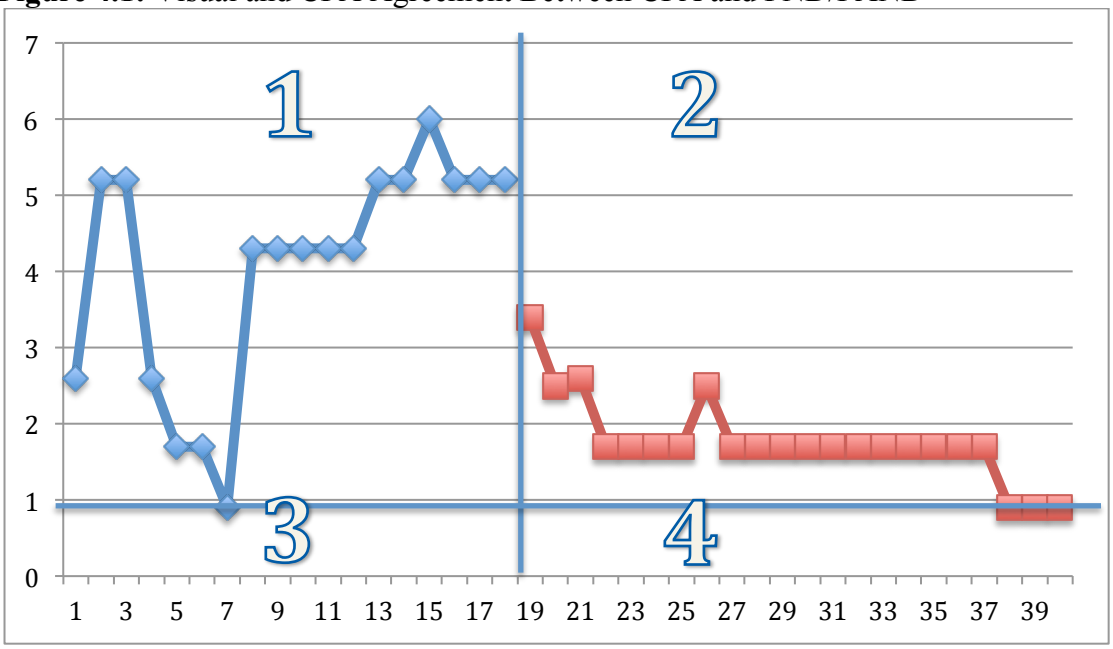
Discussion

Studies consistently indicate low agreement between visual analysis and quantitative analysis results such as effect size estimation techniques (Brossart et al., 2006; Jones, Weinrott, & Vaught, 1978; Park, Marascuilo, & Gaylord-Ross, 1990; Rojahn & Schulze, 1985). This study resulted in similar findings. Maximum agreement

occurred at a rate of 83.3% when comparing visual analysis to a CFA with PND or PAND configuration. More commonly, percentages hovered around 50%, indicating no greater than chance probability that CFA statistical significance and visual analysis raters would agree.

These findings suggest that more agreement between raters using visual analysis and CFA statistical analysis can be expected when utilizing a simple non-overlap measure for creating the 2x2 CFA configurations (i.e. PND or PAND). When CFA with PND/PAND match with visual analysis, it may be a result of clear non-overlap data configurations, as raters can easily visually identify when phase B has no overlap with phase A, as opposed to conceptualizing with median lines (i.e. PEM), trend (i.e. ECL) or by removing data points until there is no longer overlap (i.e. IRD). This is also the first hint into the type of data characteristics this sample of researchers attended to when conducting visual analysis. An example of when visual raters agreed with CFA by PND/PAND configuration, and disagreed with CFA by all other configurations (i.e. CFA by ECL, PEM, IRD, and 30% WM configurations) is presented in Figure 4.1.

