

Abstract

ROSS, SARAH GWEN. Measuring Response to Intervention: Comparing Three Effect Size Calculation Techniques for Single-Case Design Analysis. (Under the direction of Dr. John Begeny.)

Response to intervention (RTI) is increasingly being used in educational settings to make high-stakes, special education decisions. Because of this, the accurate use and analysis of single-case designs to monitor intervention effectiveness has become important to the RTI process. Effect size methods for single-case designs provide a useful way to analyze single-case data; however, little research exists that compares the various types of effect size methods. This study compared three competing single-case effect size methods: (a) the regression-based, Allison-MT method (Allison & Gorman, 1993); (b) the new, non-overlap-based, Tau-U method (Parker, Vannest, Davis, & Sauber, 2011); and (c) the standardized mean difference-based, Busk-Serlin-Model 2 (Busk & Serlin, 1992). Using previously published AB datasets that measure the Words read Correct per Minute (WCPM) variable, these three methods were compared using a repeated-measures ANOVA to determine if overall differences existed between obtained effect size scores when analyzing the same data. Additionally, the range of effect size scores obtained from each method and the effect of removing lag-1 autocorrelation from the original data were examined. Results indicated that the Tau-U method was least affected by autocorrelation, however, there was no significant difference between scores obtained by the Allison-MT and Tau-U methods after the removal of autocorrelation. With autocorrelation removed, effect size scores appear to fit within interpretation guidelines for group designs. Implications of these findings for research and practice are discussed.

Measuring Response to Intervention: Comparing Three Effect Size Calculation Techniques
for Single-Case Design Analysis

by
Sarah Gwen Ross

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina

2012

APPROVED BY:

John Begeny, Ph.D.
Chair of Advisory Committee

Ann Schulte, Ph.D.

Edward Sabornie, Ph.D.

Scott Stage, Ph.D.

Dedication

To my parents for their wisdom, patience, and unconditional love throughout my life.

“Someday, the light will shine like a sun through my skin and they will say, What have you done with your life?, and though there are many moments I think I will remember, in the end, I will be proud to say, I was one of us.” Brian Andreas

Biography

Sarah Gwen Ross grew up in Reidsville, North Carolina. After graduating from the International Baccalaureate program at Reidsville High School in 2001, Sarah attended Wake Forest University where she majored in Psychology with a minor in Biology. Sarah was a member of several research teams while attending Wake Forest, and upon graduating with a Bachelor of Arts degree in 2005, she worked as a research coordinator for a NIH-funded grant on personality theory. Sarah entered the School Psychology program at North Carolina State University in August 2006 and completed her Master of Science degree in December 2009. She completed her predoctoral child psychology internship at the Texas Child Study Center/Dell Children's Hospital in Austin, Texas. Upon completing her Doctor of Philosophy degree in August 2012, Sarah will begin a Harvard Medical School post-doctoral fellowship at the Dana Farber Cancer Institute in Boston, Massachusetts.

Acknowledgments

I would like to thank the amazing faculty whose ideas and critiques made the writing of my dissertation possible. John Begeny, thank you for helping me to remain passionate about my dissertation topic and its roots in social justice and community outreach. Your intelligence, positivity, creativity, and patience are a model for us all, and I am truly honored to have had you as an advisor. Ann Schulte, thank you for keeping me on my toes. Having your approval of my work means more to me than you know. Scott Stage, thank you for your amazing statistical knowledge and ability to turn my defense into an intriguing philosophical discussion. Ed Sabornie, thank you for bringing in an educational perspective and helping me to shift my focus from statistics to real-world application.

I would also like to thank my friends for their support along the way. When I think of all of the words of encouragement, meetings, wine nights, coffee shops, hugs, patience, study breaks, emails, and g-chats, I feel so incredibly grateful.

Finally, I would like to thank my family. The love and support you have given me is overwhelming. To my parents, thank you for giving me a unique and kind perspective on the world. You are in my heart always.

