COMPARING SINGLE-CASE DESIGN NON-OVERLAP METRICS AND VISUAL ANALYSIS EXAMINING SCHOOL-BASED INTERVENTIONS FOR STUDENTS WITH AUTISM SPECTRUM DISORDER

by

FAHAD M ALRESHEED

A DISSERTATION

Presented to the Department of Special Education and Clinical Sciences
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

September 2018

DISSERTATION APPROVAL PAGE

Student: Fahad M Alresheed

Title: Comparing Single-Case Design Non-Overlap Metrics and Visual Analysis Examining School Based Interventions for Students with Autism Spectrum Disorder

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Special Education and Clinical Sciences by:

Dr. Wendy Machalicek	Chairperson
Dr. Kent McIntosh	Core Member
Dr. Roland Good	Core Member
Dr. Kathleen Scalise	Institutional Representative

and

Janet Woodruff-Borden	Vice Provost and Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded September 2018

© 2018 Fahad M Alresheed

DISSERTATION ABSTRACT

Fahad M Alresheed

Doctor of Philosophy

Department of Special Education and Clinical Sciences

September 2018

Title: Comparing Single-Case Design Non-Overlap Metrics and Visual Analysis Examining School Based Interventions for Students with Autism Spectrum Disorder

High prevalence of individuals with autism spectrum disorder (ASD) and the legislation movement impacted the placement of students with ASD in general education settings. Hence, the increase raised the need to conduct research for ASD populations, and to examine the effectiveness of these interventions. With the increase of single case-design (SCD) studies, there is a demand to include SCD in the evaluation of evidence-based practices (EBPs), to analyze and interpret SCD results in meaningful ways beside visual analysis, and to generate effect size estimates. This dissertation contains four systematic literature reviews which examine single-case intervention research targeting academic, social communication, play, and functional life skills for children with ASD in school settings. 132 studies with 924 AB phase contrasts were analyzed using visual analysis and three non-overlap measures. Sensitivity and specificity of Tau-U, IRD, and Baseline Corrected Tau were tested on detecting intervention effects. Also, the three methods were examined in their agreement with interpretations based on the visual analysis and the effect of confounding factor on their scores. The analysis demonstrated that the three methods performed fairly well in distinguishing effective from non-effective interventions. The three non-overlap methods had a moderate to substantial level of agreement with visual analysis. The author recommended further research on the

impact of confounding factors especially baseline trend and autocorrelation as well as the use of effect size methods with high sensitivity and visual aids to improve the reliability and accuracy of visual analysis.

CURRICULUM VITAE

NAME OF AUTHOR: Fahad M Alresheed

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Texas A&M University, Commerce
King Saud University, Riyadh

DEGREES AWARDED:

Doctor of Philosophy, Special Education, 2018 University of Oregon
Master of Science, Special Education, 2013 Texas A&M University
Bachelor of Education, Special Education, 2001 King Saud University

AREAS OF SPECIAL INTEREST:

Early intervention for children with developmental delays
Parent Education and Training
Cultural Adaptation of Evidence-Based Practices

PROFESSIONAL EXPERIENCE:

Special Education Teacher, Department of Education, Saudi Arabia, 2000-2010

Undergraduate Special Education Student Supervisor, Department of Education, Saudi Arabia, 2005-2009

GRANTS, AWARDS, AND HONORS:

Academic Excellence, Texas A&M University, Commerce, 2013

Saudi Arabia Cultural Mission Scholarship, Department of Education, Saudi Arabia, 2012

PUBLICATIONS:

Alresheed, F., Machalicek, W., Sanford, A., & Bano, C. Academic and related skills interventions for autism: A 20-year systematic review of single-case research. *Review Journal of Autism and Developmental Disorders.*1.16

Scalise, K., Irvin, S., **Alresheed, F**., Zvoch, K., Yim, H.,Park, S., Landis, B., Meng, P., Kleinfelder, B., Halladay, L., & Partsafas, A. (2018). Accommodations in digital interactive stem assessment tasks: Current accommodations and promising practices for enhancing accessibility for students with disabilities. *Journal of Special Education Technology*.

Machalicek, W., *LeQuia, J., *Pinkelman, S., *Knowles, C., *Raulston, T., Davis, T., &***Alresheed, F**. (accepted, in press). Behavioral telehealth consultation with families of children with autism spectrum disorder. *Behavioral Interventions*.

Hott, B. L., **Alresheed, F. M.,** & Henry, H. R. (2014). Effects of peer tutoring interventions for students with Autism Spectrum Disorders: A meta-synthesis. *Journal of Special Education and Rehabilitation, 15*, 109-121. doi: 10.2478/JSER-2014-0007

Alresheed, F. M., Hott, B. L., & Bano, C. (2013). Single subject research: A synthesis of analytic methods. *Journal of Special Education Apprenticeship*, *2*(1), n1.

ACKNOWLEDGMENTS

This work would not have been possible without the financial support of the Saudi Arabian Cultural Mission scholarship. I am especially indebted to Dr. Wendy Machalicek and her support during my PhD program. In addition, I am grateful to Dr. Roland Good, Dr. Kent McIntosh, and Dr. Kathleen Scalise whom I have had the pleasure to work with during my dissertation. Last, I wish to thank my passed-away parents who would be happy for me because I accomplished such a worthy goal.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

Table Page

CHAPTER I

INTRODUCTION

Autism Diagnostic Criteria and Prevalence

Autism Spectrum Disorder (ASD) is a group of neurodevelopmental disorders characterized by restricted and repetitive behaviors and interests as well as deficits in social communication. These symptoms are present from early childhood (American Psychiatric Association, 2013). Often, children with ASD show comorbid disorders (Jang & Matson, 2015) and challenging behaviors such as aggression and inattention (Mazurek, Kanne & Wodka, 2013), attention deficit/hyperactivity disorder, mood disorders, and anxiety disorders (Mannion, Leader & Healy, 2013; Matson, Rieske & Williams 2013; Van Steensel, Bögels & Perrin, 2011). Other comorbid disorders include epilepsy and sleep problems (Mannion et al.). The Center for Disease Control and Prevention latest estimate shows that one in 68 children have been identified with ASD in the United States. ASD occurs across all racial, socioeconomic, and ethnic groups and it is almost five times more common among boys than girls (CDC, 2014). ASD symptoms impact several areas of development and can adversely affect the ability to function in society as well as learn new skills.

Characteristics of Autism

Deficits in social communication and repetitive behaviors can increase the risk of developing challenging behaviors thus affecting children's educational goals and the way to reach them, as well as their community involvement (Sigafoos, Arthur-Kelly & Butterfield, 2006). Types of social communication delays include limitations in initiating conversation, requesting information, listening and responding, and interacting with

others (VanMeter, Fein, Morris, Waterhouse, & Allen, 1997; Volkmar, Carter, Grossman, & Klin, 1997). Horner, Carr, Strain, Todd, & Reed (2002) identified the primary challenging behaviors seeing in children with ASD as aggression, property destruction, tantrums, self-injury, and stereotypies. Children with ASD are also at increased risk for comorbid disorders including challenging behavior like aggression and inattention (Hartley, Sikora, & McCoy, 2008), attention deficit/hyperactivity disorder, depression and mood disorders, and anxiety disorders (Ghaziuddin, Ghaziuddin, & Greden, 2002; Matson & Williams 2013; Van Steensel, et al., 2011). Social communication delays, repetitive behaviors and comorbidity disorders may interfere with academic progress and friendship development (Thiemann, & Goldstein, 2001). Children with ASD and comorbid disorders need individualized and intensive instruction, creating great challenges to the local education agencies serving increasing numbers of children with ASD (Bagatell, Mirigliani, Patterson, Reyes & Test, 2010; Centers for Disease Control and Prevention, 2010; Carnahan & Williamson, 2013).

Academic, Social Communication, Play, and Functional Life Skills

Children with ASD need to learn a variety of skills that will allow them to learn, work and live successful lives. First, academic skills, including math and prerequisite skills, and general knowledge are important for children with ASD to function in society. Unfortunately, most research in this area targets literacy and less research has been done in other academic areas (Pennington, 2010; Spencer, Evmenova, Boon & Hayes-Harris, 2014). For example, there are few studies targeting science and social studies curriculum for students with ASD. Furthermore, studies targeting prerequisite skills for academic performance are very limited in number (Alresheed, Machalicek, Sanford, & Bano,

2018). Second, social communication skills are essential for development. For example, a child with ASD will have more opportunities to engage in school activities, interact and play with others and make requests if he or she can initiate conversations (Kagohara et al., 2012). Third, play skills are important because they help the child acquire social and language skills (Pierucci, Barber, Gilpin, Crisler & Klinger, 2015). Finally, children with ASD must acquire functional life skills to be independent. Individuals lacking these skills will struggle to maintain relationships and be productive members of society (Ayres, Mechling & Sansosti, 2013).

Legislation and court decisions in the last 40 years have had great impact in the delivery of treatment of children with ASD as well as research in this field. The Education for All Handicapped Children Act passed of 1975 as well as the Americans with Disabilities Act of 1990 established the right of children with disabilities to a free and appropriate public education in the least restrictive environment (IDEA, 2004). In 2010, it was estimated that 38.5% of students with ASD were in general education 80% or more of the time they were in school (U.S. Department of Education, 2012). Furthermore, several court decisions have entitled children to rehabilitative and educational as opposed to custodial and institutional types of programs. At the same time, research underscored the importance of interaction with peers as a source of learning for children with ASD (Carter, Sisco, Chung & Stanton-Chapman, 2010). As a consequence of all these factors, there has been a greater focus on interventions implemented in school settings, particularly in general education classrooms. Programs have introduced school-based and home-based instruction (Reichow, Doehring, Cicchetti & Volkmar, 2010). Schools are a great environment for delivering interventions to students with ASD.

Children are in school many hours a day and for most of their developing years. Schools allow the possibility of delivering intensive and comprehensive interventions (Koegel, Matos-Freden, Lang & Koegel, 2012).

Effectiveness of Interventions

Considering the need for individualized instruction combined with a growing emphasis on inclusive education, higher education, employment, and independent living for children with ASD, the identification of effective interventions is critical (Callahan, Henson & Cowan, 2008; IDEA, 2004; Wong et al., 2015). A diagnosis of ASD can have great impact on individuals and their families who are confronted with a wide variety of treatments often without evidence of effect. Consequently, it is critical that we have a way to identify what has been done and what seems to be a promising intervention for these individuals. A variety of organizations have developed guidelines and methods to evaluate empirical evidence to help identify evidence-based practices (EBPs) for the treatment of individuals with ASD (Reichow, et al., 2010). Evidence-based practice is now the standard in education, especially in the special education field (Odom, 2009).

Single-Case Designs and Evidence-Based Practices

Studies employing single-case designs (SCD) have enhanced the special education field by providing evidence for practices that have benefited individuals with disabilities and their families (Horner, Carr, Halle, Mcgee, Odom and Wolery, 2005). The Council for Exceptional Children Working Group (2014) stresses the importance of SCD results for clinical and public policy decisions. Furthermore, their researchers suggest that findings from SCD should be considered for inclusion in systematic reviews and meta-analyses that aim to synthesize the existing evidence about intervention effects. Studies

using SCDs have been recently included in the process to identify evidence-based practices (EBPs). For example, What Works Clearinghouse (WWC) has introduced SCD as an acceptable methodology beside group-based designs as evidence to evaluate best practices. A panel of experts drafted a pilot version of standards to use with single-case design studies. The draft was released in 2010 (Kratochwill et al., 2013). In 2015, a second panel of experts worked to develop criteria to determine the level of intervention effectiveness for studies that met WWC standards. Other groups of researchers have already included SCD in their reviews of EBPs as well. Lately, Wong and colleagues (2015) published a review of EBPs for children, youth and young adults with ASD. The authors noted the importance of including SCD in any evaluation of EBPs.

Single-Case Design vs Group Designs

Single-case design has several advantages over other designs when is used to evaluate interventions and provide evidence of effectiveness. SCD can be used to determine the utility of an intervention for a small number of individuals. With this purpose in mind, researchers have used SCD across different disciplines such as behavior analysis, psychology, special education, and medicine. Investigators implementing SCD assign different treatments to the same individual and measure the outcomes over time (Hedges, Pustejovsky & Shadish, 2012). Group design, on the other hand, uses summative evaluation of effect for a large number of participants (Lane & Gast, 2014). While a group study is logical when the researcher wants to learn about a particular population, most educational professionals work with one child or small group of children. A well-researched intervention may prove to be ineffective with a particular child even if delivered with integrity. It is critical that education professionals assess

whether an intervention is effective for each child after implementation (Riley-Tillman & Burns, 2009). The main factor contributing to this heterogeneity is the comorbidity between diagnoses such as ASD, ADHD, anxiety disorders, and intellectual disability (Lenroot & Yeung, 2013). This heterogeneity affects not only the symptoms of ASD but also the changes over time and the responses to interventions (Levitt & Campbell, 2009).

Rigor of the Studies

As with any other type of research design, SCD is applied with different levels of methodological rigor. Considering that the quality of a SCD study can vary, several authors have proposed guidelines to guide researchers when conducting studies or when synthesizing studies employing SCD. Critical appraisal guidelines to assess the rigor of research findings on healthcare and medicine existed a long time ago (Crombie, 1996). Knowing poorly conducted from well-conducted single-case design studies is important to better interpret the outcomes of research. It helps to increase the accuracy and transparency of the studies. One way to analyze the quality of studies is to implement quality appraisal tools. Several reporting standards and quality evaluation tools for SCD studies have been developed.

Task Force on evidence-based interventions for school psychology (2003) and the quality indicators for SCDs published by Horner and colleagues (2005) were the first tools. Raters made their judgments based on the following criteria: (a) description of participants and settings; (b) dependent variables; (c) independent variables; (d) baseline; (e) experimental control and internal validity; (f) external validity; and (g) social validity. Some of the tools were generated for a specific research domain (i.e. ASD, psychosocial interventions for individuals with ASD, or social skills training of children with ASD).

These tools are the Evaluative Method by Reichow, Volkmar, and Cicchetti, (2008); the Smith, Jelen, and Patterson Scale (2009); T. Smith et al. (2007); and Wang and Parrila (2008) scales. Other exclusive tools are used for a specific discipline such as augmentative and alternative communication. In this field, researchers used the Certainty Framework developed by Simeonsson and Bailey (1991) and the Evidence in Augmentative and Alternative Communication Scales (EVIDAAC; Schlosser et al., 2008). In 2008, Logan, Hickman, Harris, and Heriza developed an evaluation scale to be used in medicine and rehabilitation. The same year, Tate and colleagues created the Single-Case Experimental Design Scale to be implemented within the field of neurorehabilitation. In 2010, Kratochwill and colleagues designed the standards of what is known as What Works Clearinghouse to be used for educational research. Two federally sponsored organizations, the National Autism Center and the National Professional Development Center for Autism Spectrum Disorder, developed their own rubrics in 2008 and 2009 (i.e. Scientific Merit Rating Scale and The National Professional Development Center on Autism Spectrum Disorders rubric) to examine single-case literature. Maggin and Chafouleas (2010) developed the Protocol for Assessing Single-Subject Research Quality. Finally, Schlosser (2011) and Schlosser, Sigafoos, and Belfiore (2009) developed exclusive tools to evaluate one treatment vs. two or more treatments. A summary of standards and quality evaluation tools for SCD studies are presented in Table 1.

Table 1

Standards and Quality Evaluation Tools for SCD Studies

Tool	Author	Year	Specific discipline
Task Force on Evidence-Based Interventions for School Psychology		2003	General
Quality Indicators form single-subject research	Horner and colleagues	2005	General
Quality Indicators in Single-Case Research on Psychosocial Interventions for Individuals with ASD	Smith and colleagues	2007	Psychosocial interventions for individuals with ASD
Single Subject Research Quality Indicators	Reichow, Volkmar, and Cicchetti,	2008	For individuals with ASD
Quality indicators for Single-Case Research on Social Skill Interventions for Children with ASD	Wang and Parrila	2008	Social skills training for individuals with ASD
Evidence in Augmentative and Alternative Communication Scales	Schlosser and colleagues	2008	Augmentative and alternative communication
Evaluating Rigor and Quality of Single-Subject Research Designs	Logan and colleagues	2008	Medicine and rehabilitation
Single-Case Experimental Design Scale	Tate and colleagues	2008	Neuro-rehabilitation.
Scientific Merit Rating Scale	National Autism Center	2008	ASD
The National Professional Development Center on Autism Spectrum Disorders rubric	National Professional Development Center for Autism Spectrum Disorder	2009	ASD
What Works Clearinghouse	Kratochwill and colleague	2010	General

PASS-RQ	Maggin and Chafouleas	2010	General

Measuring SCD – Visual Analysis

Conducting a visual analysis is usually the first step when evaluating SCD in research practices (Ray, 2015). Visual analysis is a process to evaluate a participant's performance within and between conditions using systematic procedures. Visual analysis evaluates several factors present in a graph. These factors include: (a) level; (b) trend; (c) variability; (d) immediacy of effect; and (e) overlap (Lane & Gast, 2014). The person conducting visual analysis makes a judgment about the presence or absence of a functional relationship between the intervention and the outcome, rather than about the magnitude and consistency of this outcome (Valentine, Tanner-Smith, Pustejovsky & Lau, 2016). Despite its benefits, the process of visual analysis can be subjective and ambiguous. Historically, agreement between visual analysts has been low to moderate (Ninci, Vannest, Willson, & Zhang, 2015). However, research in this area has had important limitations. Ottenbacher (1993) conducted a quantitative review of the literature examining interrater agreement (IA) for visual analysis. The review reported a mean interrater agreement index of .58. There have been no other systematic reviews of visual analysts' agreement until Ninci and colleagues (2015) published their meta-analysis. The authors' purpose was to determine the overall agreement between raters in published literature and to examine potential moderators affecting the agreement. The study provided a mean IA of .76, which was an improvement relative to Ottenbacher's findings. Although Ninci and colleagues reported a higher IA index, it is difficult to

compare both reviews since they included different studies and used differences indices of agreement.

Several limitations of the two previously mentioned reviews make it difficult to draw conclusions about overall levels of agreement and generalize their outcomes. For example, it has been observed that analysts with lower level of expertise produced higher levels of agreement (Ninci et al., 2015). However, it is worth noting that analysts' expertise has been very limited in many studies. In other studies, it has been poorly described or not described at all (Wolfe, Seaman, & Drasgow, 2016). Furthermore, Ninci and colleagues noted that studies that operationally defined the characteristics to be analyzed visually reported higher rates of agreement. However, researchers studying visual analysis agreement rarely provide raters with operational definitions (Wolfe, Seaman, & Drasgow). Although visual analysis can provide important information about the effectiveness of interventions, certain characteristics of the data (e.g., presence of trend in the baseline, high variability of the data) make it difficult for visual analysts to make accurate judgements. Although criteria have been established to guide the process, visual analysis seems to allow some degree of subjectivity. Because of these limitations, it is often recommended to use other methods to measure the effect of an intervention, such as nonparametric techniques, together with visual analysis (Kazdin, 2011).

Measuring SCD – Standardized Effect Size Methods

The fact that SCD has historically relied on visual analysis of graphs to judge the effectiveness of an intervention has complicated the inclusion of SCD studies in evidence-based practices reviews (Valentine et al., 2016). Shadish, Hedges, Horner, & Odom, 2015, pointed to the importance of using good standardized effect size measures

using scales that have the same meaning across reviewed studies. These standardized effect size measures can benefit SCD research by making research studies more accessible and more useful when informing policy decisions. This past decade has seen a growing interest in the development of new methodologies for analyzing and synthesizing data from SCDs (Shadish, 2014; Smith, 2012).