Figure 4.1: Visual and CFA Agreement Between CFA and PND/PAND



In Figure 4.1, we see complete overlap between data, as the range of data in phase A is from 1 to 6, while the range of data in phase B is 1 to 3.5. All phase B data is within the range of phase A. When using a PND/PAND configuration for CFA, the lowest data point in phase A causes a split of data points. When extending a horizontal line from phase A, through the lowest data point (7,1), the data point passes through three points in phase B (38,1; 39,1; 40;1). When this happens, this researcher chose to "split" the data points, giving half the value to the upper quadrant, and half to the lower quadrant.

As such, quadrant 1 has 17.5 observed data points, while quadrant 2 has 20.5 observed data points. Quadrant 3 has one split data point (.5 observed data points), while quadrant 4 has 1.5, (3 data points that are split by the horizontal line). Expected

values, based on percentages of observed data in phase A, equal 21.38 for quadrant 3 (17.5 observed data points in quadrant 1 divided by 18 total in phase A = 97.2 % of A phase data in quadrant 1. .972 x 22 total phase B data points = 21.38 expected data points in quadrant 3). Similar calculations yield .62 expected data points in quadrant 4, with 1.5 data points observed. CFA calculations using Stirling's formula result in *p* values of .121 for quadrant 2, and .208 for quadrant 4. Both of these values are not considered statistically significant.

When this graph was analyzed by the 6 visual raters, 2 determined "no treatment effect", 1 determined a "small" treatment effect, 2 determined a "medium" treatment effect, and 1 determined a large treatment effect. By this researcher's classification of visual analysis ratings, this was determined to be a visually not significant effect, although there was considerable disagreement between the raters. Nonetheless, only 1 rater determined that Figure 4.1 represented a large treatment effect. As a reminder, the graph was shown to visual raters without non-overlap or trend lines such as those that are presented in Figure 4.1.

For Figure 4.1, CFA by PND/PAND configurations and visual analysis agreed that this treatment effect was not significant. However, when CFA was applied to this data using other configurations including ECL, PEM, IRD, and a 30% WM, all other configurations found that this data presentation was statistically significant. This suggests that visual raters may have over relied on a simple non-overlap conceptualization for visual analysis determinations. Additionally, this emphasizes the need to conceptualize a treatment effect in multiple ways to have a more complete

understanding of what a treatment effect may look like. Most importantly for this research, this example illustrates the flexibility CFA provides researchers, and should encourage those who choose to utilize CFA to apply the statistical significance evaluation across multiple configurations to obtain a more comprehensive picture of any potential treatment effects in their data.

For example, imagine a researcher relied on a visual analysis that indicated no significant treatment effect, and verified this conclusion by conducting a CFA with PND and found no statistically significant *p* values? If the researcher stopped there, the additional information available through CFA by other configurations would be ignored possibly leading to an error in their interpretation of the treatment effect. If CFA had been conducted with other methods (e.g., 30% WM) the researcher would probably look twice at the data, recognize a potential floor effect of this data set, recognize that the majority of the phase A data is above that of the phase B data, and consider the treatment for further evaluation based on additional results.

Similar findings were obtained by using the first point biserial correlation represented in table 4.3, as largest and most frequently statistically significant correlations were found between visual analysis and CFA statistical significance by PND, PAND, and IRD configurations. As these CFA statistical significances were calculated using non-overlap method configurations, this may again suggest that visual raters were using a data overlap conceptualization rather than a holistic view integrating trend, overlap, immediacy of effect, variability, and so on. This suggests that researchers should utilize multiple methods in generating the 2x2 matrix when applying

CFA in the analysis of single-case data. CFA by simple non-overlap may serve only to verify visual analysis rather than add additional information. Relying solely on non-overlap as an indicator of treatment effect may ignore important features such as immediacy of treatment effect, changes in general levels of dependent variables, changes in dependent variable trend, and may over-emphasize data points that cause complete overlap.