Table of Contents

List of Tables	viii
List of Figures	ix
Chapter 1	1
Introduction	1
Response to Intervention and Single-Case Design in Schools.....	2
Curriculum-Based Measurement for Reading (CBM-R)	3
CBM-R validity	4
CBM-R reliability	4
CBM-R standard error of measurement	5
Single-Case Designs.....	5
AB designs.....	6
Multiple Baseline designs.....	6
Visual Analysis Techniques for Single-case Designs	7
Reliability of visual analysis.....	8
Effect Sizes and Single-Case Designs.....	10
The problem of autocorrelation and single-case designs.....	11
Categories of Single-Case Effect Sizes.....	13
Regression methods	13
Percentage of non-overlap methods	16
Standardized mean difference methods	20
Summary of Previous Research with Effect Sizes for Single-case Design Data	21
Purpose	23
Hypotheses	24
Do obtained effect sizes from the three methods differ significantly from each other? .	24
Do effect size scores change significantly when autocorrelation is removed from the data?.....	24

Do obtained effect size scores from the three methods differ significantly from each other when lag-1 autocorrelation is removed from the data?	25
What are the typical effect size ranges for each effect size method when evaluating AB designs that utilized the WCPM dependent variable?	25
Chapter 2.....	26
Method	26
Overview	26
Selection of Published Data	26
Graph inclusion criteria	28
Data Extraction.....	28
Data Analysis	29
Effect Size Analyses.....	29
Statistical Differences between Effect Size Estimates	31
Autocorrelation Removal	32
Score Distributions.....	32
Chapter 3.....	33
Results	33
Differences Between Effect Sizes	33
Differences Between Scores Produced by the Same Effect Size Method, Before and After Autocorrelation Removal	33
Differences Between Effect Sizes After Lag-1 Autocorrelation Removal	34
Descriptive Statistics.....	34
Chapter 4.....	35
Discussion	35
Differences Between Effect Size Scores	36
Differences Between Scores Produced by the Same Effect Size Method, Before and After Autocorrelation Removal	38
Differences Between Effect Size Scores After Lag-1 Autocorrelation Removal	40
Limitations and Future Directions.....	41

Overall Conclusions and Implications	43
References	45
Appendices.....	61
Appendix A	62
Appendix B	66

List of Tables

Table 1. Means of effect size estimates before and after autocorrelation removal..... 58

List of Figures

Figure 1. Box-Plot Display of Effect Size Ranges..... 57

Chapter 1

Introduction

Accurately monitoring students' response to intervention is an important task in research and practice. In practical settings, such as schools, response to intervention (RTI) is now being used to help make high-stakes decisions such as special-education placement (Riley-Tillman & Burns, 2009). Because practitioners are often faced with the task of monitoring students' intervention progress, single-case designs are useful for determining intervention effectiveness. As such, educators and researchers must be able to accurately interpret data gathered as part of single-case design analytic strategies. A variety of single-case effect size calculation techniques have evidence of being a valid way to analyze data (Brossart, Parker, Olson, & Lakshmi, 2006; Campbell, 2004; Ma, 2006; Manolov & Solanas, 2008; Manolov, Solanas, & Leiva, 2010; Parker & Brossart, 2003; Parker, Cryer, & Burns, 2006; Parker, Vannest, Davis, & Sauber, 2011).

The goal of this study was to compare three competing single-case design effect-size methods by analyzing 66 previously published AB designs. The three methods included: (a) the regression-based Allison-MT model (Allison & Gorman, 1993); (b) the new, non-overlap plus trend method, Tau-U (Parker, et al., 2011); and (c) a standardized mean difference model, or Busk-Serlin-Model 2 (Busk & Serlin, 1992). Because reading is an area commonly targeted for intervention in schools, and because words read correctly per minute (WCPM) data are commonly used to monitor students' reading progress, reading intervention studies that measured fluency using the WCPM dependent variable were identified and used in this study.

Response to Intervention and Single-Case Design in Schools

Response-to-Intervention (RTI) is an educational problem-solving model with “the primary goal of providing the most effective instruction and intervention to each student through the efficient allocation of educational resources” (Riley-Tillman & Burns, 2009, p. 3). Part of this process involves determining special education eligibility according to a student’s specific response to intervention. RTI was originally developed as part of the data-based decision-making movement fueled by Deno and Mirkin’s (1977) Data-Based Program Modification (DBPM) model. The DBPM model emphasized the individualization of education plans for students with learning and/or behavior problems and included the implementation and monitoring of interventions developed to meet the needs of those students. The RTI model was later included as part of the 2004 reauthorization of the Individuals with Disabilities Education Improvement Act (IDEIA). IDEIA states that an education agency “may use a process that determines if the child responds to scientific, research-based intervention as part of the evaluation procedures” (P.L. No. 108-446 § 614 [b][6][A]; § 614 [b][2&3]). The RTI, problem-solving process also has relevance for the No Child Left Behind (NCLB) Act of 2001, which requires that academic skills and progress of all children be measured and monitored.