Effect size is "a value that reflects the magnitude of the treatment effect" (Borenstein, Cooper, Hedges & Valentine, 2009. P.3). The American Psychological Association now highlights the importance of reporting effect sizes in research. Although statistical analysis is not often used to assess treatment effects in applied research that uses SCD, there has been recent evidence that statistical analysis can be applied to single-case studies (Hedges, Pustejovsky & Shadish, 2013). While some researchers note that effect sizes are practically a necessity for fully understanding the findings of a study and can compare studies (Shadish et al., 2015), other researchers consider reporting effect sizes best practice when presenting empirical research findings in many fields (Nakagawa & Cuthill, 2007). WWC standards suggest that SCD research must include visual analysis as well as quantitative measurement in order to be scientifically validated (Kratochwill et al., 2010). The author agrees with Parker and Hagan-Burke (2007) in regards to the importance of statistical analysis as a way to supplement visual analysis rather than substituting it.

Measuring SCD – Nonparametric Methods

The use of nonparametric methods is common in SCD research. These methods can be calculated by hand, do not require extensive training, and are easily interpretable (Vannest & Ninci, 2015). The most common type of methods that have been proposed for

use with SCDs are non-overlap indices, which rely on examination of the distributional overlap of the outcome data across phases (Parker, Vannest & Davis, 2011). In the last decade, several researchers have developed new non-overlap methods and published reviews analyzing the performance of these methods when compared to each other and to visual analysis. Several researchers have documented the superior performance of some of the methods.

The concept of overlap refers to data falling above or below an imaginary line that separates baseline and interventions phases (Vannest & Ninci, 2015). Using non-overlap methods has some advantages. They blend well with visual analysis of graphed data, are easy to use, and they do not require to make assumptions about the distribution of data. Although some of the methods might combine non-overlap with median calculation, non-overlap methods do not rely on means, medians or modes but rather rely on the value of individual data points (Parker et al., 2011). Data in applied settings are often variable, containing extreme scores and baseline trend (Vannest & Ninci).

Despite their advantages, most non-overlap methods share some weaknesses. When Parker, Vannest, and Davis compared nine non-overlap methods, they concluded that most were insensitive to outcomes at the top end of the distribution. This is a serious disadvantage when one is trying to compare two successful interventions. Furthermore, most methods are insensitive to positive baseline trend.

One of the oldest methods is the extended celeration line (ECL) or "split middle" line (White & Haring, 1980). ECL calculates the proportion of Phase B data above a "split middle" median line. Although ECL is one of the few methods that controls for

trend, it assumes that baseline data is linear. ECL has been published as PEM-T (Wolery, Busick, Reichow & Barton, 2010).

Percentage of non-overlapping data (PND) was created by Scruggs, Mastropieri and Casto (1987) for synthesizing single-case research. PND is very popular and can easily be calculated in the great majority of cases. However, several limitations have been identified by Parker, Hagan-Burke, and Vannest (2007). For example, PND ignores all baseline data except for one data point, which because of its extremity, is likely the most unreliable. Because it has a ceiling-and-floor effect, PND cannot discriminate magnitudes at the higher and lower ranges. Consequently, PND is an accurate measure of overlap only when data is stable and does not include trend or outliers.

In 2006, Ma develop the percentage of data exceeding the median (PEM). PEM identifies the overlap based on the median score of the baseline rather than a single data point like PND does. In other words, PEM is the percentage of data points in the intervention that exceeds the median of the baseline. However, if there is large variability or trend in the baseline, the median might not be a good reflection of the baseline (Vannest & Ninci, 2015). A critical advantage of PEM over PND is the fact that PEM reflects an effect size in the presence of floor or ceiling baseline data points. PEM has also some important limitations: It is insensitive to magnitude of data points above the median and it does not consider variability and trend.

Percentage of All non-overlapping data (PAND) was the variation on PND proposed by Parker et al., (2007). The change is the identification of the total number of data points that do not overlap between baseline and intervention phases. PAND uses all data from both phases, avoiding the PND focus on one unreliable data point (Parker et

al.). The authors also suggest that the index can be converted into a *Phi* effect size index. Pearson's Phi can be calculated and translated into Cohen's *d* (Schneider, Goldstein, & Parker, 2008). Like PND, PAND is insensitive at the upper end of the scale. When there is no data overlap between phases, both PND and PAND give a 100% score, regardless of the distance between the two data clusters. Neither PND nor PAND control positive baseline trend.

Percentage of zero data (PZD) is the degree to which behavior is eliminated in treatment (Harvey, Boer, Meyer, & Evans, 2009) so it is used when the treatment's goal is to eliminate negative behaviors rather than reduce them (Parker et al., 2011). It is easy to calculate but can be distorted if treatment is terminated immediately after zero data point occurs (Harvey et al., 2009). Just like PND, PZD statistics are overly sensitive to outliers and trend (Allison & Gorman, 1993).

Pairwise data overlap (PDO; Parker & Vannest, 2007) calculates the overlap of all possible paired data comparisons between baseline and intervention phases. PDO has advantages and limitations. PDO produces more reliable results than other non-parametric indices, and relates closely to established effect sizes (e.g., Pearson r, Kruskal-Wallis W). However, it takes slightly longer to calculate since it requires that individual data point results be written and added, making calculation laborious for long and crowded data series (Wendt, 2009).

To address the lack of control for baseline trend in most of the previous methods, Manolov and Solanas (2009) developed the percentage of non-overlapping corrected data (PNCD) as a modification of PND. A data-correction procedure is to be implemented prior to applying the PND in order to eliminate from the data a possible pre-existing trend

that was not related to the introduction of the intervention. Unstable baselines have been regarded as undesirable, but they can be common in applied settings in which the introduction of the treatment is subjected to factors that cannot always be controlled by the practitioners (Manolov & Solanas).

Non-overlap of all pairs (NAP; Parker & Vannest, 2009) was developed mainly to improve upon existing SCD overlap-based methods: PND, PAND, and PEM. NAP is interpreted as the percentage of all pairwise comparisons across baseline and treatment phase, which show improvement across phases or, more simply, the percentage of data, which improve across phases. The concept of score overlap is identical to that used by visual analysts of SCD graphs and is the same as is calculated in the other overlap indices, PAND, PEM and PND (Parker & Vannest). NAP is a "complete" non-overlap index as it individually compares all baseline and treatment phase data points. Using more data in the analysis increases the sensitivity of the measure (Vannest & Ninci, 2015).

Improve Rate Difference (IRD; Parker, Vannest, & Brown, 2009) is based on the risk-reduction technique used in medical research. IRD is defined as the difference in improvement rates between Phases A and B. Just like PAND, it begins by identifying the minimum number of data points needing removal to eliminate all data overlap between the phases. This method tends to be more robust than PND and PEM because it uses more data in the calculation (Vannest & Ninci, 2015).

Percent exceeding the median trend line (PEM-T; Wolery et al., 2010) is an improved version of PEM that considers the trend in the baseline data. PEM-T is calculated by using first the split-middle technique (White & Haring, 1980), a common

approach to determine trend in SCD data. After drawing a trend line in the baseline phase using split-middle technique, this line is extended to the treatment phase. The data points in the treatment phase above the trend line are counted and the percentage of non-overlap is calculated.

Tau-U (Parker, Vannest, Davis, & Sauber, 2011) is a non-overlap method that was developed to address the problems of the previous methods. Just like NAP, Tau-U uses all pairwise comparisons across baseline and treatment phases. Tau-U also adjusts for positive baseline trend, can handle small data sets, and discriminates magnitudes at the upper and lower limits (Vannest & Ninci, 2015).

Published literature regarding the benefits and limitations of the different non-overlap methods was reviewed by the investigator in order to decide which methods to include them in this review. Parker and Hagan-Burke (2007) compared PND, PEM and IRD and their agreement with visual analysis. The authors concluded that IRD was the metric with the highest agreement. IRD allows the calculation of confidence intervals and is not limited to data that meets certain assumptions. Also, IRD has been extensively used in published meta-analysis of single-case research, making it the most validated non-overlap method (Mason, Ganz, Parker, Burke & Camargo, 2012). Parker, Vannest, Davis and Sauber (2011) described three critical limitations of nonverlap measures: (a) lack of statistical power, (b) low ability to discriminate the magnitude of successful interventions, and (c) inability to account for trend. The authors presented Tau-U as a method that addresses these limitations. Tau-U is a complete measure that includes both trend and level. Because it includes level and trend, Tau-U is not likely to hit a ceiling like other non-overlap methods. This characteristic gives Tau-U higher discriminating

power with the top results. In 2011, Parker Vannest and Davis compared nine nonverlap methods. The authors concluded that Tau-U had several advantages over other nonverlap methods. Tau-U was the only method able to control Phase A monotonic trend, which is the tendency for scores to increase over time. Tau-U was one of the measures with the greatest statistical power, which is required to produce results with high precision. Finally, the authors noted that all methods shared some degree of insensitivity to results at the top end. However, Tau-U proved to be the most sensitive.

Brossart, Vannest, Davis & Patience (2014) provided suggestions for integrating visual and statistical analysis in single-case research. After reviewing the strengths and weaknesses of all non-overlap methods available, they selected IRD and Tau-U for their paper. They concluded that Tau-U was the best available method for analyzing single-case data. Both IRD and Tau-U are easily understood and interpreted by interventionists. Rakap, Snyder and Pasia (2014) conducted a review comparing 12 non-overlap methods. Based on the findings, the authors recommended the use of both IRD and Tau-U to supplement visual analysis judgements. Both measures show good levels of discriminability, are compatible with visual analysis, allow for classification of intervention magnitude, and do not require data assumptions. According to Rakap (2015) using Tau-U has two important advantages. First, it considers both level change across phases and baseline trend, which is congruent with single-case research logic. Second, it is a non-parametric technique compatible with visual analysis. In their review of non-overlap methods, Chen, Hyppa-Martin, Reichle & Symons (2016) concluded that some methods are more sensitive and more consistent with visual analysis than others. According to the authors, improvement rate difference (IRD), when compared to percent

of data points exceeding the median (PEM), percent of all non-overlapping data (PAND), Phi, non-overlap of all pairs (NAP) and Tau-novlap overlap indices, was the most consistent with visual analysis judgement and one of the most effective in differentiating the magnitude of intervention effect.

Tarlow (2016) identified some limitations in Tau-U and proposed an improved method of correcting baseline trend and measuring effect size, Baseline Corrected Tau. Tarlow's study demonstrated that both methods show to be robust to autocorrelation and correlate highly. However, the study also showed that Tau-U can yield out-of-bounds results, with scores less than -1 or greater than +1. Baseline Corrected Tau did not show this weakness. Finally, Baseline Corrected Tau showed to control baseline trend more effectively than Tau-U. Since Tarlow introduced and field-tested this new measure, no studies have been published replicating this testing and Tarlow's outcomes. A summary of the major effect size metrics for single case design are presented in Table 2.

Table 2

Major Effect Size Metrics for Single Case Designs

Metric	Author	Year	Discerption	Control trend
Extended Celeration Line (ECL)	White and Haring	1980	ECL calculates the proportion of Phase B data above a "split middle" median line	Yes
Percentage of non-overlapping data (PND)	Scruggs and colleagues	1987	PND is an accurate measure of overlap only when data is stable and does not include trend or outliers	No

Percentage of Data Exceeding the Median (PEM)	Ma	2006	PEM is the percentage of data points in the intervention that exceeds the median of the baseline	No
Percentage of All non-overlapping data (PAND)	Parker and colleagues	2007	PAND is the variation on PND except the identification of the total number of data points that do not overlap between baseline and intervention phases	No

Table 2
(continued).

Metric	Author	Year	Discerption	Control trend
Pairwise data overlap (PDO)	Parker and Vannest,		PDO calculates the overlap of all possible paired data comparisons between baseline and intervention phases	No
Percentage of zero data (PZD)	Harvey and colleagues	2009	PZD is the degree to which behavior is eliminated in treatment and used when the treatment's goal is to eliminate negative behaviors rather than reduce them	No
Percentage of non-overlapping corrected data (PNCD)	Manolov and Solanas	2009	PNCD is the same as PND except that a data-correction procedure is implemented prior to applying the PND to eliminate from the data a possible pre-existing trend	Yes
Non-overlap of all pairs (NAP)	Parker & Vannest,	2009	NAP is interpreted as the percentage of all pairwise comparisons across baseline and treatment	No

Improve Rate Difference (IRD)	Parker and colleagues	2009	IRD is based on the risk-reduction technique used in medical research. IRD is defined as the difference in improvement rates between Phases A and B.	No

Table 2 (continued).

Metric	Author	Year	Discerption	Control trend
Percent exceeding the median trend line (PEM-T)	Wolery and colleagues	2010	PEM-T is an improved version of PEM that considers the trend in the baseline data. PEM-T is calculated by using first the split-middle technique in bassline phase and then extend the line to the treatment phase	Yes
Tau-U	Parker and colleagues	2011	Tau-U is adjusts for positive baseline trend, can handle small data sets, and discriminates magnitudes at the upper and lower limit	Yes
Baseline Corrected Tau	Tarlow	2016	Baseline Corrected Tau showed to control baseline trend more effectively than Tau-U	Yes

This study will contribute to previous literature by using visual analysis as well as non-parametric measures of effect with a larger sample (i.e., 924 AB phase contrasts, 132 studies, and 369 students). The purpose of this dissertation is to answer the following four a priori research questions:

1. To what extent are IRD, Tau-U, and Baseline Corrected Tau sensitive and specific in detecting intervention effects in SCD studies?

2. To what extent do IRD, Tau-U and Corrected-Baseline Tau outcomes agree with interpretations based on visual analysis?
3. To what degree does autocorrelation, number of data points, presence of undesired trend in baseline or intervention, and degree of overlap affect interpretation in visual analysis and statistical analysis?
4. What is the typical range of non-overlap estimates from interventions deemed effective by visual analysts?

CHAPTER II

METHODS

Search Procedures

The author conducted electronic searches to identify studies targeting school-based interventions for students with ASD. The following search procedures generated a total of 9,140 studies. Inclusion and exclusion criteria were applied to this sample.

Three databases were searched including: Education Resources Information Center (ERIC), PsycINFO, and MEDLINE to find studies targeting academic, social communication, play, and functional life skills. The author defined academic skills as those skills needed to succeed in educational settings, such as post-secondary education, and to live an independent functional life. Social communication skills were defined as the use of verbal and non-verbal communication for social purposes, ability to change communication matching it to context or following the needs of listeners and following rules for conversation and storytelling. Play skills were defined as participation, engagement and appropriate behavior during a group or solitary play activity. Functional life skills were defined as the abilities that assist in living independently such as self-care, household, pre-vocational, community safety, self-determination, or health/hygiene skills.

The searches were completed using the following key words: "autis*," Asperger, or pervasive developmental disorder (pervasive developmental disorder-not otherwise specified; PDD-NOS) and "intervention," in combination with "academic," "general education," "literacy," "math*," "writing," "science," "organization," or "task engagement," "communication," "vocalization," "sign language," "speech," "augmentative and alternative communication," "AAC," "social," "conversation,"

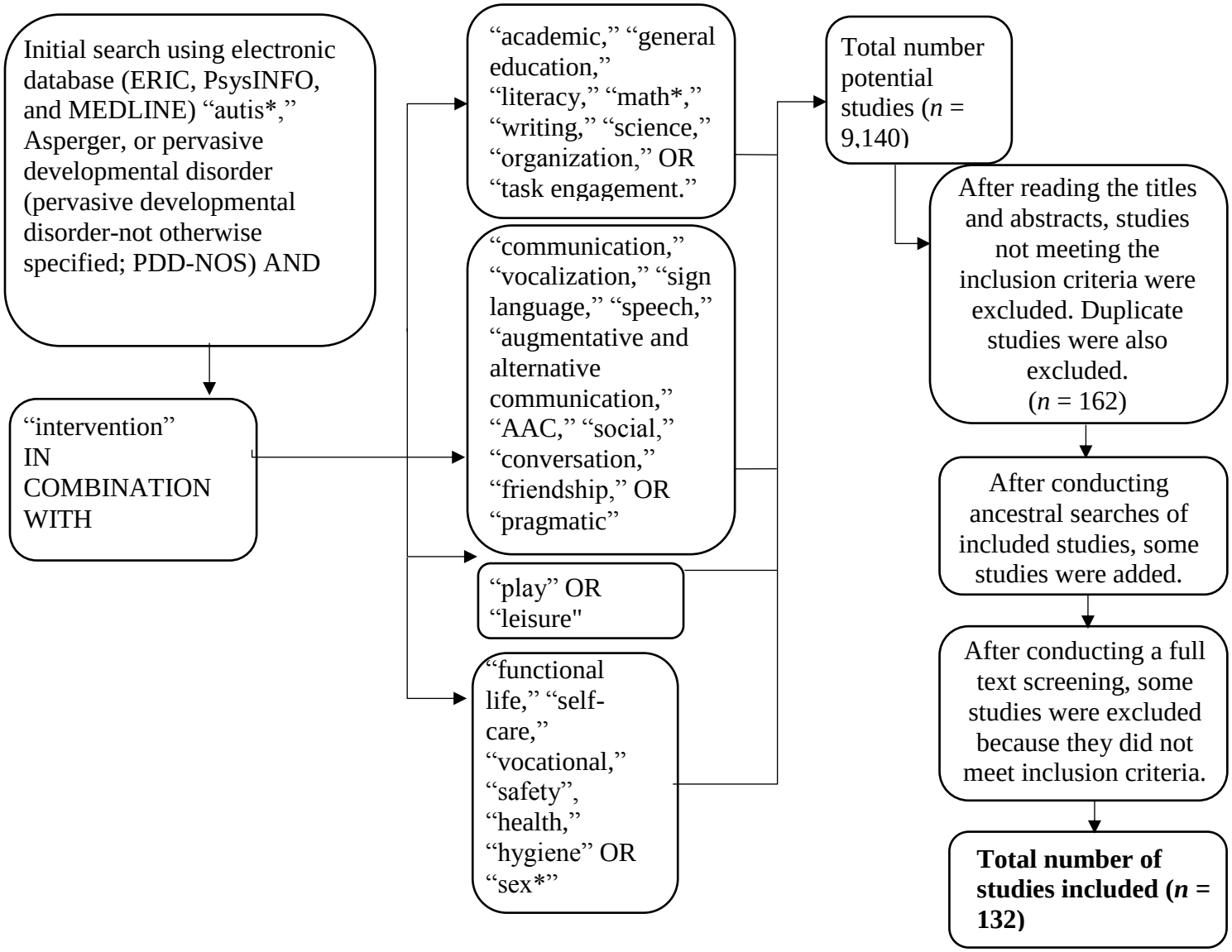

Figure 1. Search procedures for the literature review on academic, social communication, play and functional life skills interventions targeting students with ASD in school settings; *n* = number of

"friendship," "pragmatic" "play," "leisure," "functional life," "self-care," "vocational," "safety," "health," "hygiene," or "sex*." Search procedure is presented in figure 1.

The author limited the search results to English language studies published in a peer-reviewed journal between 1995 and 2014. This search initially produced 9,140 articles to review for potential inclusion. The author identified 132 studies for inclusion after eliminating duplicate articles, applying inclusion criteria and conducting ancestral searches.

Inclusion and Exclusion Criteria

The author read the method section of each study to find articles that met inclusion criteria. To be included, a study was required to (a) implement intervention to improve academic skills (i.e., literacy, mathematics, science, social studies) or related skills (i.e., prerequisite skills and engagement and task completion); social communication (i.e., conversation, nonverbal communication, vocalization, sign language, and Augmentative and Alternative Communication); play (i.e., solitary play and group play); or functional life skills (i.e., self-care, safety, health, sex, hygiene, self-determination, and vocational); (b) target at least one participant with a medical diagnosis or educational classification of ASD between the ages of three and 21 years; (c) use a single-case research design (i.e., multiple-baseline design, multiple-probe design, reversal or withdrawal design, or changing criterion design). Single-case designs are especially appropriate for special education research because they emphasize the performance of individual students. Group designs are more concerned with the effects of interventions on groups of individuals (Horner et al., 2005); and (d) conduct intervention in a private or public-school setting (i.e., preschool, elementary school, middle school, junior high or

high school, transition age program serving students aged 19 to 21 years of age). Interventions were considered school based if all sessions took place in a special education or general education classroom, playground, cafeteria, gymnasium, library, or any other area within the school campus.