In contrast to the point biserial correlation represented in table 4.3, when this analysis was run with CFA (i.e. statistically significant or not statistically significant) and Likert scaled visual analysis ratings (i.e. large treatment effect = 3, medium treatment effect = 2, and so forth) no statistically significant correlations were found between CFA and visual analysis. The non-significant correlations between visual analysis and CFA statistical significance across many non-overlap methods suggests that clinicians and researchers should consider them independent of one another, each providing a unique source of treatment effect evidence.

Point biserial correlations between effect size and visual analysis ratings suggested that the visual analysts used a visual overlap conceptualization of treatment effect rather than a more multidimensional view of these data. Said another way, it appears that visual raters may have over attended to non-overlap rather than other features that indicate a treatment effect such as immediacy of treatment effect, changes in trend, or overall changes in level of data. The results suggest that this sample of raters may have been trying to estimate a quantitative value of effect size, despite being instructed not to do this. Supporting this notion, the point biserial correlations between

effect size and visual analysis were consistently greater and more often statistically significant than those between visual analysis and CFA statistical significance (77.8% for effect size vs. 50% for statistical significance).

CFA can be used to gain information not available through visual analysis or effect size estimation techniques, as well as expand the perspective of the visual rater when used with different methods to generate the 2x2 matrix. This is especially true if the user is willing to run CFA across several different non-overlap configurations. As such, whenever time and resources permit, SCED researchers who choose to utilize CFA along side visual analysis should do so using multiple methods to generate the 2x2 matrix CFA uses to determine statistical significance.

The take home from these findings is that CFA may aid in the conceptualization of data in ways that visual analysis by itself may overlook. At a minimum, the raters who participated in this study appeared to conceptualize their data using a simple non-overlap rubric to determine treatment effectiveness. Additionally, evidence suggested CFA was able to analyze single-case data in ways that seemed to be ignored by visual analysts. As discussed, there are several factors that should go into a visual analysis. However, non-overlap may be the easier, most apparent indication of treatment effect that this sample of visual raters leaned on. Additionally, visual ratings correlated more with effect size values then with statistical significance values. Even though CFA may use non-overlap effect size methods to establish 2x2 quadrant configurations, the data suggest that visual ratings align more with effect size values than CFA *p* values. Additionally, effect size estimation through visual analysis is likely an easier task than

statistical significance estimation through visual analysis. As such, CFA clearly added another way to assess treatment effectiveness.

It should go without saying that we are not suggesting CFA should replace visual analysis, or suggesting that visual analysis is too unreliable to be used. Rather, these data suggest that if we are to truly move forward with a holistic evaluation of treatment effects, statistical significance needs to be integrated into the process. Visual analysis provides for valuable evidence, but there appears to be significant overlap between visual analysis and effect sizes using non-overlap methods. As such, these findings suggest integrating CFA into the evaluation of treatment effects. When analyzing SCED data, this researcher advocates for visual analysis, paired with a non-overlap method that appropriately aligns with the researchers data conceptualization (i.e. trend, overlap, etc.), and then applying CFA using the non-overlap methods and comparing these results with both effect size results and visual analysis. By doing this, and integrating effect size estimation tools that conceptualize treatment effects not based solely on non-overlap the investigator could achieve a more complete understanding of the treatment effect. This would require integrating CFA into the treatment effect evaluation process.

Limitations

This paper focused on applying CFA to SCED data for the purpose of treatment effect analysis. Specifically, this study compared CFA statistical significance using Stirling's formula to visual analysis, a long standing and important piece of single-case research methodology. As a review of literature indicated that this had not been done before, this study was largely exploratory in nature. This limited the study to broad-

brush strokes. As such, there are a few considerations that future researchers should consider when digging deeper into CFA and visual analysis.

Future studies may benefit from targeting specifics such as exploring if applying a CFA configuration prior to visual analysis aids or hinders treatment effect analysis. Additionally, there are many types of SCED data, and this study selected a wide range of data arrangements to provide a foundation for this exploratory comparison. This limited the study to exploring many different arrangements, without digging deep into any one aspect in particular (e.g. graphs which include trend lines versus ones that do not, graphs that have large immediate treatment effects versus those that do not, etc). Future research could dig deeper by comparing CFA to visual analysis against many graphs of one type of graphical data aspect.