Single-case designs provide a way for practitioners to accurately monitor and analyze individual student behavior. These designs have become especially critical in the RTI paradigm because of its emphasis on progress monitoring and individual student response (Riley-Tillman & Burns, 2009). Within RTI, data collected and analyzed using single-case

designs are regularly used to make high-stakes, special education decisions (Riley-Tillman & Burns).

Additionally, school psychologists and other educators may use single-case designs in experimental research aimed to evaluate the effects of one or more school-based interventions. Bliss, Skinner, Hautau, and Carroll (2008) reviewed all articles published in four school psychology journals between the years 2000-2005 and found that single-case designs represented 55% of all “causal-experimental” designs (i.e., studies that involved the manipulation of an independent variable), while group designs encompassed 45%. The authors noted that single case designs provide a logical way for school psychologists to merge practice with research.

In both research and practice, curriculum-based measurement (CBM) is a commonly used tool that can quickly and frequently be used to measure students’ academic achievement over time. CBM has also proven to be one of the most common methods for collecting data that is used in RTI and related problem-solving approaches, particularly CBM for reading (CBM-R; Burns & Gibbons, 2008).

Curriculum-Based Measurement for Reading (CBM-R)

CBM-R involves having a student read connected text aloud for one minute and then determining the number of WCPM. A comprehensive guide to administering and scoring CBM-R probes can be found on the Dynamic Indicators of Basic Literary Skills (DIBELS) website at <https://dibels.uoregon.edu/measures/orf.php#admin>.

CBM-R validity. Many researchers have examined validity estimates for the CBM-R technique (Bain & Garlock, 1992; Good, Simmons, & Kameenui 2001; Hintze & Silbergitt, 2005; Hosp & Fuchs, 2005; McGlinchey & Hixson 2004; Stage & Jacobsen, 2001). For example, Hintze and Silbergitt (2005) tracked 1,766 students longitudinally from first to third grade using the CBM-R measure. At the end of third grade, students were given the Minnesota Comprehensive Assessment (MCA). Longitudinal CBM-R scores were correlated with the 3rd grade MCA scores at a level of .58 for 1st grade spring scores, .68 for 2nd grade spring scores, and .69 for 3rd grade spring scores. Bain and Garlock (1992) examined the association of CBM-R scores with scores from the Comprehensive Test of Basic Skills (CTBS) according to grade level and found correlations of .62 for first grade passages, .79 for second grade passages, and .72 for third grade passages. Additionally, Marston (1989) summarized seven studies which evaluated the concurrent validity of CBM-R scores and found correlations in the moderate to high range (.70 to .90).

CBM-R reliability. CBM-R has also been found to be sufficiently reliable (Hintze & Shapiro, 1997; Hintze, Shapiro & Lutz, 1994; Hintze & Silbergitt, 2005; Howe & Shinn, 2002; Poncy, Skinner, & Axtell, 2005). Hintze & Silbergitt (2005), found alternate-form reliability estimates for CBM-R first to third grade passages to range from .80 to .91, while Poncy, Skinner, & Axtell (2005) found alternate-form estimates for third grade students to range from .81 to .99. Howe & Shinn (2002) reported mean reliability estimates for CBM-R passages across three probes for first through fifth grades ranging from .82 to .90. Finally, interscorer reliability for the CBM-R technique usually exceeds .90 (Graney & Shinn, 2005).

CBM-R standard error of measurement. Although appropriate levels of validity and reliability have been documented for the CBM-R technique, the presence of measurement error is an important limitation to consider when using WCPM data. Christ and Silberglitt (2007) published standard error of measurement (SEM) data for the WCPM variable. In that study, 8,200 first through fifth grade students were administered three CBM-R reading probes consecutively in the fall, winter, and spring, from 1996-2004. The researchers used median and standard deviation scores obtained from the three readings to determine the number of WCPM that can be attributed to error each time the CBM-R technique is administered. Christ and Silberglitt organized SEM reports according to grade (1st to 5th), homogeneity of scores (homogeneous, typical, and heterogeneous), and magnitude of reliability score (.89-.97). The SEM estimates for first grade samples ranged from 4 to 12 WCPM error, second grade ranged from 5 to 14 WCPM error, third and fourth grades ranged from 6 to 14 WCPM error, and fifth grade ranged from 6 to 15 WCPM error. The authors state that overall SEM estimates are likely between 5 and 9 WCPM across grades.