Exclusion criteria were (a) group design studies; (b) designs without a return to baseline following an intervention condition (e.g., ABC [i.e., Bagatell et al., 2010; Kinnealey et al., 2012], AB [i.e., Cicero & Pfadt, 2002); (c) one baseline used to evaluate effectiveness of two or more different interventions (i.e., ABAC/ACAB [i.e., Reisener, Lancaster, McMullin & Ho, 2014], and ABCD [i.e., Coleman-Martin, Heller, Cihak & Irvine, 2005); (d) alternating treatment designs (i.e., Angermeier, Schlosser, Luiselli, Harrington & Carter, 2008; Armstrong & Hughes, 2012; Cihak & Foust, 2008; Cihak, Kildare, Smith, McMahon & Quinn-Brown, 2012; Cihak & Schrader, 2008; Cihak, Smith, Cornett & Coleman, 2012; Fletcher, Boon & Cihak, 2010; Ganz, Boles, Goodwyn & Flores, 2014; Hetzroni & Ne'eman, 2013; Johnston, Buchanan & Davenport, 2009; Kurt & Tekin-Iftar, 2008; Lorah et al., 2013; Polychronis, McDonnell, Johnson, Riesen & Jameson, 2004; Riesen, McDonnell, Johnson, Polychronis & Jameson, 2003; Tincani, 2004; Wilson, 2013); (e) combination of designs (i.e., Cihak, Wright & Ayres, 2010; Couper et al., 2014; Fentress & Lerman, 2012; Marcus & Wilder, 2009; Mucchetti, 2013; Paterson & Arco, 2007; Shillingsburg, Powell & Bowen, 2013); (f) studies with fewer than three data points in baseline phases (i.e., Reagon, Higbee & Endicott, 2006; Sancho, Sidener, Reeve & Sidener, 2010; Van der Meer et al., 2013); and (g) interventions implemented in university clinics, in family homes, or in other community settings (i.e., Delano, 2007; Gunby, Carr & Leblanc, 2010; Palmen, Didden & Arts, 2008; Taylor,

Hughes, Richard, Hoch & Coello, 2004; Tekin-Iftar & Birkan, 2010). A summary of the inclusion and exclusion criteria are presented in Table 3.

Table 3

Inclusion and Exclusion Criteria for Selection of the Articles

Inclusion criteria
1. Implement intervention to improve academic or related skills, social communication, play or functional life skills.
2. Target at least one participant with a medical diagnosis or educational classification of ASD between the ages of three and 21 years.
3. Use a single-case design (i.e., multiple-baseline design, multiple-probe design, reversal design, or changing criterion design).
4. Conduct intervention in a school setting (i.e., preschool, elementary school, middle school, junior high or high school, transition age program serving students aged 19 to 21 years of age).
Exclusion criteria
1. Group design studies.
2. Designs without a return to baseline following an intervention condition (e.g., ABC and AB).
3. One baseline used to evaluate effectiveness of two or more different interventions (e.g., ABAC, ACAB and ABCD).
4. Alternating treatment designs.
5. Combination of designs.
6. Studies with fewer than three data points in baseline or treatment phases.
7. Interventions implemented in university clinics, in family homes, or in other community settings.

Data Extraction

After the implementation of inclusion and exclusion criteria, a total of 924 AB phase contrasts, 132 studies, and 369 students were recognized. These were drawn from 12 studies utilizing withdrawal designs (i.e., ABAB or ABCAC), 70 studies using multiple-baseline designs (i.e., across participants [*n* =51], across behaviors [*n* =7], across tasks [*n* =6], across settings [*n* =3], and across skills [*n* =3]) and 50 studies with multiple-probe design (i.e., across participants [*n* =22], across behaviors [*n* =11], across tasks [*n* =7], across responses [*n* =4], across routines [*n* =2], across behaviors and settings [*n* =1], across participants and settings [*n* =1], across stimuli [*n* =1], and across skills [*n* =1]). In studies using ABAB withdrawal designs, AB contrasts were identified by pairing each A phase with the consecutive B phase. One of the withdrawal studies used an ABCAC design. In this case, each A phase was paired with the following C phase. The B phase was not used. The median number of data points in Phase A (i.e., the baseline condition) was 5 (IQR: 9-4), and in Phase B (i.e., the intervention condition) was 11 (IQR: 18 -7).

Variable Coding

A word document was created for each of the 132 studies and each study was coded for the following variables: (a) type of design (i.e., withdrawal designs, multiple-baseline designs, multiple-prop designs, and changing-criterion design); (b) number of participants with ASD (students with different types of disabilities other than ASD were excluded); (c) participant demographics including age, gender, race and ethnicity; (d) diagnoses and severity (i.e., high functioning, moderate or severe); (e) type of school setting (i.e., general or special education classroom, public or private school, playground,

cafeteria, gymnasium, library, or any other area within the school campus); (f) targeted skills according to the target skills (i.e., academic skills, social communication skills, play skills, or functional life skills); (g) intervention (h) interventionist (e.g., teacher, paraprofessional, research assistant, peer); (i) inter-rater reliability procedures and scores; (j) treatment integrity procedures and scores; (k) social validity assessment and outcomes; and (l) assessment of generalization and maintenance.

Rigor of the Studies

The author used Reichow (2011) guidelines to evaluate the rigor of the studies. Like the current study, Reichow's method targets young children and children with ASD. Therefore, Evaluative Method was the closest to the objective of this review among all instruments. An Excel sheet was developed for each of the 132 included studies. The Excel sheet had a total of 13 columns. Six columns were reserved for the primary indicators (i.e., participant characteristics, independent variable, dependent variable, baseline condition, visual analysis, and experimental control). Each primary indicator was evaluated as high, acceptable, or unacceptable. The following six columns were intended for the secondary quality indicators (i.e., inter-observer agreement, kappa, blind raters, fidelity, generalization or maintenance, and social validity). Each secondary indicator was evaluated using a dichotomous scoring system. A "yes" score meant that the quality indicators were present. A "no" score was used if the indicators were not present. The last column was intended for the overall judgement, either a strong, acceptable or weak study. A study is determined to be *strong* if (a) each primary indicator is rated as high, and (b) at least three of the six secondary indicators are present. A study is considered *adequate* if (a) it has a high score on at least four primary indicators

(without any unacceptable scores in any other primary indicator), and (b) at least two of the six secondary indicators are present. A study is considered *weak* if (a) it receives a high score in fewer than four primary indicators, or (b) if fewer than two of the six secondary indicators are present.

Preparation of Data

Using a MacBook computer, the author created a folder for each study. Then, a screenshot was taken of each graph and saved it as a png image in the appropriate folder. A total of 924 AB contrasts were numbered and saved in the folders. The author extracted data from the graphs by using the UN-SCAN-IT version 5.2 (Silk, 1992) to manually digitize underlying *x, y* data points. Then, the data was saved in an Excel database sheet. To extract contrasts from the graphs using withdrawal, multiple-baseline, or multiple-probe designs, the author identified each adjacent AB pair (baseline and following intervention phase). Each pair was treated separately. Similarly, when a combination of withdrawal and multiple-baseline or multiple-probe designs was present, each adjacent AB pair was extracted.

Effect Size Calculation

Tau-U

Tau-U scores were calculated for each contrast. Parker et al. (2011) introduced Tau-U as an alternative to regression-based and existing non-overlap models of statistical analysis of single-case research data. Tau-U combines assessment of non-overlap between A-B experimental phases while controlling for positive baseline trend. Parker and colleagues published different variations of Tau-U in their original paper. However, in 2011, they presented a revised version of the metric. In this revision, Tau-U seems to

refer to the variation that compares A and B phases and adjusts for phase A trend. Tau-U is now defined as:

$$\text{Tau-U} = \text{Tau} - t_{A.}$$

To calculate Tau-U, the author pasted the data saved in the Excel sheet into the Tau-U calculator lab at the Single Case Research website (www.singlecaseresearch.org). Baselines were corrected, and both baselines and comparison phases were weighted to obtain final Tau-U scores. Tau-U scores range from 0% to 100% and can be interpreted using the following criteria: (a) 65% or lower suggests a weak effect; (b) between 66% and 92% suggests a medium to high effect; and (c) 93% to 100% suggests a strong effect (Parker & Vannest, 2009).

Baseline Corrected Tau

Tarlow's method was selected as the second effect size metric because the author wanted to replicate Tarlow's outcomes. Baseline Corrected Tau process to calculate effect size involves three different steps. First, monotonic baseline trend is identified and its statistical significance analyzed with Kendall's Tau rank correlation coefficient. Next, baseline trend (if there is one) is corrected across A and B phases using Theil-Sen estimator. Finally, effect is calculated using Kendall's (1962) method. A dummy code phase variable is correlated with either the original or the corrected data from the previous steps.

To calculate Baseline Corrected Tau, the author first determined whether each AB contrast presented a statistically significant monotonic baseline trend conducting a Tau analysis for the A phases only. When baseline trend was present, the author removed the trend by performing a Theil-Sen regression of a time (x) variable onto the baseline

observations (*y*). Finally, Baseline Corrected Tau analysis was conducted to calculate effect size and yield a Tau value. These calculations were made using the Web-Based Calculator developed by the author (Tarlow, 2016) and available at http://www.ktarlow.com/stats/tau.

IRD

Improvement rate difference (IRD) was selected as a third effect size metric. The author's decision was based on the outcomes of Chen and colleagues (2016) review of effect size metrics. The authors concluded that IRD was one of the best effect size metrics in differentiating the magnitude of intervention effect as well as the most consistent with visual analysis judgement. To calculate IRD, the author first calculated two improvement rates (IRs). The IR for each phase is defined as the number of "improved data points" divided by the total data points in that phase:

$$IR = \frac{\text{\# improved data points}}{\text{\# total data points}}$$

Improved data points in the baseline and the treatment phase have different definitions. An improved data point in baseline is any data point that ties or exceeds all data points in the treatment phase. An improved data point in the treatment phase is any data point that exceeds all data points in the baseline phase. "Exceeds" refers to higher levels of behaviors we wish to increase (e.g., homework completion) and to lower levels of behaviors we wish to decrease (e.g., tantrums). Improved data points are identified visually (Parker et al., 2009).

The author then calculated the difference of the improvement rate of the treatment phase minus the improvement rate of the baseline phase: IRD = IRT - IRB. IRD scores

range from 0 to 100% or 1.00. A negative IRD score indicates deterioration below baseline levels (Parker et al., 2009). From comparing visual ratings with IRD, Parker et al. (2009) estimated tentative benchmarks. Very small and questionable effects scored about .50 and below. Moderate-size effects had IRD scores of around .50 to .70. Effects rated as large and very large generally received IRD scores of .70 or .75 and higher.

Visual Analysis

Visual analysis was conducted to determine whether there was an effect present in each phase contrast and study. The author used visual analysis procedures based on the procedures used by Petersen-Brown et al. (2012). Ten raters independently rated the 924 AB phase contrasts and the 132 studies. The raters were doctoral and master students who had completed graduate courses on SCD methodology and had experience in research employing SCD. See Table 4 for a summary of raters' demographic information. Raters were asked to use the singlecase.org website to complete the Visual Analysis Training Program before being asked to judge the graphs in the current study. The purpose of the training program was to assess their reliability judging graphs using visual analysis techniques. Raters were asked to judge multiple-baseline (MBD) and ABAB designs and have a minimum correlation average score of 90% compared to the experts. After completing the training program, the author conducted a 20-minute workshop for each individual rater to introduce the two rubrics (i.e., contrast rubric and study rubric) that they would use in the visual analysis of the data and modeled the procedure to judge the contrasts and the studies. Raters assessed integrity of the visual analysis instruction delivered by the author. Each step was scripted with space provided for the participants to write *yes* or *no* for each material and procedure as it was covered and taught. The

integrity checklist contained five materials and 11 procedures, a total of 16 variables. The participants recorded *yes* if the procedure was followed by the author or the materials were used. They used *no* for steps not followed or materials not used. The author counted the number of "yes" and "no" answers for each participant and calculated the percentage. Then, an average across participants was calculated. The overall integrity score was 92%. The most common score was 100% and the median was 100%. The range of scores was from 70% to 100%.

Table 4

Summary of Raters' Demographic Information

Participant	School department	Position	Terminal degree, certification	Correlation scores in online training
Rater 1	School Psychology	PhD student	M.A. BCBA	91%
Rater 2	School Psychology	PhD student	M.A. BCBA	96%
Rater 3	Special education	Faculty member	Ph.D.	90%
Rater 4	Special education	PhD student	M.S.	91%
Rater 5	Special education	PhD student	M.A. BCBA	90%
Rater 6	Special education	PhD student	M.S. BCBA	99%
Rater 7	Special education	PhD student	M.A. BCBA	90%
Rater 8	School Psychology	Master student	B.A.	90%
Rater 9	Special education	Master student	B.S. RBT	90%
Rater 10	School Psychology	Master student	B.S.	93%

Inter-rater 1	Special education	PhD student	M.S. RBT	92%
Inter-rater 2	School Psychology	Master student	B.A.	90%
Inter-rater 3	School Psychology	PhD student	M.S. BCBA	92%

Note. B.A.= Bachelor of arts; B.S.= Bachelor of science; M.A.= Master of arts; M.S.= Master of science; Ph.D.= Doctor of philosophy; BCBA= Board certified behavior analyst; and RBT= Registered behavior technician.

These raters evaluated changes in level, trend, variability, immediacy, and overlap between phases A (baseline) and B (treatment). A change of level was determined by looking at the change of mean between phases A and B. A change of trend was defined as a change in slope showing a systematic increase or decrease of performance over time. Variability was defined as the degree to which performance fluctuates around the mean or the slope of phase A versus phase B. Immediacy of effect referred to whether there was a difference between the last three data points of the A phase and the first three data points of the B phase. Although the majority of the studies involved acquiring skills, a few studies involved decreasing certain negative behaviors. In all features, the change had to be positive if the intended outcome was to increase behavior and the change had to be negative if the intended outcome was to decrease behavior. Overlap was defined as to whether there was any overlap between data points in phase A and data points in phase B.

The author took a screen shot of each contrast. The screen shots were resized so they were each the same dimension (i.e., 3.5 X 3.5 inches) and were saved as a portable network graphics (PNG) image. A Word document was created for each study. Then, all contrast graphs from the same study were copied together in the Word document

associated with that study with an accompanying contrast rubric for each graph. Each graph was given an individual number for identification. At the end of the Word document, all of the graphs were pasted again together with an overall study rubric. The contrast rubrics included level, trend, variability, immediacy, and overlap. Each feature was to be scored using two categories (i.e., change, no change) except for overlap where raters scored using "yes" or "no". For the study rubrics, the author used a 1 to 7-point scale, where 1 indicates there is no functional relationship between the treatment and the outcome; 4 represents moderate functional relationship; and 7 indicates a strong relationship. The author used the rubric published by the singlecase.org website. The rubrics included narrative anchors to assist raters in selecting the right scores (i.e., weak, moderate, and strong relationship).

The 924 contrast graphs were distributed as equally as possible among the raters. Each rater received a mean of 92.4 graphs and 13.2 studies. The median for the graphs was 92.5 and for the studies was 13.5. The graph mode was 86 and the study mode was 13. Finally, the range of graphs was 86 to 99. The range of studies was 6 to 19. Each rater was provided with a printed booklet containing all of the phase contrasts and studies they were assigned to analyze. They were asked to evaluate each contrast and complete the rubrics indicating how many features showed a change. If raters detected change in at least three of the five features in one contrast, an intervention effect was coded for that specific contrast by the author. After the raters rated each contrast in each individual study, they made a holistic judgement of the intervention effect for the whole study. This judgement was recorded in the 7-point scale rubric provided. Based on the benchmarks

provided by the singlecase.org website, a score of 1 or 2 indicated weak effect. A score of 3, 4, or 5 indicated medium effect. A score of 6 or 7 indicated strong effect.

Statistical Analysis

Following the procedures used in Petersen-Brown et al. (2012) as well as Chen et al. (2016), receiver operating characteristic (ROC) analysis, kappa coefficient, and Pearson's correlation coefficient were used to answer the first two questions in the current study.

Sensitivity and Specificity

Receiver Operating Characteristic (ROC) was originally used in signal analysis theory (Swets, 1988). According to the theory, the two most important accuracy measurements of a predictor are the true-positive proportion and the false-positive proportion. In other words, "hits" versus "false alarms". A metric or predictor will always require a trade-off between the two measurements unless the metric or predictor is perfect. When casting a net to catch as many true positives as possible, one will mistakenly catch some negatives (i.e., type I error). See Figure 2 for a confusion matrix showing the four possible proportions in ROC analysis.

Confusion Matrix		**Visual Analysis**		
		Effect	**No effect**	
Non-overlap metric	**Effect**	a True- positive (TP) Correct	b False-positive (FP) Type I Error	*a+b*
	No effect	c False-negative (FN) Type II Error	d True- negative (TN) Correct	*c+d*
		a+c	*b+d*	*a+b+c+d=N*
		Precision = a/(a+b) *Positive Predictive Value*		

	True-Positive Proportion = a/(a+c) *Sensitivity*
	True-Negative Proportion = d/(b+d) *Specificity*
	False-Positive Proportion = b/(b+d) *1-Specificity*

Figure 2. Confusion matrix for relative operating characteristic (ROC) analysis.

ROC was later extended to be used in the fields of medicine and psychology to analyze the accuracy and performance of diagnostic tools. Specifically, ROC analysis measures the sensitivity and specificity of a metric. Sensitivity refers to the metric's ability to detect true positives and specificity refers to the metric's ability to detect true negatives. ROC analysis produces a ROC curve and the area under the curve (AUC) shows the accuracy of the metric. A metric's accuracy is determined by the relationship between sensitivity and specificity. When sensitivity increases, specificity decreases, and vice versa. An AUC above 0.80 has been demonstrated in past research (e.g., Muller et al. 2005) to indicate a reasonable metric.

The author used the SPSS software package to obtain ROC analysis outcomes. First, the author inserted all the scores for each of four variables (i.e., Tau-U, IRD, Baseline Corrected Tau, visual analysis) into SPSS Data Editor. The values for the visual analysis variable were adjusted to two possible values (0 for "no effect" and 1 for "effect"). The measurements for the three non-overlap methods were adjusted as "scale", while "nominal" was used for visual analysis. Second, ROC curve was selected from the Analyze drop-down menu. Third, visual analysis was entered as "state variable" with a value of 1. The three non-overlap methods were entered as "test variable". SPSS provided the ROC curve and the table with the AUC scores. With this process, the author was able to create a ROC curve that showed the ability of the non-overlap metrics to rank the positive cases versus the false positive cases (1-specificity) based on a gold standard

(i.e., visual analysis). In this study, true positives are the phase contrasts correctly classified by the metrics as showing effect. A large number of true positives indicate a high level of sensitivity. False positives are the phase contrasts incorrectly classified by the metrics as showing effect. A large number of false positives indicate a low level of specificity.

To calculate sensitivity and specificity, the outcomes of the three non-overlap methods must be compared with visual analysis. Consequently, they must have the same type of outcomes, in this case dichotomous outcomes (effect vs no effect). To accomplish this, a cutoff score must be selected for each method. The author used ROC analysis to find the best cutoff score for each non-overlap method. Three different cutoff scores were selected for each method and sensitivity and specificity were calculated for each of the three cutoff scores (i.e., Tau-U = 0.61; IRD = 0.65; Baseline Corrected Tau = 0.41). The equations are the following:

Prevalence of Effect = T_{effect}/ Total × 100

Specificity is the fraction of those with no effect which will have a negative test result: Specificity: D/(D+B) × 100. *Specificity=true negatives/(true negative + false positives).*

The previous procedures will also allow the calculation of type I (false positives) and type II (false negatives) errors made by the three non-overlap metrics included in the review.