One limitation of the visual analysis was that only one small group of visual raters was used. As all raters were trained the same way at the same university, this may have systematically effected the accuracy of their ratings. Future studies may benefit from a diverse sample of visual raters.

A significant limitation was found when classifying "visually significant" results. This study asked raters to analyze graphs and determine either a "no", "small", "medium", or "large" treatment effect was present rather than asking for a visual analysis of statistical significance. There may be a few concerns with this conceptualization. First, this is not as direct of a comparison as it could be, as conceptualizations of "treatment effect" and "treatment significance" are two very different constructs. Also, this arrangement may have led to large variability in visual

analyst ratings. Further research comparing statistical significance estimations by visual raters and CFA statistical significance may benefit from using a different visual rater classification system or rating scheme, and may result in further clarification in how visual analysis and CFA statistical significance compare. Related to this limitation was the decision to select a non-significant *p* value over a significant value. Although this did not occur frequently (5 graphs total when using PND or PAND), this may have effected comparison results between visual analysis and CFA.

Conclusion

The purpose of this research was to compare CFA using Stirling's formula to existing visual analysis practices, and to explore what CFA may add to the decision making process. Based on evidence presented in this article, it appears that CFA may add a new perspective by providing treatment effect evidence that visual raters may not consider. First, visual analyses aligned more with effect size estimations than it did with statistical significance values. This suggests that visual analysis and many non-overlap effect size methods may have some overlap in the information they provide, but that CFA using other methods of 2x2 matrix generation can step outside of these boundaries and offer evidence of treatment effects that are not considered by visual analysts.

Also of interest, CFA more consistently aligned with visual analysis using simple non-overlap methods for effect size calculations. This is an issue if researchers rely on evidence from CFA by a simple non-overlap configuration and also evidence from a simple non-overlap effect size estimation method. Results from this research suggest that these two pieces of evidence may be a replication of one another, rather than

different in the way they evaluate a treatment effect. This research suggests that visual raters should be careful when applying CFA using a non-overlap 2x2 configuration such as PND/PAND which emphasizes simple overlap, and should strongly consider applying additional CFA configurations such as IRD or ECL to make more informed determinations. CFA by PND/PAND and visual analysis may only serve to repeat what is already known, rather than add new information. More research is needed to determine if this is the case.

Based on previous research, the high rate of disagreement between CFA and visual analysis was expected. Additionally, this article aligned with previous research that showed visual analysis inter-rater reliability to be quite low. The 6 raters who participated in this study showed wide variability in their visual analysis ratings. Only one graph resulted in all 6 raters choosing the same rating, selecting a "medium treatment effect". It is interesting to note that the findings suggest that visual raters may use a simple non-overlap effect conceptualization of treatment effectiveness rather than a more holistic approach to visual analysis, but this merits further investigation.

Even with poor reliability between raters, the take away from this study is that one may improve the process of treatment evaluation by integrating multiple perspectives or methods when evaluating SCED data. This research suggested that CFA could be that additional piece of data needed for treatment effect evaluation, as CFA provided a statistical significance perspective that was not covered in other methods. Additional research regarding the integration of CFA with visual analysis when making treatment effect decisions may help to determine how best to go about this combination

(i.e. which method should be considered first, do CFA *p* values positively or negatively influence visual analysis decision accuracy, etc.).