Single-Case Designs

Broadly speaking, single-case designs (sometimes referred to as small-N designs) are employed to monitor the most basic experimental analysis of behavior: documenting an organism's response to a change in environment (Richards, Taylor, Ramasamy, & Richards (1998). In using these designs, behavior should be monitored before and after the environmental change to ensure that changes in behavior can accurately be attributed to

changes in the environment (Richards, et al.). Single-case designs are useful in research and practice to evaluate the effects of particular interventions with an individual (Riley-Tillman & Burns, 2009). The six most common single-case designs are: AB-designs, Reversal designs, Extended ABA designs, Alternating Treatments designs, Multiple Baseline designs, and Changing Criterion designs (Richards, et al. 1999). Given the methodology of the present study, AB and Multiple Baseline designs are described briefly below, whereas an overview of all six designs are described in Appendix A.

AB designs. An AB design involves repeated measurement of behavior throughout a baseline (A) and intervention (B) phase of experimentation (Barlow, Nock, & Hersen, 2009). The A phase involves documentation of the natural rate of the chosen behavior, while the rate of behavior during an intervention is documented in phase B. Intervention effectiveness can be tentatively assumed if there is a desired change in behavior from the A to B phase. However, researchers and practitioners should be cautious when using or interpreting AB designs because it is difficult to distinguish between changes that occur as a result of intervention from changes that could have occurred regardless of phase change.

Multiple Baseline designs. Multiple Baseline designs involve the combination of several AB designs (with A representing a baseline phase and B representing an intervention phase), either from multiple participants, multiple behaviors within a single participant, or multiple settings within a single participant (Morgan & Morgan, 2009). The unique component of Multiple Baseline designs is that each AB design has a baseline of varying lengths. Intervention effectiveness is determined if there is a desired change from the baseline to intervention phase for all participants, regardless of length of baseline condition.

Multiple Baseline designs can be useful when it is undesirable or unethical to withdraw an intervention already in place and to avoid potential intervention carry-over effects. This design may be especially valuable for practitioners interested in monitoring the effects of an intervention on multiple students over time.

Visual Analysis Techniques for Single-case Designs

When single-case design data are presented graphically, several characteristics of the data should be examined and combined in order to determine if intervention effects are present. These characteristics include: level, latency of change, variability, and trend.

One should compare the *level*, or the mean or median, of the data during the baseline phase to that of the intervention phase (Morgan & Morgan, 2009). If there is a clear difference between the levels of the baseline and intervention phases, one can make preliminary judgments that the intervention was effective. *Latency of change* refers to how quickly behavior changes after a new phase is introduced (e.g., from baseline to intervention or from intervention 1 to intervention 2). Although it is not necessary for behavior to change immediately with the onset of intervention, the more quickly a change in behavior occurs, the more confident one can be that the behavior change is due to the new phase (Riley-Tillman & Burns, 2009).

Variability refers to the amount of variation in range and/or to consistency in a set of data (Riley-Tillman & Burns, 2009). Variability within a phase can be examined visually and is sometimes expressed by giving a high-low range of scores. Variability can be examined within and between phases. If behavior is highly variable within an intervention phase there

is a strong likelihood that the behavior is not under the control of an in-place intervention, or there is the possibility that the behavior is not properly defined or is being measured inaccurately (Morgan & Morgan, 2009).

In behavioral research, the goal of an intervention may sometimes be to decrease the amount of variability in an individual's behavior. For example, if a student displays highly variable behavior in the baseline phase, the goal of the intervention may be to decrease the variability of the student's behavior. Therefore, a decrease in variability from baseline to intervention may mean that the intervention is responsible for the change in stability of the target behavior.

Finally, *trend* is the tendency for the data to show systematic increases or decreases over time (Kazdin, 2010). If there is an increase in the trend of the data from baseline to intervention, one can tentatively conclude that positive results of the intervention are being displayed (if, in fact, the goal is to increase a particular behavior). However, it is important to be certain that a trend in the intervention phase is not merely a continuation of a positive trend in the baseline phase (Morgan & Morgan, 2009).