Kappa Coefficients

As seen in the previous section, ROC analysis can evaluate the sensitivity and specificity of a metric. Sensitivity and specificity are calculated for each possible score of

the metric. Consequently, ROC analysis can also help select a specific value or score to use as a threshold that provides the best trade-off between sensitivity and specificity. By selecting a threshold (Tau-U = 0.61; IRD = 0.65; Baseline Corrected Tau = 0.41) cutoff score, the author was able to transform the metrics into dichotomous measures that could differentiate intervention effects based on the visual analysis results. To determine the extent to which the non-overlap methods of IRD, Tau-U, and Baseline Corrected Tau outcomes agreed with interpretations based on visual analysis, the author used Kappa coefficient to analyze the agreement between the effect size metrics outcomes and visual analysis judgements for all 924 phase contrasts included in this review.

Kappa (Cohen, 1960) is a statistic that evaluates inter-rater reliability. It can be used with ordinal or nominal measurement scales. Kappa counts for the probability of agreement based on chance. Kappa is used to evaluate the inter-rater agreement between two raters or between two types of classification systems on a dichotomous outcome. Kappa is calculated by looking to the relative observed agreement and subtracting the probability based on chance and dividing by 1 minus the probability based on chance. Kappa can be expressed as the following equation: $k = (Pr(a) – Pr(e))/ (1- Pr(e))$.

The author used the SPSS software package to obtain Kappa coefficient outcomes. First, the scores for each one of three non-overlap metrics were entered and computed separately. For each metric, the author inserted the scores for two variables (e.g., Tau-U and visual analysis) into SPSS Data Editor. The values for both the visual analysis variable and the metric variable were adjusted to two possible values (0 for “no effect” and 1 for “effect”). The three non-overlap variables and the visual analysis variable were adjusted as “nominal”. Second, Crosstabs option was chosen from the

Descriptive Statistics at the Analyze drop-down menu. Third, the scores from the visual analysis were entered in the Column box and each one of the scores of the three non-overlap metrics were entered in the Row box. Lastly, Kappa coefficient was selected for the Statistics option. SPSS provided two tables (i.e., Crosstabulation and Symmetric Measures) for each one of the three non-overlap metrics. Kappa coefficient calculated the agreement between visual analysis and non-overlap metrics and provided the corresponding Kappa coefficient percentages. Kappa coefficient can be interpreted using different guidelines (i.e., Landis & Koch, 1977; Fleiss, 1981; Altman, 1991). The author used Landis and Koch guidelines following Petersen & Brown (2012) and Chen et al. (2016) examples. Values from 0.0 to 0.2 show slight agreement, 0.21 to 0.40 show fair agreement, 0.41 to 0.60 show moderate agreement, 0.61 to 0.80 show substantial agreement, and 0.81 to 1.0 show almost perfect or perfect agreement.

Test for Normality and Pearson Correlation

A normality test was conducted to determine if the data roughly fit a bell curve shape and Pearson correlation was used to determine the strength and direction of a linear relationship between IRD, Tau-U and Baseline Corrected Tau, and visual analysis outcomes. The author used correlation coefficient *r* for all 924 phase contrasts included in this review.

Pearson's correlation coefficient (Cohen, 1988) is a statistical measure of the strength of a linear relationship between paired data. Usually, the coefficient goes by *r* and has three outcomes. Correlations are never lower than -1 or higher than 1. A correlation of -1 indicates that the two variables are perfectly, negatively, and linearly related. A correlation coefficient of 1 means that the two variables are perfectly,

positively, and linearly related. Last, a correlation of 0 means that the two variables do not have any linear relation. According to Cohen (1988), the strength can be assessed by the following general guidelines: (a) $0.10 < | r | < 0.30$ small / weak correlation; (b) $0.30 < | r | < 0.50$ medium / moderate correlation; and (c) $0.50 < | r |$ large / strong correlation.

The author used the SPSS software package to obtain test for normality and Pearson's correlation coefficient outcomes. To test for normality, the scores for the four variables (i.e., visual analysis and three non-overlap metrics outcomes) were entered into SPSS Data Editor. Then, Explore option was chosen from the Descriptive Statistics at the Analyze drop-down menu. Last, the visual analysis and the three non-overlap metrics outcomes were entered in the dependent variables box. To obtain Pearson's correlation coefficient outcomes, the scores for each one of three non-overlap metrics were entered and computed separately. For each metric, the author inserted the scores for two variables (e.g., Tau-U and visual analysis) into SPSS Data Editor. The values for the visual analysis variable were adjusted to two possible values (0 for "no effect" and 1 for "effect") and the values for the metric variable were simply entered. The three non-overlap variables and the visual analysis variable were adjusted as "scale". Second, Bivariate option was chosen from the Correlate at the Analyze drop-down menu. Third, the scores from the visual analysis and the three non-overlap metrics were entered in the variables box. Lastly, Pearson coefficient was selected for the Correlation Coefficients option and two-tailed was selected for the test of significance. SPSS provided two tables (i.e., Test of Normality and Correlations) for each one of the three non-overlap metrics and for the visual analysis outcomes.

Factors Affecting Analysis

Previous research suggests that agreement between visual and statistical analysis can be expected when the data is clear. However, the presence of certain patterns, such as short data sets, trend, and variability in the data can complicate analysis. The use of non-overlap methods can help keep visual analysis objective (Brossart et al., 2014). Following Brossart and colleagues' recommendations, the author identified the graphs that demonstrated the following patterns: (a) baseline trend, (b) short data phases (i.e., between 3 and 6 data points), (c) overlap, and (d) autocorrelation. The purpose was to analyze the impact of these common patterns in the interpretation of visual analysis of the data and the scores of the selected non-overlap methods.

The author analyzed the effect of number of data points in the interpretation of visual and statistical analysis. First, two separate Excel sheets were created, one for baseline phases and the other for intervention phases. Each Excel sheet contained the names of the contrasts (e.g., unique ID number), the number of data points, and the outcomes of visual analysis and the three non-overlap methods. Based on the accepted standards for single-case designs, three to five data points reported in an experimental phase were considered short baseline or intervention phase, respectively. Second, the author used a unique color to code for those contrasts that were considered to have a weak effect by visual analysis and the three methods. A different unique color was used for the contrasts considered to have a strong effect by visual analysis and the three methods. To separate weak from strong effect, the author used the same cutoff scores that were used in the previous research questions (Tau-U = 0.61; IRD = 0.65; Baseline Corrected Tau = 0.41). Third, the author counted all of the contrasts in each of four

categories (i.e., short/long phase and weak/strong effect) for visual analysis, Tau-U, IRD, and Baseline Corrected Tau. Fourth, the mean and the median of Tau-U, IRD and Baseline Corrected Tau scores were calculated for both categories (i.e., short/long phase). Finally, the author analyzed the outcomes to identify possible patterns in the data.

The author analyzed the effect of overlap in the interpretation of visual and statistical analysis. To make sure the visual analysis judgements were reliable when identifying overlap, the author used all of the 272 contrasts rated by two visual analyst raters for inter-rater reliability. Then, the author evaluated the agreement between the two raters. The author divided the contrasts into three categories: (a) contrasts with overlap according to both raters, (b) contrasts without overlap according to both raters, and (c) contrasts for which there was no agreement regarding overlap. Then, the author counted the number of contrasts in each category and calculated the percentage of agreements versus disagreements by adding the agreements for overlap and non-overlap categories together, independently dividing each of the two categories (i.e., agreements and disagreements) by the total number of contrasts (n = 272) and multiplying by 100 to obtain percentage scores.

Once reliability was established, the author used all 924 contrasts to analyze the relationship between overlap and agreement between visual analysis and the three non-overlap methods. A cutoff score was selected for each non-overlap method. The author selected the cutoff score with the highest Kappa score identified in the analysis of research question 2 (Tau-U = 0.61; IRD = 0.65; Baseline Corrected Tau = 0.41). All the scores at the cutoff or above were considered to show effect. All the scores below the cutoff were considered to show no effect. Regarding visual analysis, all the contrasts that

scored 3, 4 and 5 were considered to show effect. All the contrasts that scored 0, 1 and 2 were considered not to show effect. Next, the author counted the number of contrasts showing effect and no effect that were in agreement with visual analysis. This step was completed for each non-overlap method. Finally, the author calculated the overall percentage of agreements for each non-overlap method. To accomplish this step, the author divided the number of total agreements by the number of contrasts with overlap and multiplied the result by 100. One category was not considered relevant in the analysis (i.e., contrasts without overlap that were considered not effective by both visual analysts) because the category contained only one contrast and none of the non-overlap methods identified as not effective.

To see how the presence of trend in the baseline affected the agreement between visual analysis judgements and the scores of Tau-U, IRD, and Baseline Corrected Tau, the author used the same cutoff scores that were used to answer the previous analysis (Tau-U = 0.61; IRD = 0.65; Baseline Corrected Tau = 0.41). Once the scores of the three methods were converted into dichotomous outcomes (i.e., effect, no effect), the author counted the number of contrasts identified as effective and not effective by visual analysts raters and by each of the three methods. A simple percentage was calculated to find the agreement between visual analysis and the methods on the number of effective and non-effective contrasts. Finally, the author calculated the overall percentage of agreements for each non-overlap method. To accomplish this step, the author divided the number of total agreements by the number of contrasts with trend and multiplied the result by 100.

Microsoft Excel was used to identify the contrasts with autocorrelation (i.e.

positive and negative). Lag 1 autocorrelation coefficients for the baseline and treatment phases were calculated. The author analyzed the relationship between autocorrelation and number of data points, trend, visual analysis rater judgements, and the three non-overlap methods scores. The author looked for any patterns related to these relationships present in the data. The data was analyzed at two different levels: positive vs negative autocorrelation and different magnitudes of autocorrelation. The author first divided all of the phase contrasts into three categories: contrasts with positive autocorrelation, contrasts with negative autocorrelation, and contrasts without autocorrelation. Then, the relationships between autocorrelation and average of data points, presence of trend in the baseline, outcomes of visual analysis, and outcomes of non-overlap methods were analyzed for each category. The same procedure was employed for both autocorrelation in the baseline and the intervention phases.

Each of the previous categories were further divided into large, medium, and small magnitude of autocorrelation based on commonly used guidelines for the interpretation of correlation outcomes. See Table 5 for a summary of these guidelines. The same procedures used for comparing positive vs negative autocorrelation were used to compare each level of autocorrelation magnitude.

Range of Effective Interventions

Finally, the author identified the typical range of non-overlap estimates for IRD, Tau-U, and Baseline-Corrected Tau that are considered effective according to the visual analyst raters in this study. In order to identify the typical range of non-overlap estimates from interventions deemed effective by visual analysis, the author used only the phase

Table 5

Guidelines to Interpreting Pearson's Correlation Coefficient

	Coefficient, *r*	
Strength of association	Positive	Negative
Small	0.10 to 0.30	-0.10 to -0.30
Medium	0.30 to 0.50	-0.30 to -0.50
Large	0.50 to 1.00	-0.50 to -1.00

contrasts that were rated by two visual analysis raters (i.e., the 272 contrasts included in IRR). Also, only the phase contrasts that were scored by the two raters with a 5 were included (contrasts without confounding factors). The author decided to use only this limited set of contrasts because it was determined that factors such as trend, overlap, and immediacy of effect had a strong effect on the non-overlap estimates to the point that there was no clear pattern. All non-overlap methods showed a range between 0.00 and 1.00 for the contrasts with confounding factors.

Inter-rater Reliability

Inter-rater reliability (IRR) was calculated for inclusion, coding, rigor of the studies, effect size computation, and visual analysis using the same procedure. See Table 6 for a description. IRR was calculated by dividing the total number of agreements by the total number of agreements plus disagreements and multiplying by 100 (Cohen &

Swerdlik, 2005). In case of disagreement, the primary rater outcome was used for consistency.

Inclusion

The primary author independently reviewed the titles, abstracts and method section of each study identified for inclusion against the exclusion criteria. From the initial finding of 9,140 articles yielded by electronic searches, a total of 132 were included for this review. A sample of 46 (35%) articles was randomly selected for IRR. A faculty member reviewed each of the 46 articles to confirm that they met the inclusion criteria. Participants agreed on 44 of 46 studies. The process yielded an IRR of 96% agreement.

Coding

A master's student randomly coded 35% of the studies (46 out of 132) using the same coding procedures as the author. Then, the author used the same formula used in the inclusion inter-rater reliability to calculate coding inter-rater reliability. The master's student reached agreement on 40 of the 46 studies and the agreement coefficient was 0.89%.

Rigor of the Studies

A second-year doctoral student reevaluated 35% of the studies (46 out of 132). The 46 studies were randomly selected to compute IRR on the rigor of the studies. The doctorate student reached agreement on 38 of the 46 studies and the agreement coefficient was 0.84%.

Effect Size Computation

The author randomly selected 36% of the studies (48 studies) to compute IRR on the calculation of Tau-U, IRD, and Corrected-Baseline Tau. Three doctoral students were randomly assigned to each of the non-overlap methods and independently made the calculations. The students reached agreement on 43 of 48 studies and the agreement coefficient was 0.90%. The author recalculated the effect sizes of the five studies for which there was no agreement and corrected them prior to using them in the analysis.

Visual Analysis

One master's student and two doctoral students independently analyzed 30% of the studies and phase contrasts to determine the reliability of the visual analysis ratings. The author randomly selected 48 studies and almost equally distributed 272 phase contrasts among the participants to compute IRR on visual analysis. Each participant was provided with a booklet that contained phases, graphs and their rubrics. Participants agreed on visual analysis results for 216 out of the 272 independently analyzed phase contrasts (IRR = 79%).

Table 6

Description of Inter-Rater Reliability for Inclusion, Coding, Rigor of Studies, Effect Size Computation, and Visual Analysis

Inter-rater reliability	*N* of studies	Rater position	*N* of raters	Percentage
Inclusion	46 (35%)	Faculty member	1	96%
Coding	46 (35%)	Master's student	1	89%
Rigor of studies	46 (35%)	Doctoral student	1	84%

Effect size computation	48 (36%)	Doctoral students	3	90%
Visual analysis	48 (36%)	Master's and Doctoral students	3	79%

Note. N = number; a total of 132 studies were included for this review.

CHAPTER III

RESULTS

Visual Analysis and Non-Overlap Methods

There were four questions in this study: (a) to what extent are IRD, Tau-U, and Baseline Corrected Tau sensitive and specific in detecting intervention effects in SCD studies? (b) to what extent do IRD, Tau-U and Corrected-Baseline Tau outcomes agree with interpretations based on visual analysis? (c) to what degree does autocorrelation, number of data points, presence of undesired trend in baseline or intervention, and degree of overlap affect interpretation in visual analysis and statistical analysis? (d) what is the typical range of non-overlap estimates from interventions deemed effective by visual analysts?

In order to answer the previous study questions, two approaches were first used to measure the effectiveness of the interventions included in all the studies of this review. The first approach was visual analysis in which participants examined 924 graphs and decided whether an effect existed or did not exist. Following What Works Clearinghouse standards for evaluating SCD, ten trained visual analyst raters evaluated changes in level, trend, variability, immediacy of change between the baseline and treatment phase and overlapping between the two phases. The second approach was calculating effect size using Tau-U, IRD, and Baseline Corrected Tau metrics. After calculating the effectiveness of the interventions using visual analysis, Tau-U, IRD, and Baseline Corrected Tau, additional methods were used to answer each question of this review.

Sensitivity and Specificity

To answer question one, “To what extent are IRD, Tau-U, and Baseline Corrected Tau sensitive and specific in detecting intervention effects in SCD studies?", Receiver Operating Characteristic (ROC) analysis was used to assess the extent to which Tau-U, IRD, and Baseline Corrected Tau can find the same dichotomous outcome of visual analysis. The author used ROC analysis and the area under the curve (AUC) to examine the sensitivity and specificity of each metric to predict group membership (i.e., effect or no effect) established by visual analysis. A ROC curve shows the sensitivity and specificity for each possible cutoff score.

After visually analyzing all 924 phase contrasts, visual analyst raters identified 788 contrasts as being positive for change (effect) and 136 contrasts as being negative for change (no effect). Table 7 shows the results of visual analysis.

Table 7

Visual Analysis Judgement on the Effectiveness of the 924 Phase Contrasts

Visual Analysis	Number of Phases
Positive	788
Negative	136

In Figure 3, sensitivity and specificity of Tau-U, IRD and Baseline Corrected Tau are graphically plotted. The Y-axis represents sensitivity (true positive rate) which is the probability that the phase contrast would be identified as effective when it is truly

effective according to visual analysis. X-axis represents specificity (true negative rate) which is the probability that the phase contrast would be identified as not effective when it is truly not effective according to visual analysis. Each metric was represented with a curve and was compared to the reference line that represents zero sensitivity and zero specificity. When a classifier cannot distinguish between the two groups, the area under the curve (AUC) is .50. However, when there is a perfect separation of the two groups, the area under the ROC curve is 1.00tud. Most importantly, the AUC of each metric was reported to understand how accurate the metric was to correctly classify the contrasts graphs as effective or not effective at each possible cutoff score.

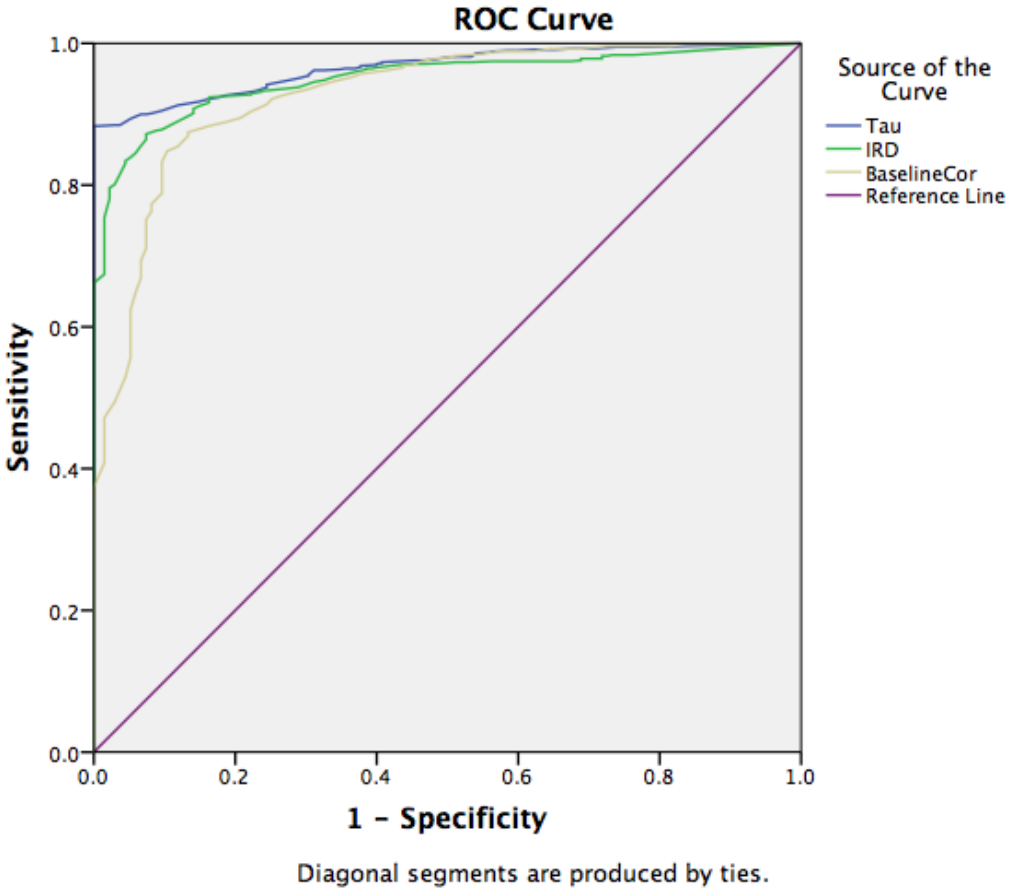

Figure 3. Sensitivity and specificity of Tau-U, IRD and Baseline Corrected Tau.