CFA may fill a gap in treatment effect evaluation by providing statistical significance, a concept that raters don't appear to naturally consider. As such, this researcher proposes integrating CFA into visual analysis procedures, and doing so across multiple non-overlap configurations. By applying CFA with a few different configurations (e.g. one method that emphasizes highest data point such as PND and another that emphasizes A phase trend such as ECL), a researcher would gain a wealth of statistical significance information that could go a long way in making decisions regarding treatment effects. Additionally, it may be important to apply CFA to other methods besides those that only look at non-overlap configurations. This may provide treatment effect evidence not covered by visual analysis. It is important to remember that one statistical analysis should not be the sole determining factor of an overall treatment effect. Factors such as design, context, visual analysis of graphed data, and multiple effect size calculations based on different techniques should be considered in order to accurately evaluate a treatment effect. Although additional research is warranted, CFA should be strongly considered for integration when choosing visual analysis for SCED data evaluation.

CHAPTER V

SUMMARY

This venture started with a simple question. What can we do to improve our treatment effect decisions from SCEDs? When exploring the possibilities, one stood out that could provide a new perspective and new evidence for data evaluation. Traditionally a large group exploratory analysis, CFA had the potential to analyze single-case design data. It appeared that, if translated into a four-quadrant system, non-overlap effect size estimation tools aligned with data configurations necessary to run CFA.

This research began with published SCED data, which was compared across eleven different evaluation tools, including CFA using Stirling's approximation of the binomial, and visual analysis. This broad introductory application would ensure that an adequate initial examination of CFA was conducted while exposing its limitations as well. As a result of this initial investigation it was hoped that we could determine if CFA was worth using in the analysis of data from SCEDs..

To begin, CFA was compared to three different categories of evaluation tools. The first comparison was with non-overlap effect size measures that produce statistical significance values. This provided a direct comparison, as CFA analysis also produces statistical significance values. This first study also examined issues related to the application of CFA to single-case data. Could CFA be applied to SCED data? How would expected and observed values be calculated? Could CFA be compared to

methods like Kendall's Tau, which doesn't split graphical data into a four-quadrant system? Can CFA be applied to published SCED data? With few complications, CFA was shown to perform as well as comparable measures of statistical significance. In addition, CFA was shown to provide additional, valuable information that existing methods may not provide.

Next, CFA was compared to non-overlap tools that provide for an estimate of effect size, but that do not provide a statistical significance value. CFA was shown to again be a value added procedure for researchers choosing to use SCEDs. CFA could be applied in the same context as the non-overlap measure, provide statistical significance values, and increase the amount of evidence researchers had at their disposal to evaluate treatment effects. These comparisons suggested that CFA performed well with non-overlap methods that easily produce a 2x2 matrix for analysis. When the four quadrants are not present, such as in Kendall's Tau, a measure of central tendency proved a potential substitute. A 30% WM allowed for the evaluation of single-case data when a 2x2 configuration was not a part of the effect size method. However, the correlations between CFA run with a 2x2 WM configuration and established statistical significance value methods were consistently smaller than those of CFA run with a 2x2 configuration based on the non-overlap methods used to calculate the statistical significance values. Although this offers a substitution when effect size methods do not provide a 2x2 configuration that CFA can use, more research is needed to determine if a 30% WM is the best method available. For now, a 30% WM may be an appropriate substitute until further research clarifies this process.

The next step was to compare CFA to visual analysis, a signature method of SCED data evaluation. Visual analysts were instructed to rate graphs holistically, looking for indicators that the treatment had an effect, and then determining if this effect was small, medium, or large. CFA once again proved valuable to the SCED data analysis process. The results suggested that visual raters were utilizing a simple non-overlap conceptualization of treatment effect as the results aligned well with non-overlap estimators of effect size. Alternatively, visual analysis results were moderately correlated with CFA results, indicating that CFA was providing a measure related to visual analysis, but independent enough to provide evidence visual ratings may not already cover.

In sum, CFA appeared to perform well when compared to other statistical methods and visual analysis. Also, CFA was shown to provide unique information for the evaluation of treatment effects. This suggests that CFA should be utilized as an additional statistical significance analysis tool. Caution, as always, should be exerted when considering using one measure alone, and this is the same with CFA.