Reliability of visual analysis. Although visual analysis of single-case designs can be useful in monitoring student behavior, many researchers across time have objected to using it as the sole method of single-case design analysis (DeProspero & Cohen, 1979; Edgington, 1984; Houle, 2009; Kazdin, 2010). Some researchers suggest that visual analysis is highly subjective and influenced by biases and differences in individuals (Morgan & Morgan, 2009). They argue that there are no true standards for making judgments about treatment effects (Wolery & Harris, 1982), leading to inconsistencies in interpretation (Brossart, et al.,

2006; DeProspero & Cohen, 1979; Furlong & Wampold, 1982, 1981; Harbst, Ottenbacher, & Harris, 1991; Ottenbacher, 1990; Ottenbacher & Cusick, 1991). For example, DeProspero and Cohen (1979) asked 250 article reviewers to visually analyze 36 reversal designs for level of experimental control (i.e., the degree to which there was a significant change in behavior between each phase of the design). They found the mean interrater agreement to be .61, which represents moderate agreement.

In another study, Brossart and colleagues (2006) recruited 15 graduate students in educational psychology to examine thirty-five computer-generated AB graphs depicting a fictional elementary student's level of cooperation over time. The graphs differed in variability, mean level, trend, and gap between baseline endpoint and intervention start-point. The graduate students were asked to rate how convinced they were that there had been an improvement in behavior due to intervention. Individual rater-to-group correlations across the thirty-five graphs averaged .58.

Important to the reliability of visual analysis in school-based settings is teacher background in single-case designs and, more generally, using graphical representations of data. Begeny and Martens (2006) examined the extent of training that masters level elementary, secondary, and special education university students had in assessment techniques, such as using single-case designs. A total of 97% of elementary, 86% of secondary, and 86 % of special education masters students had little to no training in single-case designs. Additionally, 84% of elementary, 94% of secondary, and 78% of special education students had little to no experience with making instructional decisions using a graphical chart.

Overall, if special education decisions are being made based on single-case designs displayed graphically, practitioners must be conscious of the fact that educators may have little to no experience in the area, and that even those with training and experience, interpretations of data may meaningfully differ between educators. Because visual analysis may sometimes be inconsistent or inaccurate, effect size measurements are beginning to gain popularity as a supplement to visual interpretation of single-case data. Additionally, according to the Institute of Educational Sciences Single Case Design Standards, effect size analysis should be completed when visual analysis suggests moderate to strong effects (Kratochwill, et al., 2010).

Effect Sizes and Single-Case Designs

The American Psychological Association (2010) suggests that all manuscripts submitted for publication include effect size calculations to aid in interpretation of intervention outcome. An APA Board of Scientific Affairs report states that "...reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses" (Wilkinson and Task Force on Statistical Inference, 1999, p. 599). Also, several recent articles have promoted the use of effect size measurements with single-case designs to monitor an individual's response to an intervention (Brossart, et al., 2006; Campbell, 2004; Ma, 2006; Manolov & Solanas, 2008; Manolov, et al., 2010; Parker & Brossart, 2003; Parker, et al., 2006; Parker, et al., 2011) and to facilitate meta-analyses of single-case

research (Bonner & Barnett, 2004; Busse, Kratochwill, & Elliott, 1995; Jensen, Clark, Kircher, & Kristjansson, 2007).

Brossart and colleagues (2006) list three major advantages of using effect sizes for group designs. First, effect sizes provide an index of the strength of association between an intervention and behavior, allowing for inferences to be made regarding how much of the behavior can be explained by the intervention (Rosnow & Rosenthal, 1989). Second, effect sizes provide a continuous versus a dichotomous index of treatment success, which supports the incremental change of treatment components instead of simpler “continue or discontinue” decisions. Third, effect sizes are not systematically affected by sample size, so strong effects may be found even with relatively small samples. Additionally, effect sizes allow for the calculation of confidence intervals, which provide even more information about the dependability of results (Kirk, 1996). Parker and Hagan-Burke (2007) add that effect sizes used with single-case designs (a) improve measurement precision when results are not large and obvious; (b) provide a means of comparing intervention success across single-case studies, either at a local level or in a more elaborate meta-analysis; and (c) provide an objective summary of results when visual judgments do not agree.