The AUC for Tau-U was found to be 0.97. The AUC for IRD was 0.95. The AUC for Baseline Corrected Tau was 0.93. Each of the three metrics can be considered an excellent metric based on this analysis. However, Tau-U had the largest AUC with 0.97 and IRD was close behind. Baseline Corrected Tau had the smallest AUC. These differences between metrics turned out to be statistically significant. Each of the three metrics showed a *p*-value of less than .05 so they are statistically significant ROC curves. See Table 8 for a summary of the AUC values. A test with a 0.90 AUC or higher is considered an excellent test. A test between 0.80 and 0.90 is a good test. A test between 0.70 and 0.80 is a fair test. A test between 0.60 and 0.70 is considered a poor test (Zweig & Campbell, 1993). Less than that is an unpredictable test and is not recommended for use.

Table 8

AUC for Tau-U, IRD and Baseline Corrected Tau

	Area	Std. Error	Sig.	95% CI	
				LL	UL
Tau-U	0.97	0.05	0.05	0.96	0.98
IRD	0.95	0.07	0.05	0.94	0.97
Baseline Corrected Tau	0.93	0.12	0.05	0.91	0.95

Note. CI= Confidence interval; Sig.= Significance level; Std. Error= Standard error; LL= lower limits; UL= upper limits.

The author selected three different cutoff scores to report the degree of sensitivity and specificity of each non-overlap method. The three cutoff scores for Tau-U were: 0.61, 0.70, and 0.79. The three cutoff scores for IRD were: 0.65, 0.72, and 0.79. The three cutoff scores for Baseline Corrected Tau were: 0.41, 0.50, and 0.59.

The author calculated sensitivity and specificity for each of the three cutoff scores for each metric. Regarding sensitivity, Tau-U identified 719 (91%), 696 (88%), and 634 (80%) of true positives; IRD identified 715 (91%), 667 (85%), and 617 (78%) of true positives; and Baseline Corrected Tau identified 712 (90%), 648 (82%), and 547 (69%) of true positives. Regarding specificity, Tau-U identified 119 (88%), 135 (99%), and 135 (99%) of true negatives; IRD identified 116 (85%), 128 (94%), and 133 (98%) of true negatives; and Baseline Corrected Tau identified 105 (77%), 122 (90%), and 127 (93%) of true negatives.

Regarding the incidence of type I errors, the three non-overlap methods found the following contrasts to be positive when visual analysis determined them to be negative (the percentage of errors related to the total number of non-effective contrasts according to visual analysis is also reported): Tau-U = 17 (13%), IRD = 20 (15%), Baseline Corrected Tau = 31 (23%). Regarding the incidence of type II errors, the three non-overlap methods found the following contrasts to be negative when visual analysis determined them to be positive (the percentage of errors related to the total number of effective contrasts according to visual analysis is also reported): Tau-U = 69 (8%), IRD = 73 (9%), Baseline Corrected Tau = 76 (10%).

Kappa Coefficients

To answer question two, "to what extent do IRD, Tau-U and Corrected-Baseline Tau outcomes agree with interpretations based on visual analysis?", Kappa coefficients were calculated. Based on judgement of the ten visual analyst raters, 788 phase contrasts were found to be effective and only 136 were not effective. Kappa coefficients were calculated for the three cutoff scores selected earlier for each non-overlap method. Cutoff scores and Kappa coefficients for Tau-U were: 0.61 (Kappa = 0.68), 0.70 (Kappa = 0.69), and 0.79 (Kappa = 0.54). The three cutoff scores and Kappa coefficients for IRD were: 0.65 (Kappa = 0.66), 0.72 (Kappa = 0.59), and 0.79 (Kappa = 0.50). The three cutoff scores and Kappa coefficients for Baseline Corrected Tau were: 0.41 (Kappa = 0.59), 0.50 (Kappa = 0.52), and 0.59 (Kappa = 0.37). The author selected the highest Kappa coefficient for each method. This coefficient was always the one obtained using the lowest cutoff score.

The selected Kappa coefficients for the three non-overlap methods were all above 0.41 and ranged from 0.59 - 0.68. In other words, Baseline Corrected Tau fell within a moderate agreement level compared to visual analysis. Tau-U and IRD showed a substantial agreement with visual analysis. Baseline Corrected Tau effect size had the lowest number of agreements with a Kappa coefficient score of 0.59. The highest was Tau-U effect size with a Kappa coefficient score of 0.68. Last, IRD effect size had a Kappa coefficient score of 0.66. All of the three Kappa scores were associated with *p* values less than 0.05. Therefore, all the three non-overlap methods were statistically significant and had moderate to substantial level of agreement. See Table 9 for Kappa coefficient values.

Table 9

Kappa Coefficient for Tau-U, IRD and Baseline Corrected Tau

Effect size	Kappa value	*N* of cases	Std. Error	Sig.
Tau-U	0.68	924	0.03	0.05
IRD	0.66	924	0.03	0.05
Baseline Corrected Tau	0.59	924	0.04	0.05

Note. N= Number of phase contrasts; Sig.= Significance level; Std. Error= Standard error.

Table 10 shows the crosstabulation of the relationship between Tau-U and visual analysis. It demonstrated the sensitivity and specificity of Tau-U. Visual analysis found 788 phase contrasts to be effective, 719 phase contrasts were also found by Tau-U to be effective with a sensitivity of 91%. Of the 136 phase contrasts that visual analysts found not effective, 119 of these phase contrasts were determined by Tau-U to be not effective with a specificity of 88%.

Table 10

Crosstabulation Between Visual Analysis and Tau-U

		Visual analysis		
		No effect	Effect	Total
	N	119	69	188

Tau-U	No effect	%	88%	08%	20%
		N	17	719	736
	Effect	%	13%	91%	80%

Note. N= Number of phase contrasts.

Table 11 displays the crosstabulation of the relationship between IRD and visual analysis. Furthermore, it showed the sensitivity and specificity of IRD compared to the standard measurement in this study of visual analyst rater judgements. Of 788 phase contrasts that were found to be effective by visual analysis, 715 were also found by IRD to be effective with a sensitivity of 91%. Of 136 phase contrasts that were found not effective by visual analysis, 116 were determined by IRD to be not effective with a specificity of 85%.

Table 11

Crosstabulation Between Visual Analysis and IRD

			Visual analysis		
			No effect	Effect	Total
IRD	No effect	*N*	116	73	189
		%	85%	09%	20%
	Effect	*N*	20	715	735
		%	15%	91%	80%

Note. N= Number of phase contrasts.

Table 12 presented the crosstabulation of the relationship between Baseline Corrected Tau and visual analysis. From this table, sensitivity and specificity of Baseline Corrected Tau were identified. Of 788 phase contrasts that were found to be effective by visual analysis, 712 were also found by Baseline Corrected Tau to be effective with a sensitivity of 90%. Of 136 phase contrasts that were found not to be effective by visual analysis, 105 were determined by Baseline Corrected Tau to be not effective with a specificity of 77%.

Table 12

Crosstabulation Between Visual Analysis and Baseline Corrected Tau

			Visual analysis		
			No effect	Effect	Total
Baseline Corrected Tau	No effect	*N*	105	76	181
		%	77%	10%	20%
	Effect	*N*	31	712	743
		%	23%	90%	80%

Note. N= Number of phase contrasts.

Sample Characteristics

A Shapiro-Wilk's test ($p < .05$) showed that the scores of visual analyses and three non-overlap metrics were not normally distributed. Visual analysis had a skewness

of -2.07 (ES = 0.08) and a kurtosis of 2.03 (ES = 0.16). Tau-U had a skewness of -1.50 (ES = 0.08) and a kurtosis of 1.22 (ES = 0.16). IRD had a skewness of -1.60 (ES = 0.08) and a kurtosis of 1.85 (ES = 0.16). Baseline Corrected Tau had a skewness of -0.80 (ES = 0.08) and a kurtosis of 0.17 (ES = 0.16). In conclusion, skewness and kurtosis for all four variables were skewed and kurtoic, and differed significantly from normality. So, the null hypothesis for this test of normality is that the scores were not normally distributed. Table 13 summarizes the results of the normality test of visual analysis and the three non-overlap metrics.

Table 13

Normality Test of Visual Analysis and Three Non-Overlap Metrics

	Shapiro-Wilk Test of Normality		
Variables	Statistic	df	Sig.
Tau-U	0.73	924	0.05
IRD	0.75	924	0.05
Baseline Corrected Tau	0.94	924	0.05
Visual Analysis	0.42	924	0.05

Pearson Correlation

A Pearson's correlation was computed to assess the relationship between visual analysis and three non-overlap metrics (i.e., Tau-U, IRD, and Baseline Corrected Tau). Overall, there was a strong, positive correlation between visual analysis and three non-

overlap metrics. Also, there was a strong, positive correlation between the three non-overlap metrics themselves.

There was a positive correlation between visual analysis and Tau-U, r = 0.74, n = 924, p = 0.05. Also, there was a positive correlation between visual analysis and IRD, r = 0.67, n = 924, p = 0.05. Last, there was a positive correlation between visual analysis and Baseline Corrected Tau, r = 0.63, n = 924, p = 0.05. Table 14 summarizes the results of the bivariate correlations between the four single-case design measurements.

Table 14

Bivariate Correlations Between Four Single-Case Design Measurements

Measure	1	2	3	4
1.Visual Analysis	____	0.74	0.67	0.63
2.Tau-U	0.74	____	0.90	0.83
3.IRD	0.63	0.90	____	0.79
4.Baseline Corrected Tau	0.63	0.83	0.79	____

Factors Affecting Analysis

To answer question three “to what degree does autocorrelation, number of data points, presence of undesired trend in baseline or intervention, and degree of overlap affect interpretation in visual analysis and statistical analysis?", the author analyzed the relationship between the number of data points and visual and statistical analysis. This

analysis showed that the contrast effect sizes were weaker when the phases were longer. In other words, the more data points, the weaker the effect shown for the contrast. The pattern was the same for both the baseline and intervention phases.

The relationship between two variables (i.e., number of data points in the baseline and methods) was analyzed first. See table 15 for a summary. After counting how many contrasts had short and long baselines that either showed strong or weak effect, data for short baselines were the following: 346 strong effect and 73 weak effect (Tau-U), 349 strong effect and 71 weak effect (IRD), 339 strong effect and 82 weak effect (Baseline Corrected Tau), 351 strong effect and 69 weak effect (visual analysis). Data for long baselines were the following: 390 strong effect and 114 weak effect (Tau-U), 383 strong effect and 121 weak effect (IRD), 404 strong effect and 100 weak effect (Baseline Corrected Tau), 384 strong effect and 121 weak effect (visual analysis).

Table 15

Relationship Between Number of Data Points in the Baseline, Metrics and Visual Analysis

	Short baseline		Long baseline	
	Strong effect	Weak effect	Strong effect	Weak effect
Tau-U	346	73	390	114
IRD	349	71	383	121
Baseline Corrected Tau	339	82	404	100
Visual analysis	351	69	384	121

The relationship between the number of data points in the intervention phases and effect sizes was analyzed next. See table 16 for a summary. After counting how many contrasts had short and long interventions that either showed strong or weak effect, data for short interventions were the following: 119 strong effect and 23 weak effect (Tau-U), 123 strong effect and 19 weak effect (IRD), 128 strong effect and 14 weak effect (Baseline Corrected Tau), 117 strong effect and 25 weak effect (visual analysis). Data for long interventions were the following: 617 strong effect and 165 weak effect (Tau-U), 612 strong effect and 170 weak effect (IRD), 615 strong effect and 167 weak effect (Baseline Corrected Tau), 618 strong effect and 164 weak effect (visual analysis).

Table 16

Relationship Between Number of Data Points in the Intervention, Metrics and Visual Analysis

	Short intervention		Long intervention	
	Strong effect	Weak effect	Strong effect	Weak effect
Tau-U	119	23	617	165
IRD	123	19	612	170
Baseline Corrected Tau	128	14	615	167
Visual analysis	117	25	618	164

The relationship between the number of data points, the mean and the median scores for each method and length of phase were then calculated. See tables 17 and 18 for a summary of baseline data and intervention data. Mean scores for baselines were the following: 0.82 short baseline and 0.77 long baseline (Tau-U), 0.84 short baseline and 0.76 long baseline (IRD), 0.58 short baseline and 0.60 long baseline (Baseline Corrected Tau). Median scores for baselines were the following: 1.00 short baseline and 0.91 long baseline (Tau-U), 1.00 short baseline and 0.88 long baseline (IRD), 0.62 short baseline and 0.66 long baseline (Baseline Corrected Tau).

Table 17

Metrics' Mean and Median Scores for Short and Long Baselines

	Short baseline		Long baseline	
	Mean	Median	Mean	Median
Tau-U	0.82	1.00	0.77	0.91
IRD	0.84	1.00	0.76	0.88
Baseline Corrected Tau	0.58	0.62	0.60	0.66

Mean scores for interventions were the following: 0.82 short intervention and 0.79 long intervention (Tau-U), 0.87 short intervention and 0.80 long intervention (IRD), 0.69 short intervention and 0.59 long intervention (Baseline Corrected Tau). Median scores for interventions were the following: 1.00 short intervention and 0.90 long

intervention (Tau-U), 1.00 short intervention and 0.88 long intervention (IRD), 0.75 short intervention and 0.57 long intervention (Baseline Corrected Tau).

Table 18

Metrics' Mean and Median Scores for Short and Long Interventions

	Short intervention		Long intervention	
	Mean	Median	Mean	Median
Tau-U	0.82	1.00	0.79	0.90
IRD	0.87	1.00	0.80	0.88
Baseline Corrected Tau	0.69	0.75	0.59	0.57

The author evaluated the reliability of visual analysis concerning the detection of data overlap in phase contrast graphs. See table 19 for a summary. A total of 272 contrast graphs were analyzed by three additional raters for inter-rater agreement. Consequently, each graph was rated by a total of two raters. In 253 (93%) graphs, the two raters agreed in their judgement of whether overlap was present or was not present. They disagreed on 19 (7%) graphs. This outcome indicated that visual analysis judgements were reliable when trying to detect data overlap in the graphs. Regarding the effect of data overlap on the ability of Tau-U, IRD, and Baseline Corrected Tau to agree with visual analysis effect judgements, the results were similar for the three non-overlap methods.

Table 19

Agreement Between Two Visual Analysis Raters Identifying Overlap or Lack of Overlap on 272 Phase Contrasts

	Overlap	Non-overlap	Overall Agreement
Agreement	148	105	93%
Disagreement	19		7%

The author found 530 contrasts with data overlap and 394 contrasts without overlap. See table 20 for a summary. The contrasts were divided into four categories: no overlap and effect, no overlap and no effect, overlap and effect, and overlap and no effect. When scoring contrast graphs without data overlap, all methods were able to identify 98% of the 393 contrast graphs showing effect. There was only one graph without data overlap and showing no effect according to visual analysis. All the methods scored this graph incorrectly because the three methods identify it as effective. However, when the graphs had some degree of overlap, the ability of the methods to agree with the visual analysis was much lower. Visual analysis determined that there were 530 graphs with data overlap. Of these graphs, 342 showed effect and 188 showed no effect. Tau-U successfully identified 260 (76%) graphs showing effect and 135 (71%) graphs showing no effect (overall agreement = 75%). IRD successfully identified 242 (70%) graphs showing effect and 140 (74%) graphs showing no effect (overall agreement = 72%). Baseline Corrected Tau agreement with visual analysis regarding effect/no effect was somewhat different. The method successfully identified 278 (81%) graphs showing effect and 109 (58%) graphs showing no effect. The overall agreement was 73%. See table 21

and table 22 for a summary of agreement between visual analysis and the three non-overlap methods on contrasts with overlap and without overlap, respectively. The three methods had exactly the same agreement with visual analysis when the contrasts had no overlap.

Table 20

Visual Analysis Judgements on the Overlap Between Baseline and Intervention Phases for 924 Phase Contrasts

Phase contrast	*N*
With overlap	530
Without overlap	394

Note. *N*= Number; Within the overlap, 342 showed effect and 188 showed no effect

Overall, visual analysis identified 252 phase contrasts with trend and 672 contrasts without trend. Among the contrasts with baseline trend, there were 117 contrasts showing effect and 135 contrasts showing no effect. Among the contrasts without baseline trend, 618 contrasts were effective and 54 were not effective. In general, Tau-U, IRD, and Baseline Corrected Tau showed a high level of agreement with visual analysis when identifying effective contrasts showing baseline trend. Tau-U agreed on 112 contrasts (96%), Baseline Corrected Tau agreed on 110 contrasts (94%), and IRD agreed on 109 contrasts (93%). The percentage of agreement when identifying non-effective contrasts was much lower. Tau-U agreed on 91 contrasts (67%), IRD agreed on 90 contrasts (67%) and Baseline Corrected Tau agreed on 84 contrasts (62%). The overall

Table 21

Agreement Between Visual Analysis and the Three Non-Overlap Methods on Contrasts with Overlap

Overlap methods	Effect	No effect	Overall Agreement
Tau-U	260 (76%)	135 (71%)	75%
IRD	242 (71%)	140 (74%)	72%
Baseline Corrected Tau	278 (81%)	109 (58%)	73%

Note. Visual analyses found 530 phases with overlap. 342 contrasts were effective and 188 were not effective according to visual analysis.

Table 22

Agreement Between Visual Analysis and the Three Non-Overlap Methods on Contrasts without Overlap

Overlap methods	Effect	No effect	Overall Agreement
Tau-U	385 (98%)	0 (0%)	98%
IRD	384 (98%)	0 (0%)	98%
Baseline Corrected Tau	385 (98%)	0 (0%)	98%

Note. Visual analysis found 394 phases without overlap. 393 contrasts were effective and 1 was not effective according to visual analysis.

Table 23

Visual Analysis Judgements for 252 Phase Contrasts with Trend

Overlap methods	*N*	Effect	No effect	Overall Agreement
Tau-U	252	112 (96%)	91(67%)	81%
IRD	252	109 (93%)	90 (67%)	79%
Baseline Corrected Tau	252	110 (94%)	84 (62%)	77%

Note. N= number of phases; visual analyses found 252 phases with trend. 117 contrasts were effective and 135 were not effective according to visual analysis.

agreement was the following: Tau-U = 203/252 (81%), IRD = 199/252 (79%), and Baseline Corrected Tau = 194/252 (77%). See table 23 for a summary. When Baseline Corrected Tau was compared to Tau-U, the agreement was higher: They agreed on 112 (96%) contrasts considered effective by visual analysis and 117 (87%) contrasts considered non-effective by visual analysis. The overall agreement was 229/252 (91%).

In general, Tau-U, IRD, and Baseline Corrected Tau showed a higher level of agreement with visual analysis when identifying effective contrasts without baseline trend. Tau-U agreed on 555 contrasts (90%), Baseline Corrected Tau agreed on 553 contrasts (89%), and IRD agreed on 556 contrasts (90%). The percentage of agreement when identifying non-effective contrasts was much lower. Tau-U agreed on 29 contrasts (54%), IRD agreed on 28 contrasts (52%) and Baseline Corrected Tau agreed on 25 contrasts (46%). The overall agreement was the following: Tau-U = 584/672 (87%), IRD = 584/672 (87%), and Baseline Corrected Tau = 578/672 (86%). See table 24 for a summary of visual analysis judgements for 672 phase contrasts without trend.

Table 24

Visual Analysis Judgements for 672 Phase Contrasts without Trend

Overlap methods	*N*	Effect	No effect	Overall Agreement
Tau-U	672	555 (90%)	29 (54%)	87%
IRD	672	556 (90%)	28 (52%)	87%
Baseline Corrected Tau	672	553 (89%)	25 (46%)	86%

Note. N= number of phases; visual analyses found 672 phases with no trend. 618 contrasts were effective and 54 were not effective according to visual analysis

Among the 924 phase contrasts included in the review, the author identified a large number of contrasts with autocorrelation either in the baseline or intervention phases. Overall, there were 887 (96%) contrasts with autocorrelation either in the baseline, the intervention or both. There were 5 (0.5%) contrasts without autocorrelation. Thirty-two (3.5%) contrasts were not included in this analysis because they presented errors in the calculation of autocorrelation. There was a tendency to get an error message in the Excel sheet when the values in either baseline or intervention phase were the same (e.g., 5, 5, 5). See table 25 for a summary.