CFA does not provide for an estimation of the effect size, nor does it provide a qualitative analysis such as that obtained when using visual analysis. Instead of replacing any one measure, it is suggested that CFA be added to the mixture of analytic tools. If using a non-overlap tool that provides for statistical significance, here we suggest using CFA as an additional check to the values produced, and to take into account the statistical significance values produced using CFA. When choosing measures that do not provide for statistical significance, CFA should be used to maintain

the context of the non-overlap technique. Finally, when using visual analysis, CFA should be added to the mix using multiple overlap techniques, as CFA appears to provide statistical significance evidence not considered when using visual analysis. Researchers choosing to use CFA along with visual analysis should also apply CFA across 2x2 configurations that go beyond non-overlap (i.e. PND/PAND) but also consider other factors as well (i.e. IRD, ECL, WM).

To continue forward and be applied to more scenarios, CFA will require additional research. For example, other methods besides Stirling's formula could be compared and may avoid the divide by zero problem inherent in Stirling's formula. Also, this research focused on non-parametric, non-overlap methods of SCED data analysis, and did not venture into parametric or regression based techniques. As such, it is clear that CFA's full potential will likely remain untapped unless further studies take place. Despite this need for additional research, it appears that CFA can be applied to non-overlap configurations of SCED data, can provide valuable treatment effect evidence, and can do so in a simple and accessible manner for those with limited statistical knowledge. This user-friendliness fits with the accessibility of single-case research, and likely sets the stage for CFA to be applied across a wide range of behavioral psychology disciplines.

REFERENCES

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single-case. *Behavior Research and Therapy, 31*, 621-631. doi: 10.1016/0005-7967(93)90115-B

Beretvas, S. N., & Chung, H. (2008). An evaluation of modified R^2-change effect-size indices for single-subject experimental designs. *Evidence-based Communication Assessment and Intervention, 2,* 120-128. doi: 10.1080/17489530802446328

Bergman, L. R. & von Eye, A. (1987). Normal approximations of exact tests in configural frequency analysis. *Biometric Journal, 29,* 849-855. doi: 10.1002/bimj.4710290714

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30,* 531-563. doi: 10.1177/0145445503261167

Brossart, D. F., Parker, R. I., & Castillo, L. (2011). Robust regression for single-case data analysis: How can it help? *Behavior Research Methods*, 43, 710-719. doi: 10.3758/s13428-011-0079-7

Brossart, D.F., Vannest, K. J., Davis, J. D., & Patience, M. A. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychological Rehabilitation: An international Journal*, 24, 464-491. doi: 10.1080/09602011.2013.868361

Busk, P. L., & Serlin, R. C. (1992) Meta-analysis for single-case research. In Kratochwill, T. R. & Levin, J. R. (Eds.), *Single-case research design and analysis: New direction for psychology and education*, (pp. 187-212). Hillsdale, NJ; Erlbaum.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification 28,* 234-246. doi: 10.1177/0145445503259264

Campbell, J.M. & Herzinger, C. V. (2010). Statistics and single subject research methodology. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences,* (pp. 91-109). New York, NY: Routledge.

Donoho, D. L. & Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness, *The Annals of Statistics, 20,* 1803-1827. doi: 10.1214/aos/1176348890

Goodman, L. A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data. *Journal of the American Statistical Association, 86,* 1085-1111. doi: 10.1080/01621459.1991.10475155

He, X., Simposon, D. G. & Portnoy, S. (1990). Breakdown robustness of tests. *Journal of the American Statistical Association, 85*, 446-452. doi: 10.1080/01621459.1990.10476219

Huitema, B.E. & McKean, J.W. (2000). Design specification issues in time-series intervention models. *Educational & Psychological Measurement, 60,* 38-58. doi: 10.1177/00131640021970358

Indurkhya, A., & von Eye, A. (2000). The power of tests in configural frequency analysis. *Psychologische Beiträge, 32,* 723-737. Retrieved from http://www.statmodel.com/bmuthen/alka2cfa.pdf

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). *Applied linear regression models,* 4th ed. New York, NY: McGraw-Hill.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification,* 30, 598-617. doi: 10.1177/0145445504272974

Ma, H. H. (2009). The effectiveness of intervention on the behavior of individuals with autism: A meta-analysis using percentage of data points exceeding the median of baseline Phase (PEM). *Behavior Modification, 3*, 339-359. doi: 10.1177/0145445509333173

Parker, R.I., Brossart, D.F., Callicott, K.J., Long, J.R., De-Alba, R.G., Baugh, F.G., et al. (2005). Effect sizes in single case research: How large is large? *School Psychology Review, 34,* 116-132. Retrieved from http://drsmorey.org/bibtex/upload/Parker:etal:2005.pdf

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34,* 189-211. doi: 10.1016/S0005-7894(03)80013-8

Parker, R.I., Hagan-Burke, S., & Vannest, K. (2007). Percent of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education,* 40, 194-204. doi: 10.1177/00224669070400040101

Parker, R. I. & Vannest, K. J. (2009) The improvement rate difference for single-case research. *Exceptional Children, 2,* 135-150. doi: 10.1177/001440290907500201

Parker, R.I., Vannest, K.J., & Davis, J.L. (2010). Effect size in single case research: A review of nine non-overlap techniques. *Behavior Modification*, 35, 302-322. doi: 10.1177/0145445511399147

Parker, R. I., Vannest, K. J., Davis, J. L. & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42,* 284-299. doi: 10.1016/j.beth.2010.08.006

Preston, D., & Carter, M. (2009). A review of the efficacy of the picture exchange communication system intervention. *Journal of Autism and Developmental Disorders, 39,* 1147-1486. doi: 10.1007/s10803-009-0763-y

Sackett, D. L., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (Eds.). (1997). *Evidence-based medicine: How to practice and teach EBM.* London: Churchill Livingstone. doi: 10.1016/S0146-0005(97)80013-4

Scruggs, T. E., Mastropieri, M. A. & Casto, G. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial and Special Education,* 8, 24-33. doi: 10.1177/074193258700800206

Smith, S., Vanest, K.J., Davis, J.L. (2011). Seven reliability indices for high stakes decision-making: description, selection, and simple calculation. *Psychology in the Schools, 48*, 1064-1075. doi: 10.1002.pits.20610

Solanas, A., Rumen, M., & Patrick, O. (2011). Estimating slope and level change in N = 1 designs. *Behavior Modification, 34,* 195-218. doi: 10.1177/0145445510363306

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing.* New York: Wiley. doi: 10.1002/9781118165485

Tarlow. K. R. (2014). Kendall's Tau and Tau-U for single-case research. R script. Retrieved from http://www.ktarlow.com/stats

Vannest, K.J., Parker, R.I., & Gonen, O. (2011). Single Case Research: web based calculators for SCR analysis. (Version 1.0) [Web-based application]. College Station, TX: Texas A&M University.

Vickers, A. (2005). Parametric versus non-parametric statistics in the analysis of randomized trials with non-normally distributed data. *BMC Medical Research Methodology, 5:35.* doi: 10.1186/1471-2288-5-35

Von Eye, A. (1990). *Introduction to Configural Frequency Analysis: The search for types and antitypes in cross-classifications.* Cambridge, UK: Cambridge University Press.

Von Eye, A. (2002). *Configural frequency analysis: methods, models, and applications.* Mahwah, NJ: Erlbaum.

Von Eye, A. (2007). Configural frequency analysis. *Methodology, 3,* 170-172. doi: 10.1027/1614-2241.3.4.170

Von Eye, A., Mair, P. & Mun, E. Y. (2010). *Advances in Configural Frequency Analysis.* New York, NY: Guilford Press.

White, O.R. & Haring, N.G. (1980) *Exceptional Teaching.* 2nd ed. Columbus, Ohio: Charles E. Merrill.

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing, 3rd Edition.* Waltham, MA.: Academic Press.