Table 25

Autocorrelation in the Baseline and Intervention for 924 Phase Contrasts

Phase contrast	*N*
With autocorrelation	887 (96%)

Without autocorrelation	5 (0.5%)

Note. N= Number; Thirty-two (3.5%) contrasts were not included in this analysis because they presented errors in the calculation of autocorrelation.

The author identified 552 contrasts with autocorrelation in the baseline. In the baseline, 213 (38%) phase contrasts had positive autocorrelation and 339 (61%) phase contrasts had negative autocorrelation. Four contrasts (1%) did not show any autocorrelation. The author found 844 contrasts with autocorrelation in the intervention. Among all, 604 (71%) phase contrasts had positive autocorrelation and 240 (28%) contrasts had negative autocorrelation. Five (1%) contrasts did not show autocorrelation. The contrasts without any correlation were not included in the following analyses because of the limited number. See table 26 for a summary.

Table 26

Positive and Negative Autocorrelation in Baseline and Intervention

Phase contrast	Baseline	Intervention
Positive autocorrelation	213 (38%)	604 (71%)
Negative autocorrelation	339 (61%)	240 (28%)
No autocorrelation	4 (1%)	5 (1%)

Note. 552 contrasts with autocorrelation in the baseline and 844 contrasts with autocorrelation in the intervention.

The author first analyzed the relationship between the number of data points and the presence of autocorrelation (i.e., positive and negative autocorrelation, no

autocorrelation). In the baseline, it was determined that the contrasts with positive autocorrelation had an average of 8.4 data points per contrast. The average for the contrasts with negative autocorrelation was 6.8 data points. The average number of data points for contrasts without autocorrelation was 9.5. Regarding the intervention phases, it was determined that the contrasts with positive autocorrelation had an average of 18 data points per contrast. The average for the contrasts with negative autocorrelation was 10 data points. The average number of data points for contrasts without autocorrelation was 7.2. See table 27 for a summary.

Table 27

Relationship Between Number of Data Points and Autocorrelation

Phase contrast	Baseline	Intervention
Positive autocorrelation	8.4	18
Negative autocorrelation	6.8	10
No autocorrelation	9.5	7.2

Note. Numbers are the average of data points.

The author noticed that, in general, contrasts with a small number of data points tended to concentrate around the large autocorrelation values (i.e., 1.00). When the average number of data points was calculated for each of the three autocorrelation values (i.e., large, medium, small), this fact was confirmed. The author calculated the averages for baseline positive and negative autocorrelation, as well as intervention positive and negative autocorrelation. These four averages are the following: (a) small autocorrelation:

baseline positive = 10, baseline negative = 10, intervention positive = 17, intervention negative = 13; (b) medium autocorrelation: baseline positive = 9, baseline negative = 7, intervention positive = 17, intervention negative = 10; (c) large autocorrelation: baseline positive = 6.7, baseline negative = 4, intervention positive = 19, intervention negative = 5. See table 28 for a summary.

Table 28

Relationship Between Number of Data Points and Three Autocorrelation Values

	Autocorrelation		
Autocorrelation	Small	Medium	Large
Positive baseline	10	9	6.7
Negative baseline	10	7	4
Positive intervention	17	17	19
Negative intervention	13	10	5

Note. Numbers are the average of data points per phase.

The relationship between baseline trend and autocorrelation in either the baseline or the intervention was then analyzed for all the contrasts. A total of 653 (74%) autocorrelated contrasts had also baselines trend. Trend was present in 339 (68%) of the 556 contrasts with autocorrelation in the baseline. Positive and negative autocorrelation had a similar percentage of trend in the baseline (69%, 67%). Trend was present in 179 (74%) of the 849 contrasts with autocorrelation in the intervention phase. In this case, the

percentages had a wider range, with 81% of contrasts with positive autocorrelation and 56% of contrasts with negative autocorrelation in the intervention. The analysis of the relationship between small, medium and large autocorrelation and trend shows that 47% of contrasts with positive autocorrelation and trend have a large autocorrelation. This tendency is present with both baseline and intervention correlations.

The relationship between visual analysis judgements and the presence of autocorrelation was also analyzed. Phase contrasts with autocorrelation in the baseline or intervention phases had a similar amount of effective judgements for positive and negative autocorrelation. Of the 213 positively autocorrelated baselines, 169 (79%) of the contrasts were considered effective by visual analysis. Of the 339 negatively autocorrelated baselines, 281 (83%) of the contrasts were considered effective by visual analysis. Of the 604 positively autocorrelated intervention phases, 536 (89%) of the contrasts were considered effective by visual analysis. Of the 240 negatively autocorrelated intervention, 192 (80%) of the contrasts were considered effective by visual analysis. When adding the analysis of the three autocorrelation values (large, medium, small), a pattern emerges. Overall, the larger the magnitude of the autocorrelation, the higher the number of effective judgements by visual analyst raters. The only exception to this tendency were the contrasts that had negative autocorrelation in the baseline. Regardless of the magnitude of the autocorrelation, the percentage of effective judgements was the same (83%). The proportion of effective vs not effective contrasts was the same for the three magnitudes (large, medium, small). See table 29 for a summary.

Table 29

Relationship Between Visual Analysis and the Presence of Autocorrelation

Phase contrast	Effective	Not effective
Positive autocorrelation baseline	169 (79%)	44 (21%)
Negative autocorrelation baseline	281 (83%)	58 (17%)
Positive autocorrelation intervention	536 (89%)	68 (11%)
Negative autocorrelation intervention	192 (80%)	48 (20%)

Note. 556 contrasts with autocorrelation in the baseline and 849 contrasts with autocorrelation in the intervention.

The relationship between autocorrelation and the scores produced by the three non-overlap methods (i.e., Tau-U, IRD, and Baseline Corrected Tau) was analyzed. The outcomes were consistent for the three metrics. When the contrasts had a positive baseline autocorrelation, the average of the metrics scores were consistent with medium to large effect size for the contrast (Tau-U = 0.75, IRD = 0.76, Baseline Corrected Tau = 0.54). When a negative baseline autocorrelation was present, the average contrast scores were also consistent with medium to large effect size (Tau-U = 0.79, IRD = 0.80, Baseline Corrected Tau = 0.56). When the contrasts had a positive intervention autocorrelation, the metrics average contrast scores were consistent with medium to large effect size (Tau-U = 0.81, IRD = 0.80, Baseline Corrected Tau = 0.58). When a negative intervention autocorrelation was present, the average contrast scores were similar (Tau-U = 0.78, IRD = 0.78, Baseline Corrected Tau = 0.61). See table 30 for a summary.

Table 30

Relationship Between Autocorrelation and the Scores of Non-Overlap Methods

	Non-overlap methods		
Autocorrelation	Tau-U	IRD	Baseline Corrected Tau
Positive baseline	0.75	0.76	0.54
Negative baseline	0.79	0.80	0.56
Positive intervention	0.81	0.80	0.58
Negative intervention	0.78	0.78	0.61

Note. Scores are medium to large for the three metrics whit the presence of autocorrelation.

Range of Effective Interventions

To answer question four, "what is the typical range of non-overlap estimates from interventions deemed effective by visual analysts?" the author identified the contrasts considered effective by visual analysis. Then, calculated the equivalent range of scores for the three metrics. Regarding the typical range of non-overlap estimates from interventions deemed effective by visual analysis, the following ranges were identified: Tau-U ranged between 0.65 and 1.00; IRD ranged between 0.65 and 1.00; and Baseline Corrected Tau ranged between 0.45 and 1.00. See table 31 for a summary.

Table 31

Typical Range of Non-Overlap Methods Compared to Visual Analysis

	From	To
Tau-U	0.65	1.00
IRD	0.65	1.00
Baseline Corrected Tau	0.45	1.00

Note. The cutoff scores selected were 0.61 for Tau-U, 0.65 for IRD, and 0.41 for Baseline Corrected Tau

CHAPTER IV

DISCUSSION

This study utilized the outcomes of 132 single-case studies evaluating the effectiveness of interventions for students with ASD in school settings to compare the performance of visual analysis and three effect size metrics (Tau-U, IRD, and Baseline Corrected Tau). The author examined a total of 924 phase contrasts from the sample of 132 studies. First, several aspects of the metrics and visual analysis performance are discussed. Then, the implications for research are discussed. Finally, the author summarizes the limitations and suggestions for future research.

Sensitivity and Specificity

The first question of this review concerns the sensitivity and specificity of the three selected non-overlap methods. Although the area under the curve (AUC) of these methods showed them to be good at distinguishing between effective and non-effective interventions, Tau-U performed better than IRD and Baseline Corrected Tau. One might wonder how a method like Tau-U is so closely related to visual analysis when raters do not always agree on their visual analysis judgements. One possible reason is publication bias. Most published studies have found the interventions to be effective (Kittelman, Gion, Horner, Levin, & Kratochwill, 2017). This fact has been confirmed in this study where 788 out of 924 studies showed effective interventions. It is easy for different analytical methods to agree when the interventions are clearly effective. On the other hand, it should be mentioned that the agreement between the non-overlap methods and visual analysis is probably lower than indicated by the AUC because visual analysis is not 100% reliable.

The outcomes of this study showed the three non-overlap metrics performed well identifying effective interventions (high sensitivity) but tended to incorrectly classify some interventions as effective when they were not effective according to visual analysis. The metrics caught many true positives but mistakenly caught many false positives in the process. To incorrectly classify an intervention as effective when it is not effective is normally referred to as "type I error" or a "false positive".

Type I and Type II Errors

The findings of this study showed that the three non-overlap methods and visual analysis made both type I and type II errors. The three statistical methods had a tendency to make type I errors as opposed to type II errors. This can be explained by the fact that the methods had a high level of sensitivity and they reported some false positives in the process. On the other hand, visual analysis has shown to be less sensitive and make more type II errors but made less type I errors. The high sensitivity of the methods can be beneficial to identify interventions that show small effect visual analysis might not be able to detect. Kazdin (1982) warns that visual analysis should be complemented with statistical analysis because small effects may be important and trend and variability in the data can prevent visual analysts from seeing the effects. The current study confirms Kazdin statement regarding the difficulty of detecting small effects. The author analyzed the agreement between the two visual analysis raters who scored the 272 contrasts included in the inter-rater reliability. The outcomes showed that the two raters had a higher degree of agreements when scoring very effective contrasts or contrasts that were not effective at all. They did not agree as much when scoring contrasts showing a small effect. These outcomes also confirm the findings of Ximenes, Manolov, Solanas, &

Quera (2009). These authors concluded that human judges are fairly good at detecting graphs that show no treatment effect.

Considering that it can be challenging for visual analysis to detect a small effect when a graph has several confounding factors, the author concluded that it is an advantage to use a statistical metric that has a high degree of sensitivity detecting treatment effects. Because of this, the author selected the cutoff score for each non-overlap method that proved to have a high degree of sensitivity but an acceptable degree of specificity. This cutoff score was later used to answer the second question of this review. The cutoff scores selected were: 0.61 for Tau-U, 0.65 for IRD, and 0.41 for Baseline Corrected Tau. Any intervention receiving these scores or above would be considered effective. Tau-U, IRD and Baseline Corrected Tau had a similar level of performance in terms of detecting true positives (sensitivity) using these cutoff scores, with a range of 90% to 91% of effective interventions detected. The metrics were less consistent when ruling out true negatives (specificity). The three metrics showed a lower level of success and the percentages of true negatives ranged from 78% to 88% of non-effective contrasts detected. Tau-U was the metric with the highest sensitivity and specificity. Overall, Baseline Corrected Tau showed the lowest sensitivity and specificity, but differences were small.

Until this point, the author has advocated for sensitive statistical methods to compensate visual analysis difficulty detecting small effect. However, the author also realized that sensitive statistical methods make type I errors which is the tendency to identify false positives. Consequently, the author asked the following question: Which type of error is most important to minimize when evaluating interventions for students

with disabilities? In general, researchers agree that minimizing type I errors is more important. In most fields, type II errors are considered less serious. For example, failing to report an effective intervention can be less damaging than reporting an ineffective intervention as effective. Therefore, most research designs will focus their attention on minimizing type I errors. The use of ineffective interventions can cost money and even be harmful to children. Missing effective interventions, on the other hand, will simply leave things the way they are and will not make things worse. According to the proponents of visual analysis and the findings of this study, visual analysis seems to be a more conservative approach to the evaluation of intervention effects reducing the possibility of making type I errors (Allison et al., 1992; Kazdin, 2011). Considering these facts, the author recommends the use of visual analysis as the primary method of evaluation of intervention effects in order to avoid type I errors. However, because of the relatively low reliability of visual analysis, the author also recommends the use of statistical methods as adjunctive methods. Incorporation of statistical analysis will assist in identifying interventions with small yet clinically significant effects that might go undetected with visual analysis alone as well as assist in identifying no effect when confounding factors are present.

Kappa Coefficients

To calculate the agreement between visual analysis outcomes and the outcomes of Tau-U, IRD and Baseline Corrected Tau, a cutoff score had to be selected for each metric to separate effective from non-effective interventions. Three cutoff scores for each metric had been selected to answer question one so Kappa coefficients were calculated for each of these three cutoff scores. It was noticed that the value of the cutoff scores and the level

of agreement was negatively correlated. The lower the cutoff score, the higher the agreement. It was also noticed that the value of the cutoff score and the sensitivity were also negatively correlated. The lower the cutoff score, the higher the sensitivity. This indicated that sensitivity of the methods and agreement with visual analysis were related. More sensitivity meant more agreement with visual analysis. Furthermore, because sensitivity and specificity are negatively correlated, more specificity meant less agreement with visual analysis. With the intention to make the three metrics as sensitive as possible, the author decided to use the lowest cutoff scores because they have the highest level of sensitivity. After calculating Kappa using the lowest cutoff score, it was found that the three non-overlap methods had a moderate to substantial level of agreement with visual analysis. Tau-U had the highest agreement and Baseline Corrected Tau had the lowest.

Although the three tested metrics in this study showed to be promising methods based on their high sensitivity and specificity, they should not be used as the only method of analysis. The present study confirms previous research regarding the high level of agreement between visual analysis and Tau-U and IRD. However, no matter how much the metrics agree with visual analysis, they must be used only as a complement to visual analysis. For example, non-overlap methods usually account for two or three elements of visual analysis (i.e., level, trend, variability, overlap, immediacy of effect, and consistency across similar phases) but not all the elements at the same time. Even a parametric method such as Standardized Mean Difference cannot account for all the features of a single-case design. Only visual analysis can do that. Single-case researchers should use visual analysis as a primary method examining the effect of interventions until

a new statistical method that will fit all single-case design requirements is developed (Kennedy, 2005).

A clear advantage of using both visual and statistical analysis at the same time is to have more confidence on the results. Researchers can trust their outcomes when both visual and statistical analysis agree on the results. On the other hand, disagreement leads to a lack of confidence and requires further investigation on the effectiveness of the intervention. The outcomes of this study show that there was sometimes a lack of agreement between visual analysis and the three non-overlap metrics, especially when confounding factors were present in the graphs. These findings prompted the author to suggest the use of both visual analysis as a primary method of analysis for single-case research and one or two statistical methods, either parametric or non-parametric methods to support the findings. Moreover, the author suggests to request a second visual analysis judgement by a third expert rater in case of disagreement between visual and statistical analysis. In the current study, the author conducted this second visual investigation and discovered that, in some cases, visual analysis raters had been too conservative in their judgement regarding effect. Taking this step before reporting research findings will increase the confidence on the outcomes of published studies.

Factors Affecting Analysis

Type II errors originally found in this review were of particular interest to the author. After visually reviewing the judgements of the contrasts selected for interrater reliability, the author realized that 32 contrasts have been rated as "not effective" by the visual analysis raters but showed effect by the non-overlap methods. The author concluded that the contrasts showed effect and that the raters had not detected this effect.

The author analyzed the possible causes. Two possible reasons for the errors of judgement were considered: (a) visual analysts were wrong in their judgement; (b) there were some confounding factors that made visual analysis difficult (i.e., overlap, trend, small number of data points, variability, more than one variable line in the graph). Consistent with Ninci and colleagues (2015) findings, visual analysts in this review have shown to have a moderate degree of interrater reliability (79%) and the graphs included a variety of confounding factors. Consequently, the author considered both reasons to be true. Furthermore, it was hypothesized that the confounding factors might be causing the low percentage of inter-rater reliability. The author, then, analyzed the incidence of the confounding factors in each of the 32 contrasts mistakenly rated as not effective by the visual analysts. It was found that all 32 contrasts had at least one confounding factor. The author found a total of 85 factors distributed among the 32 contrasts, including 1 contrast that exhibited five factors and three contrasts that exhibited four factors at the same time. All the 32 contrasts had some degree of overlap. Variability seemed to be the second most common confounding factor (20 contrasts) especially variability combined with trend (9 contrasts). In addition, the effect shown on the contrasts was often small and difficult to see. The following sections will analyze the impact of some of the confounding factors seen in the contrasts included in this review.

Number of Data Points

Regarding the number of data points and their effect on visual judgement or statistical scores, it was found that contrasts with longer phases consistently showed lower scores and contrasts with shorter phases showed higher scores. This was true for visual analysis judgements, as well as Tau-U, IRD, and Baseline Corrected Tau

calculations. The author hypothesized on the reasons for this pattern. The author hypothesized that longer baselines might allow the participants to show their true skills and the internal validity threats might appear in the baseline so that any patterns that will show in the intervention will first appear in the baseline. For example, maturation, other treatments, and uncontrolled events during a long occurring baseline phase will make the scores in the baseline higher. Consequently, the intervention will show a weaker effect. Furthermore, a longer intervention phase might have the effect of lowering the performance of the participants because there will be more opportunities for uncontrolled events to happen that might lower the scores (e.g., loss of interest, getting tired). The author concluded that a longer baseline was much impactful than a longer intervention when it came to controlling internal validity threats. Any decisions regarding the length of the data series ultimately must be made by the researcher considering a variety of factors. Some of these factors include: the severity and the topography of the participants' behaviors and the variability in the data. However, this pattern suggests the need for further research on the impact of number of data points on visual analysis judgements and scores from non-overlap methods.

Overlap

Regarding the presence of overlap and its effect on visual analysis judgement, it was found that agreement between raters was high. The two raters judging each graph agreed on 93% of their judgements. This outcome suggests that overlap is a fairly easy feature to detect by visual analysis compared to other features such as trend. Regarding the effect of overlap on the performance of statistical methods for all contrasts, it was found that the agreement between the three non-overlap methods and visual analysis

decreased over 20% when the contrasts had some degree of overlap. When the graphs had some degree of overlap, Tau-U and IRD made a similar amount of type I (false positives) and type II (false negatives) errors, while Baseline Corrected Tau ability to discriminate between graphs showing effect and not showing effect was characterized by a stronger ability to detect effect but a higher level of mistakes (Type I errors) as well. Despite the larger number of mistakes, Baseline Corrected Tau showed to be more effective than Tau-U and IRD identifying effective interventions

Trend

The author analyzed the agreement between visual analysis judgements and Tau-U, IRD and Baseline Corrected Tau scores regarding effectiveness or lack of effectiveness of the treatments. In general, it was noticed that there was good agreement between visual analysis outcomes and the scores of the three methods in identifying effective contrasts with baseline trend. However, the agreement was much lower when the methods had to identify non-effective contrasts. Baseline Corrected Tau, however, showed the highest agreement in both cases. It was expected that Baseline Corrected Tau and Tau-U should perform better than IRD when trend is present in the baseline. Both methods were developed to account for trend in the baseline. The author documented that Tau-U and Baseline Corrected Tau agreed on 91% of their judgements regarding effect or lack of effect. On the other hand, the agreement between visual analysis and the three non-overlap methods was higher when the contrasts lacked a baseline trend, which was an expected outcome.

Although two non-overlap indices (i.e., Tau-U, Baseline Corrected Tau) are supposed to be very similar in form detecting and correcting trend in the baseline, they

often performed differently when applied to the same set of data containing baseline trend. The author found that almost half the graphs that were corrected for trend by Baseline Corrected Tau did not agree in the final outcomes with the outcomes of Tau-U. So, it is encouraged to use more than one effect size measure when reporting effectiveness of a study's outcomes besides visual analysis. Another reason to suggest the use of a second effect size measure is the fact that baseline trends are not always apparent to the visual analyst (Chiu & Roberts, 2018).

Autocorrelation

Most statistical methods used in single-case designs do no account for autocorrelation and this might increase the number of type I or type II errors. For example, Brossart and colleagues (2006) compared five methods commonly used in single-case studies using the same data. They found that all methods were highly affected by autocorrelation.

The author found more autocorrelation in the intervention phases (849) than in the baselines (556) of the 924 contrasts. The reason for this difference was the fact that 369 contrasts were removed from the analysis of baseline autocorrelation due to errors in the Excel software. Only 76 were removed from the intervention phases. The software produced errors because of an excessive number of data points with the same value. Negative autocorrelation tended to be more concentrated in the baseline phases (61%) as opposed to the intervention phases (28%). However, positive autocorrelation was much more concentrated in the intervention phases (71%) as opposed to the baseline phases (38%). Also, it was established that contrasts with positive autocorrelation had a higher number of data points in both the baseline and the intervention phases than the contrasts

with negative autocorrelation. The analysis of the number of data points per phase related to large, medium, and small autocorrelation indicates that a lower number of data points were related to a large autocorrelation. The contrasts with large positive autocorrelation in the interventions were the only exception. Forty-nine (8%) of these contrasts had an unusually large number of data points (i.e., between 30 and 90). Similarly, the contrasts with medium positive autocorrelation in the interventions had a contrast with 108 data points, an unusually large number. These outliers affected category averages. Overall, longer phases seemed to be correlated with positive and small autocorrelation.

Outcomes show that the presence of both baseline trend and positive autocorrelation in the intervention phase was more frequent than the combination of trend and negative autocorrelation or autocorrelation in the baseline. Furthermore, more than half (51%) of contrasts showed a large autocorrelation. These outcomes suggest that a relationship between baseline trend and autocorrelation in the intervention might exist. No other strong pattern was identified. However, slight differences seemed to indicate that more effect was identified when the baseline had negative correlation and when the intervention had positive correlation. Overall, the larger the magnitude of the autocorrelation, the higher the number of effective judgements. Overall, Tau-U and IRD showed a similar pattern. They both produced lower average scores when the autocorrelation was positive in the baseline but produced higher average scores when the autocorrelation was positive in the intervention phases. On the other hand, Baseline Corrected Tau produced higher average scores when the autocorrelation was negative regardless of baseline or intervention autocorrelation. Regarding the effect of autocorrelation magnitude, a clear pattern emerged. The stronger the magnitude, the

higher the average scores produced by Tau-U, IRD, and Baseline Corrected Tau. The only exception was found on contrasts with a negative baseline autocorrelation. Baseline Corrected Tau was not affected by the differences in magnitude since it produced the same average score for the contrasts with small, medium, and large autocorrelation.

Range of Effective Interventions

The typical range of Tau-U identified by the investigator from interventions deemed effective by visual analysis seemed consistent with Tau-U criteria determined by Parker and Vannest (2009). The typical range for an effective intervention found by the author of the current review was 0.65 to 1.00. Parker and Vannest considered any scores above 0.65 to be effective. Regarding IRD, the author's results (0.65 to 1.00) were less consistent. IRD benchmarks (Parker et al., 2009) indicate that any effect rated as moderate had scores of around .50 to 0.70. Large and very large effects generally received scores of .70 or higher. Consistent with Tarlow's (2016) findings, the author found that Baseline Corrected Tau gave smaller results than Tau-U on 200 (99%) time series and equal results on 2 (1%) time series. The author found that a Baseline Corrected Tau of 0.45 or higher indicated an effective intervention. Considering that the current review included 924 data sets produced by real research studies, the author considers these meaningful outcomes. Comparing the performance of the three non-overlap metrics, the author concluded that Baseline Corrected Tau showed the best overall performance due to its ability to avoid the effects of autocorrelation, trend, and overlap. Baseline Corrected Tau has been developed as an improved extension of Tau index (Parker, Vannest, Davis, & Sauber, 2011).

Limitations of the Study

There were four limitations to this study worth mentioning. One of the limitations is that only lag 1 autocorrelation coefficients for the baseline and treatment phases were calculated and reported. The investigator decided not to detrend the baseline and intervention phases, correct the autocorrelation values for bias, and test the significance of the autocorrelation. Performing these tasks would have required lengthening the study timeline, but performing all the steps would have improved the accuracy of the outcomes. Second, visual analysis IRR was 79% which was a moderate percentage. This was true despite the fact the participants had previously completed graduate level single-case design methodology classes (i.e., between 4 and 16 credits), successfully completed the online training, and attended a workshop delivered by the author. Third, this review of single-case designs excluded alternating treatment designs and combined designs. This decision was made because a number of alternating treatment designs did not include baselines; they were few in number and tended to produce different results than the results obtained using the other designs (Chen et al., 2016). Most of the studies included in this review used multiple-baseline designs or multiple-probe designs. This is not surprising given that school-based interventions often target the acquisition of skills. The acquisition of skills cannot be successfully measured using withdrawal designs that require returning to baseline when the intervention is removed. Finally, the author would have liked to include more effect-size metrics. Including more statistical methods would have benefited the research literature.

Future Directions

The outcomes of this review highlighted some important needs for future research. Matching the outcomes of Ninci and colleagues (2015), visual analysts have been shown to have a moderate degree of interrater reliability (79%) when the graphs included a variety of confounding factors. In this study, visual analysts had difficulty identifying small effects when the graphs had a combination of confounding factors (e.g., trend, variability, small number of data points). Because of this, many of the phase contrasts analyzed by the raters were judged to be ineffective when they effective according to non-overlap estimates. Considering the number of type II errors made by visual analysis, the author recommends the use of non-overlap metrics with a high degree of sensitivity that can detect small effects to complement visual analysis. Research has shown that human judges are not always effective even with contextual information to interpret the graphs (Brossart, Parker, Olson, & Mahadevan, 2006) so most researchers advocate the use of both visual and statistical analysis to inform each other (Brossart et al., 2006; Franklin, Allison, & Gorman, 2014). Furthermore, the investigator recommends using a sensitive metric because higher levels of sensitivity are related to higher levels of agreement with visual analysis.

The author also recommends the use of visual aids or contextual information to make visual judgements more accurate. Examples of visual aids are the dual-criteria (DC) method developed by Fisher, Kelley, & Lomas, (2003) and the SMIQR developed by Manolov, Jamieson, Evans, & Sierra (2015). Finally, the investigator agrees with the recommendation provided by Ninci and colleagues (2015) regarding the advantages of operationally defining the characteristics to be analyzed visually. According to the

researchers, when characteristics are operationally defined, visual analysis reports higher rates of inter-rater agreement (citation). Unfortunately, raters are not often provided with such definitions (citation). One method worth to examine is the use of alternative forms of graphical display. Always, SCDs graphs are displayed with connected line graphs or occasionally with bar graphs. Morales, Domínguez, and Jurado (2001) examined 3000 AB graphs using three different graphical displays, 1000 graphs for each display (i.e., line graphs, bar graphs and Tukey box plots). Interestingly, Tukey box plots was the graphical type that participants made less errors when data were presented. It is worth to test and compare again these mentioned three graphical types and other graphical types for future studies.

Third, the author recommends the use of visual analysis as the primary method of evaluation of intervention effects in order to avoid type I errors. However, it must be kept in mind that visual analysis is not a method without reliability issues. Considering this, the author recommends the adjunctive use of statistical methods following visual analysis suggesting a functional relation. This gated approach of visual analysis followed by statistical analysis is outlined in the pilot What Works Clearinghouse standards for Single-Case Research (Kratochwill et al., 2010). Furthermore, Baseline Corrected Tau and Tau-U did not agree in the final outcomes even though they are supposed to account for baseline trend. So, it is encouraged to use more than one effect size measure when reporting effectiveness of a study's outcomes besides visual analysis. In conclusion, visual analysis should be the primary method of analysis for single-case research and one or two statistical methods, either parametric or non-parametric, should be used to support the findings. Although non-overlap analysis showed to be efficient, simple to calculate,

and agree with visual analysis, more investigation is needed on parametric methods such as Dynamic Multilevel Analysis.

A fourth recommendation relates to factors affecting the outcomes of visual and statistical analysis. Researchers would benefit from investing in further research concerning the impact of features like number of data points, variability, autocorrelation, and trend. The impact of overlap might not need the same investment due to the fact that overlap is easily detected by visual analysis, like the findings of this review supports. The current review concludes that the other confounding factors, such as baseline trend, number of data points, and autocorrelation, do have an impact in both visual and statistical analysis. The author used 924 real data sets. Although this is one of the strengths of the study, future researchers should use artificial sets in order to manipulate the data and better identify any patterns related to the number of data points or the other important factors affecting outcomes. As mentioned in the limitations, further autocorrelation testing is recommended to know the impact of this factor on the statistical analysis outcomes. Regarding the issue of further research on autocorrelation and its impact on analysis, the author recommends to investigate any possible relationships between autocorrelation, the length of the intervention, and both the intervention (independent variable) and the behaviors observed (dependent variables).

Fifth, the author suggests to request another visual analysis judgement by a third expert rater in case of disagreement between visual and statistical analysis. Taking this step before reporting research findings will increase the confidence on the outcomes of published studies.

Sixth, the author recommends investigating and testing a variety of effect size metrics especially those that account for trend, as well as investigate their agreement with visual analysis. The proper way to analyze data from SCDs remains a topic under investigation, with no single approach proving to be superior in every instance (Brossart et al 2014). The metrics suggested are: Extended Celeration Line (ECL; White and Haring, 1980), Percentage of Non-Overlap Corrected Data (PNCD; Manolov & Solanas, 2009), and Percent Exceeding the Median (PEM; Wolery et al, 2010), Ratio of Distance (RD; Carlin & Costello, 2108) and Tau-U (Parker et al., 2011). Most statistical methods used in single-case designs do no account for autocorrelation and this might increase the number of type I or type II errors. Two advanced methods might be a solution to this problem. One is Interrupted Time-Series Analysis (ITSA) but required between 50 and 100 observations per phase and to apply this procedure in SCDs most of the time will might be impossible. The second method is Dynamic Multilevel Analysis (DMA) and can be used with applies to small datasets (Machalicek, & Horner, 2018). Although the three metrics used in this review performed well, the author found Baseline Corrected Tau (Tarlow, 2016), to have better general performance with real data involving confounding factors such as trend and autocorrelation so it is suggested that future research continues to work on testing this effect size metric.

Finally, the author is proposing a new package for visual analysis training. The package includes five important steps all visual analysis raters should follow in order to guarantee a more accurate and reliable judgement. The five steps are the following: Although there is no need for statistical knowledge to conduct visual analysis, raters should first acquire some basic knowledge in statistics (e.g., multilevel analysis, time-

series analysis) before conducting visual analysis. Second, learning about confounding factors (e.g., trend, autocorrelation) that affect the interpretation of visual analysis before conducting visual analysis will contribute to improve inspection of the graphs. Third, raters should acquire knowledge on visual analysis variables (i.e., level, trend, variability, overlap, immediacy of effect, and consistency across similar phases) their definitions, and the rationale of each variable. Fourth, knowledge of visual aids that can help detect effect when confounding factors are present will increase reliability and minimize type I and II errors. Fifth, completing the Visual Analysis Training Program using singlecase.org website before conducting visual analysis will improve visual analysis judgment.

Conclusion

This study utilized the outcomes of 132 single-case studies evaluating the effectiveness of interventions for students with ASD in school settings to compare the performance of visual analysis and three effect size metrics (Tau-U, IRD, and Baseline Corrected Tau). The investigator examined a total of 924 phase contrasts from the sample of 132 studies. The findings of this study showed that the three non-overlap methods and visual analysis made both type I and type II errors. The three statistical methods had a high degree of sensitivity, agreement with visual analysis and performed well detecting effective interventions but had a tendency to make type I errors. On the other hand, visual analysis has shown to be less sensitive and make more type II errors than type I.

The high sensitivity of the methods can be beneficial to identify interventions that show small effect visual analysis might not be able to detect. With the intention to make the three metrics as sensitive as possible, the investigator decided to use the lowest cutoff scores to answer the study questions because they have the highest level of sensitivity.

After calculating Kappa using the lowest cutoff score, it was found that the three non-overlap methods had a moderate to substantial level of agreement with visual analysis. In general, researchers agree that minimizing type I errors is more important in the field of education. For example, failing to report an effective intervention can be less damaging than reporting an ineffective intervention as effective. Considering these facts, the author recommended the use of visual analysis as the primary method of evaluation of intervention effects and a minimum of two statistical methods as secondary methods in order to help identify interventions that might be difficult to be detected by a more conservative visual analysis. These procedures can increase confidence on the outcomes of published studies.

After finding that visual analysts struggle to detect effects on the graphs, a possible reason for the conservative judgements were confounding factors present in all graphs which made visual analysis difficult (i.e., overlap, trend, small number of data points, variability, more than one variable line in the graph). In addition, the effect shown on the contrasts was often small and difficult to see. Regarding the number of data points and their effect on visual judgement or statistical scores, it was found that contrasts with longer phases consistently showed lower scores and contrasts with shorter phases showed higher scores. This was true for visual analysis judgements, as well as Tau-U, IRD, and Baseline Corrected Tau calculations. Regarding the effect of overlap on the performance of statistical methods for all contrasts, it was found that the agreement between the three non-overlap methods and visual analysis decreased over 20% when the contrasts had some degree of overlap. In general, it was noticed that there was good agreement between visual analysis outcomes and the scores of the three methods in identifying effective

contrasts with baseline trend. However, the agreement was much lower when the methods had to identify non-effective contrasts. These outcomes suggest that a relationship between baseline trend and autocorrelation in the intervention might exist. No other strong pattern was identified. However, slight differences seemed to indicate that more effect was identified when the baseline had negative correlation and when the intervention had positive correlation. Overall, the larger the magnitude of the autocorrelation, the higher the number of effective judgements.

Considering that the current review included 924 data sets produced by real research studies, the author considered these outcomes to be very meaningful so he mentioned six important implications for future research.

REFERENCES CITED

Allison, D. B., Franklin, R. D., & Heshka, S. (1992). Reflections on visual inspection, response guided experimentation, and Type I error rate in single-case designs. The Journal of Experimental Education, 61(1), 45-51.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case∗. *Behaviour Research and Therapy*, *31*(6), 621-631.

Alresheed, F., Machalicek, W., Sanford, A., & Bano, C. (2018). Academic and related skills interventions for autism: A 20-year systematic review of single-case research. *Review Journal of Autism and Developmental Disorders,* 1-16.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Angermeier, K., Schlosser, R. W., Luiselli, J. K., Harrington, C. & Carter, B. (2008). Effects of iconicity on requesting with the Picture Exchange Communication System in children with autism spectrum disorder. *Research in Autism Spectrum Disorders, 2*(3), 430-446.

Armstrong, T. K. & Hughes, M. T. (2012). Exploring computer and storybook interventions for children with high functioning autism. *International Journal of Special Education*, *27*(3), 88-99.

Ayres, K. M., Mechling, L. & Sansosti, F. J. (2013). The use of mobile technologies to assist with life skills/independence of students with moderate/severe intellectual disability and/or autism spectrum disorders: Considerations for the future of school psychology. *Psychology in the Schools, 50(*3), 259-271.

Bagatell, N., Mirigliani, G., Patterson, C., Reyes, Y. & Test, L. (2010). Effectiveness of therapy ball chairs on classroom participation in children with autism spectrum disorders. *American Journal of Occupational Therapy*, *64*(6), 895-903.

Borenstein, M., Cooper, H., Hedges, L. & Valentine, J. (2009). Effect sizes for continuous data. *The handbook of research synthesis and meta-analysis*, *2*, 221-235.

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*(5), 531-563.

Brossart, D. F., Vannest, K. J., Davis, J. L. & Patience, M. A. (2014). Incorporating non-overlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychological Rehabilitation, 24*(3-4), 464-491.

Callahan, K., Henson, R. K. & Cowan, A. K. (2008). Social validation of evidence-based practices in autism by parents, teachers, and administrators. *Journal of Autism and Developmental Disorders, 38*(4), 678-692.

Carlin, M. T., & Costello, M. S. (2018). Development of a Distance-Based Effect Size Metric for Single-Case Research: Ratio of Distances. *Behavior Therapy*.

Carnahan, C. A. & Williamson, P. S. (2013). Does compare-contrast text structure help students with autism spectrum disorder comprehend science text? *Exceptional Children, 79*(3), 347-363.

Carter, E. W., Sisco, L. G., Chung, Y. & Stanton-Chapman, T. L. (2010). Peer interactions of students with intellectual disabilities and/or autism: A map of the intervention literature. *Research & Practice for Persons with Severe Disabilities, 35*, 36–79.

Centers for Disease Control and Prevention. (2010). *Nutrition*. Retrieved from http://www.cdc.gov/nutrition/

Centers for Disease Control and Prevention. (2014). Prevalence of Autism Spectrum Disorder among children aged 8 years. *Morbidity and Mortality Weekly Report*. Retrieved from http://www.cdc.gov/mmwr/

Chen, M., Hyppa-Martin, J. K., Reichle, J. E. & Symons, F. J. (2016). Comparing single case design overlap-based effect size metrics from studies examining speech generating device interventions. *American Journal on Intellectual and Developmental Disabilities, 121*(3), 169-193.

Cicero, F. R. & Pfadt, A. (2002). Investigation of a reinforcement-based toilet training procedure for children with autism. *Research in Developmental Disabilities*, 23(5), 319-331.

Cihak, D. F. & Foust, J. L. (2008). Comparing number lines and touch points to teach addition facts to students with autism. *Focus on Autism and Other Developmental Disabilities, 23*, 131-137.

Cihak, D. F., Kildare, L. K., Smith, C. C., McMahon, D. D. & Quinn-Brown, L. (2012). Using video social stories™ to increase task engagement for middle school students with autism spectrum disorders. *Behavior Modification, 36*(3), 399-425

Cihak, D. F. & Schrader, L. (2008). Does the model matter? Comparing video self-modeling and video adult modeling for task acquisition and maintenance by adolescents with autism spectrum disorders. *Journal of Special Education Technology, 23*(3), 9-20.

Cihak, D. F., Smith, C. C., Cornett, A. & Coleman, M. B. (2012). The use of video modeling with the picture exchange communication system to increase independent communicative initiations in preschoolers with autism and developmental delays. *Focus on Autism and Other Developmental Disabilities, 27*(1), 3-11.

Cihak, D. F., Wright, R. & Ayres, K. M. (2010). Use of self-modeling static-picture prompts via a handheld computer to facilitate self-monitoring in the general education classroom. *Education and Training in Autism and Developmental Disabilities*, 136-149.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

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

Cohen, R. J. & Swerdlik, M. E. (2005). *Psychological testing and assessment: An introduction to tests and measurement*. New York, NY: McGraw-Hill.

Couper, L., Van der Meer, L., Schäfer, M. C., McKenzie, E., McLay, L., O'Reilly, M. F. & Sutherland, D. (2014). Comparing acquisition of and preference for manual signs, picture exchange, and speech-generating devices in nine children with autism spectrum disorder. *Developmental Neurorehabilitation, 17*(2), 99-109.

Coleman-Martin, M. B., Heller, K. W., Cihak, D. F. & Irvine, K. L. (2005). Using computer-assisted instruction and the nonverbal reading approach to teach word identification. *Focus on Autism and Other Developmental Disabilities, 20*(2), 80-90

Council for Exceptional Children Working Group. (2014). Council for Exceptional Children: Standards for evidence-based practices in special education. *TEACHING Exceptional Children, 46*(6), 206–212. http://doi.org/10.1177/0040059914531389

Crombie, I. K. (1996). *The pocket guide to critical appraisal.* London, United Kingdom: BMJ Publishing Group.

Delano, M. E. (2007). Use of strategy instruction to improve the story writing skills of a student with Asperger syndrome. *Focus on Autism and Other Developmental Disabilities, 22*(4), 252-258.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (2014). Design and analysis of single-case research. Psychology Press.

Fentress, G. M. & Lerman, D. C. (2012). A comparison of two prompting procedures for teaching basic skills to children with autism. *Research in Autism Spectrum Disorders, 6*(3), 1083-1090.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36*(3), 387-406.

Fletcher, D., Boon, R. T. & Cihak, D. F. (2010). Effects of the TouchMath program compared to a number line strategy to teach addition facts to middle school students with moderate intellectual disabilities. *Education and Training in Autism and Developmental Disabilities*, *45,* 449-458.

Ganz, J. B., Boles, M. B., Goodwyn, F. D. & Flores, M. M. (2014). Efficacy of handheld electronic visual supports to enhance vocabulary in children with ASD. *Focus on Autism and Other Developmental Disabilities, 29*(1), 3-12.

Ghaziuddin, M., Ghaziuddin, N., & Greden, J. (2002). Depression in persons with autism: Implications for research and clinical care. *Journal of Autism and Developmental Disorders, 32*(4), 299-306

Gunby, K. V., Carr, J. E. & LeBlanc, L. A. (2010). Teaching abduction-prevention skills to children with autism. *Journal of Applied Behavior Analysis, 43*(1), 107-112.

Hartley, S. L., Sikora, D. M., & McCoy, R. (2008). Prevalence and risk factors of maladaptive behaviour in young children with autistic disorder. *Journal of Intellectual Disability Research, 52*(10), 819-829.

Harvey, S. T., Boer, D., Meyer, L. H., & Evans, I. M. (2009). Updating a meta-analysis of intervention research with challenging behaviour: Treatment validity and standards of practice. *Journal of Intellectual and Developmental Disability, 34*(1), 67-80.

Hedges, L. V., Pustejovsky, J. E. & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*(3), 224-239.

Hedges, L. V., Pustejovsky, J. E. & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, *4*(4), 324-341

Hetzroni, O. E. & Ne'eman, A. (2013). Influence of colour on acquisition and generalisation of graphic symbols. *Journal of Intellectual Disability Research, 57*(7), 669-680.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165-179.

Horner, R. H., Carr, E. G., Strain, P. S., Todd, A. W., & Reed, H. K. (2002). Problem behavior interventions for young children with autism: A research synthesis. *Journal of Autism and Developmental Disorders*, 32, 423–446.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).

Jang, J. & Matson, J. L. (2015). Autism severity as a predictor of comorbid conditions. *Journal of Developmental and Physical Disabilities, 27*(3), 405-415.

Johnston, S. S., Buchanan, S. & Davenport, L. (2009). Comparison of fixed and gradual array when teaching sound-letter correspondence to two children with autism who use AAC. *Augmentative and Alternative Communication, 25*(2), 136-144.

Kagohara, D., van der Meer, L., Achmadi, D., Green, V., O'Reilly, M., Lancioni, G. & Sigafoos, J. (2012). Teaching picture naming to two adolescents with autism spectrum disorders using systematic instruction and speech-generating devices. *Research in Autism Spectrum Disorders, 6*, 1224–1233.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings*. Oxford University Press.

Kazdin A.E. Single case research designs: Methods for clinical and applied settings. New York: Oxford; 1982.

Kendall, M. G. (1962). *Rank correlation methods* (3rd ed.). New York, NY: Hafner.

Kennedy, C. H. (2005). Single-case designs for educational research. Pearson/Allyn & Bacon.

Kinnealey, M., Pfeiffer, B., Miller, J., Roan, C., Shoener, R. & Ellner, M. L. (2012). Effect of classroom modification on attention and engagement of students with autism or dyspraxia. *American Journal of Occupational Therapy, 66*(5), 511-519.

Kittelman, A., Gion, C., Horner, R. H., Levin, J. R., & Kratochwill, T. R. (2018). Establishing Journalistic Standards for the Publication of Negative Results. *Remedial and Special Education, 39*(3), 171-176.

Koegel, L., Matos-Freden, R., Lang, R. & Koegel, R. (2012). Interventions for children with autism spectrum disorders in inclusive school settings. *Cognitive and Behavioral Practice, 19*(3), 401-412.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*(1), 26-38.

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. What works clearinghouse.

Kurt, O. & Tekin-Iftar, E. (2008). A comparison of constant time delay and simultaneous prompting within embedded instruction on teaching leisure skills to children with autism. *Topics in Early Childhood Special Education, 28(*1), 53-64.

Lane, J. D. & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation, 24*(3-4), 445-463.

Lenroot RK, Yeung PK (2013): Heterogeneity within autism spectrum disorders: what have we learned from neuroimaging studies? *Frontiers in Human Neuroscience. 7.*

Levitt, P. & Campbell, D. B. (2009). The genetic and neurobiologic compass points toward common signaling dysfunctions in autism spectrum disorders. *The Journal of Clinical Investigation, 119*(4), 747-754.

Logan, L. R., Hickman, R. R., Harris, S. R., & Heriza, C. B. (2008). Single-subject research design: Recommendations for levels of evidence and quality rating. *Developmental Medicine & Child Neurology, 50,* 99-103.

Lorah, E. R., Tincani, M., Dodge, J., Gilroy, S., Hickey, A. & Hantula, D. (2013). Evaluating picture exchange and the iPad™ as a speech generating device to teach communication to young children with autism. *Journal of Developmental and Physical Disabilities, 25*(6), 637-649.

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.

Machalicek, W., & Horner, R. H., (2018) Special issue on advances in single-case research design and analysis, *Developmental Neurorehabilitation, 21*:4, 209-211, DOI: 10.1080/17518423.2018.1468600

Maggin, D. M., & Chafouleas, S. M. (2010). PASS-RQ: Protocol for assessing single-subject research quality. Unpublished research instrument.

Mannion, A., Leader, G. & Healy, O. (2013). An investigation of comorbid psychological disorders, sleep problems, gastrointestinal symptoms and epilepsy in children and adolescents with autism spectrum disorder. *Research in Autism Spectrum Disorders, 7*, 35–42.

Manolov, R., Jamieson, M., Evans, J. J., & Sierra, V. (2015). Probability and visual aids for assessing intervention effectiveness in single-case designs: A field test. *Behavior Modification, 39*(5), 691-720.

Manolov, R., & Solanas, A. (2009). Percentage of non-overlapping corrected data. *Behavior Research Methods, 41*, 1262-1271.

Marcus, A. & Wilder, D. A. (2009). A comparison of peer video modeling and self-video modeling to teach textual responses in children with autism. *Journal of Applied Behavior Analysis, 42*(2), 335-341.

Mason, R. A., Ganz, J. B., Parker, R. I., Burke, M. D. & Camargo, S. P. (2012). Moderating factors of video-modeling with other as model: A meta-analysis of single-case studies. *Research in Developmental Disabilities, 33*(4), 1076-1086.

Matson, J. L., Rieske, R. D. & Williams, L. W. (2013). The relationship between autism spectrum disorders and attention-deficit/hyperactivity disorder: an overview. *Research in Developmental Disabilities, 34*(9), 2475-2484

Matson, J. L., & Williams, L. W. (2013). Differential diagnosis and comorbidity: Distinguishing autism from other mental health issues. *Neuropsychiatry, 3*(2), 233-243.

Mazurek, M. O., Kanne, S. M. & Wodka, E. L. (2013). Physical aggression in children and adolescents with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 455–465.

Morales, M., Domínguez, M. L., & Jurado, T. (2001). The influence of graphic techniques in the evaluation of the effectiveness of treatment in time-series design. *Quality and Quantity, 35*(3), 277-290.

Mucchetti, C. A. (2013). Adapted shared reading at school for minimally verbal students with autism. *Autism, 17*(3), 358-372.

Muller, M. P., Tomlinson, G., Marrie, T. J., Tang, P., McGeer, A., Low, D. E., ... & Gold, W. L. (2005). Can routine laboratory tests discriminate between severe acute respiratory syndrome and other causes of community-acquired pneumonia? *Clinical infectious diseases, 40*(8), 1079-1086.

National Professional Development Center on Autism Spectrum Disorders. (2009). *Evidence-based practices for children and youth with ASD*. Retrieved June 14, 2011 from http://autismpdc.fpg.unc.edu/sites/autismpdc.fpg.unc.edu/files/EBP_Update_Reviewer_Training_printversion.pdf.

Nakagawa, S. Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews* 82, 591–605.

Ninci, J., Vannest, K. J., Willson, V., & Zhang, N. (2015). Interrater agreement between visual analysts of single-case data: A meta-analysis. *Behavior Modification*, 39(4), 510-541.

Odom, S. L. (2009). The tie that binds: Evidence-based practice, implementation science, and outcomes for children. *Topics in Early Childhood Special Education, 29*(1), 53-61.

Ottenbacher, K. J. (1993). Interrater agreement of visual analysis in single-subject decisions: Quantitative review and analysis. *American Journal on Mental Retardation*.

Parker, R. I. & Hagan-Burke, S. (2007). Median-based overlap analysis for single case data. *Behavior Modification, 31*(6), 919–936.

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

Parker, R. I. & Vannest, K. (2009). An improved effect size for single-case research: Non-overlap of all pairs. *Behavior Therapy, 40*(4), 357-367.

Parker, R. I., & Vannest, S., (2007). *Pairwise data overlap for single case research.* Unpublished manuscript.

Parker, R. I., Vannest, K. J. & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children, 75*(2), 135-150.

Parker, R. I., Vannest, K. J. & Davis, J. L. (2011). Effect size in single-case research: A review of nine non-overlap techniques. *Behavior Modification, 35*(4), 303–322.

Parker, R. I., Vannest, K. J., Davis, J. L. & Sauber, S. B. (2011). Combining non-overlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*(2), 284-299.

Palmen, A., Didden, R. & Arts, M. (2008). Improving question asking in high-functioning adolescents with autism spectrum disorders: Effectiveness of small-group training. *Autism*, *12*(1), 83-98.

Paterson, C. R. & Arco, L. (2007). Using video modeling for generalizing toy play in children with autism. *Behavior Modification, 31(*5), 660-681.

Pennington, R. C. (2010). Computer-assisted instruction for teaching academic skills to students with autism spectrum disorders: A review of literature. *Focus on Autism and Other Developmental Disabilities, 25*, 239–248.

Petersen-Brown, S., Karich, A. C. & Symons, F. J. (2012). Examining estimates of effect using non-overlap of all pairs in multiple baseline studies of academic intervention. *Journal of Behavioral Education, 21*(3), 203-216.

Pierucci, J. M., Barber, A. B., Gilpin, A. T., Crisler, M. E. & Klinger, L. G. (2015). Play assessments and developmental skills in young children with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities, 30*(1), 35-43.

Polychronis, S., McDonnell, J., Johnson, J., Riesen, T. & Jameson, M. (2004). A comparison of two trial distribution schedules in embedded instruction. *Focus on Autism and Other Developmental Disabilities, 19*(3), 140-151.

Rakap, S. (2015). Effect sizes as result interpretation aids in single-subject experimental research: Description and application of four non-overlap methods. *British Journal of Special Education, 42*(1), 11-33.

Rakap, S., Snyder, P. & Pasia, C. (2014). Comparison of non-overlap methods for identifying treatment effect in single-subject experimental research. *Behavioral Disorders, 39*(3), 128-145.

Ray, D. C. (2015). Single-case research design and analysis: Counseling applications. *Journal of Counseling & Development, 93*, 394–402.

Reagon, K. A., Higbee, T. S. & Endicott, K. (2006). Teaching pretend play skills to a student with autism using video modeling with a sibling as model and play partner. *Education and Treatment of Children*, 517-528.

Reichow, B. (2011). Development, procedures, and application of the evaluative method for determining evidence-based practices in autism. In Reichow et al. (Eds), *Evidence-based practices and treatments for children with autism* (pp. 25-39). Springer US.

Reichow, B., Doehring, P., Cicchetti, D. V. & Volkmar, F. R. (Eds.). (2010). *Evidence-based practices and treatments for children with autism*. Springer Science & Business Media.

Reichow, B., Volkmar, F., & Cicchetti, D. (2008). Development of the evaluative method for evaluating and determining evidence-based practices in autism. *Journal of Autism and Developmental Disorders, 38,* 1311–1319. doi:10.1007/ s10803-007-0517-7

Reisener, C. D., Lancaster, A. L., McMullin, W. A. & Ho, T. (2014). A preliminary investigation of evidence-based interventions to increase oral reading fluency in children with autism. *Journal of Applied School Psychology*, *30*(1), 50-67.

Riesen, T., McDonnell, J., Johnson, J., Polychronis, S. & Jameson, M. (2003). A comparison of constant time delay and simultaneous prompting within embedded instruction in general education classes with students with moderate to severe disabilities. *Journal of Behavioral Education, 12*(4), 241-259

Riley-Tillman, T.C. *&* Burns, M.K. *(2009). Evaluating educational interventions: Single case design for measuring response to intervention.* New York*:* Guilford.

Sancho, K., Sidener, T. M., Reeve, S. A. & Sidener, D. W. (2010). Two variations of video modeling interventions for teaching play skills to children with autism. *Education and Treatment of Children, 33*(3), 421-442.

Schlosser, R. W. (2011). *EVIDAAC Single-Subject Scale*. Retrieved from http://www.evidaac.com/ratings/Single_Sub_ Scale.pdf

Schlosser, R. W., Raghavendra, P., Sigafoos, J., Eysenbach, G., Blackstone, S., & Dowden, P. (2008). EVIDAAC Systematic Review Scale. *Unpublished manuscript, Northeastern University, Boston, MA.*

Schlosser, R. W., Sigafoos, J., & Belfiore, P. (2009). *EVIDAAC Comparative Single-Subject Experimental Design Scale (CSSEDARS).* Retrieved from http://www.evidaac.com/rat- ings/CSSEDARS.pdf

Schneider, N., Goldstein, H., & Parker, R. (2008). Social skills interventions for children with autism: A meta-analytic application of percentage of all non-overlapping data (PAND). *Evidence-Based Communication Assessment and Intervention*, *2*(3), 152-162.

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

Shadish, W. R. (2014). Statistical analyses of single-case designs: The shape of things to come. *Current Directions in Psychological Science*, *23*(2), 139–146.

Shadish, W. R., Hedges, L. V., Horner, R. H. & Odom, S. L. (2015). The role of between-case effect size in conducting, interpreting, and summarizing single-case research (NCER 2015-002). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Shillingsburg, M. A., Powell, N. M. & Bowen, C. N. (2013). Teaching children with autism spectrum disorders to mand for the removal of stimuli that prevent access to preferred items. *The Analysis of Verbal Behavior, 29*(1), 51.

Sigafoos, J., Arthur-Kelly, M. & Butterfield, N. (2006). Enhancing everyday communication for children with disabilities. Baltimore, MD: Paul H Brookes Publishing Co.

Silk, J. (1992). Silk Scientific [UN-SCAN-IT]. Utah: Orem

Simeonsson, R. J., & Bailey Jr, D. B. (1991). Evaluating programme impact: Levels of certainty. In *Early intervention studies for young children with special needs* (pp. 280-296). Springer US.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*, 510–550.

Smith, T., Scahill, L., Dawson, G., Guthrie, D., Lord, C., Odom, S., . . .Wagner, A. (2007). Designing research studies on psychosocial interventions in autism. *Journal of Autism and Developmental Disorders, 37*, 354–366.

Smith, V., Jelen, M., & Patterson, S. (2009). Video modeling to improve play skills in a child with autism: A procedure to examine single-subject experimental research. *Evidence- Based Practice Briefs, 4*, 1–13.

Spencer, V. G., Evmenova, A. S., Boon, R. T. & Hayes-Harris, L. (2014). Review of research-based interventions for students with autism spectrum disorders in content area instruction: Implications and considerations for classroom practice. *Education and Training in Autism and Developmental Disabilities, 49*, 331–353.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*(4857), 1285-1293.

Tarlow, K. R. (2016). An Improved Rank Correlation Effect Size Statistic for Single-Case Designs: Baseline Corrected Tau. *Behavior Modification*, 0145445516676750.

Task Force on Evidence-Based Interventions in School Psychology. (2003). *Procedural and coding manual for review of evidence-based interventions*. Retrieved from http://www.indiana.edu/~ebi/documents/_workingfiles/ EBImanual1.pdf

Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single- subject designs and n-of-1 trials: Introducing the single-case experimental design (SCED) scale. *Neuropsychological Rehabilitation, 18,* 385-401.

Taylor, B. A., Hughes, C. E., Richard, E., Hoch, H. & Coello, A. R. (2004). Teaching teenagers with autism to seek assistance when lost. *Journal of Applied Behavior Analysis, 37*(1), 79-82.

Tekin-Iftar, E. & Birkan, B. (2010). Small group instruction for students with autism general case training and observational learning. *The Journal of Special Education, 44*(1), 50-63.

Thiemann, K. S., & Goldstein, H. (2001). Social stories, written text cues, and video feedback: Effects on social communication of children with autism. *Journal of Applied Behavior Analysis, 34*(4), 425-446.

Tincani, M. (2004). Comparing the picture exchange communication system and sign language training for children with autism. *Focus on Autism and other Developmental Disabilities, 19*(3), 152-163.

U.S. Department of Education, National Center for Education Statistics. (2012). *Digest of education statistics, 2011* (NCES 2012-001), Chapter 2.

Valentine, J. C., Tanner-Smith, E. E., Pustejovsky, J. E. & Lau, T. S. (2016). Between-case standardized mean difference effect sizes for single-case designs: a primer

Van der Meer, L., Kagohara, D., Roche, L., Sutherland, D., Balandin, S., Green, V. A. &. Sigafoos, J. (2013). Teaching multi-step requesting and social communication to two children with autism spectrum disorders with three AAC options. *AAC: Augmentative and Alternative Communication, 29*(3), 222-234.

Van Steensel, F. J., Bögels, S. M. & Perrin, S. (2011). Anxiety disorders in children and adolescents with autistic spectrum disorders: A meta-analysis. *Clinical Child and Family Psychology Review, 14*(3), 302-317.

VanMeter, L., Fein, D., Morris, R., Waterhouse, L., & Allen, D. (1997). Delay versus deviance in autistic social behavior. *Journal of Autism and Developmental Disorders, 27,* 557–569.

Vannest, K. J. & Ninci, J. (2015). Evaluating Intervention Effects in Single-Case Research Designs. *Journal of Counseling & Development*, *93*(4), 403-411.

Volkmar, F., Carter, A., Grossman, J., & Klin, A. (1997). Social development in autism. In D. Co- hen & F. Volkmar (Eds.), *Handbook of autism and pervasive developmental disorders* (pp. 173–194). New York: Wiley.

Wang, S.-Y., & Parrila, R. (2008). Quality indicators for single-case research on social skill interventions for children with autistic spectrum disorder. *Developmental Disabilities Bulletin, 36*, 81–105.

Wendt, O. (2009, May). Calculating effect sizes for single-subject experimental designs. Paper presented at the Ninth Annual International Campbell Collaboration Colloquium, Oslo, Norway.

White, O. R., & Haring, N. G. (1980). *Exceptional teaching: A multimedia training package.* Columbus, OH: Merrill.

Wilson, K. P. (2013). Teaching social-communication skills to preschoolers with autism: Efficacy of video versus in vivo modeling in the classroom. *Journal of Autism and Developmental Disorders, 43*(8), 1819-1831.

Wolery, M., Busick, M., Reichow, B. & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education, 44*(1), 18-28.

Wolfe, K., Seaman, M. A., & Drasgow, E. (2016). Interrater agreement on the visual analysis of individual tiers and functional relations in multiple baseline designs. *Behavior Modification, 40*(6), 852-873.

Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fettig, A., Kucharczyk, S., … Schultz, T. R. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of Autism and Developmental Disorders, 45*(7), 1951-1966.

Ximenes, V. M., Manolov, R., Solanas, A., & Quera, V. (2009). Factors affecting visual inference in single-case designs. *The Spanish Journal of Psychology, 12*(2), 823-832.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*, 561–577.