The problem of autocorrelation and single-case designs. Before discussing the varieties of effect size calculation techniques available for use with single-case research, it is important to discuss the concept of serial dependence (displayed by the presence of autocorrelation) and its effect on single-case data. Statistical serial dependence is present when the future of a variable is predictable to some degree from its own values or from the passage of time (Matyas & Greenwood, 1996). Serial dependence can be quantified by

examining the presence of autocorrelation in a data series. Autocorrelation is the degree to which values of an observed behavior at time t are correlated with values at time $t - i$ (Matyas & Greenwood). Lag-1 autocorrelations, give i a value of 1, and therefore examine the relationship of a behavior at a particular time with the behavior measured the time before. The presence of positive autocorrelation is a problem for most parametric and non-parametric statistics because of the resulting decrease in the standard error estimate and increase in the test statistic, leading to increased Type I (false positive) error rates and effect sizes (Brossart et al., 2006). Scheff (1959) showed that an autocorrelation of .30 can lead to Type I error rate increases from 5% to 12%, and an autocorrelation of .20 can lead to an increase from 5% to 10%.

As reported by Parker and colleagues (2011) a variety of studies have found that one-third or more of published datasets from single-case designs are autocorrelated to an undesirable degree (>.20 to .25; Matyas & Greenwood, 1996; Parker, et al., 2005; Sharpley & Alavosius, 1988; Suen & Ary, 1989), and it is therefore appropriate to control for or “remove” autocorrelation when using this type of data. Several authors have examined the extent to which single-case effect size scores change (or remain the same) with the removal of autocorrelation (Brossart et al., 2006; Manolov, et al., 2010; Manolov & Solanas, 2008; Parker & Brossart, 2003; Parker, et al. 2007; Parker et al., 2011), and findings for each effect size method will be discussed later.

Categories of Single-Case Effect Sizes

The majority of effect size calculations that have been studied for use with single-case designs (a) are regression-based, (b) measure the percentage of non-overlapping data, or (c) use standardized mean difference methods. Standardized mean difference methods are most similar to effect size calculations used for group designs (e.g., Cohen's d) and are described in detail by Busk & Serlin (1992).

Regression methods. Most statistical developments for single-case research are based in regression models (Brossart, et al. 2006). In a review of many available analytic techniques for single-case data, Faith and colleagues (1996) concluded that regression approaches are the best available, though “no gold standard” (p. 253). Regression-based effect sizes are based on R^2 or η^2 (R^2 applied to categorical predictors) (Brossart, et al., 2006), and Cohen (1988, p. 22) offers at least four ways to interpret these types of effects: (a) “the proportion of a client’s score variance explained by phase differences, (b) the reduction in uncertainty (percent increase in prediction ability) due to phase differences, (c) the percent of nonoverlap of client scores between phases, and (d) the percent of scores of one phase exceeded by the upper half of the scores of the other phase.” The original Cohen (1988) effect size interpretations included categories for large ($R^2 = .25$), medium ($R^2 = .09$), and small ($R^2 = .01$) effects. However, single-case design effect sizes tend to be much larger and depend on a variety of variables including: (a) formulae used to calculate (e.g., simple versus multiple regression), (b) type of predictor (e.g., continuous versus categorical), and (c) whether data trend is being controlled (Brossart et al., 2006). R^2 and η^2 effect sizes may also change according to context. For example, scores may differ according to (a) the reliability

and sensitivity-to-growth of the outcome measures, and their administration schedule; (b) the participant's response levels; and (c) the nature, intensity, and timetable of the intervention (Mitchell & Hartmann, 1981).

There are several regression-based effect size calculation techniques that have been studied for use with single-case designs over the past 30 years including: (a) binomial test on extended Phase A baseline (BINOM; White & Haring, 1980); (b) Last Treatment Day (LTD; White, Rusch, Kazdin, & Hartmann, 1989); (c) Gorsuch's trend effect size (GORSUCH; Faith et al., 1996; Gorsuch, 1983); (d) Center's mean plus trend difference (CENTER-MT), which can also be calculated for mean differences only (Center, Skiba, & Casey, 1985-1986); and (e) Allison's mean plus trend difference (ALLISON-MT), which can also be calculated for mean differences only (Allison & Gorman, 1993). Brossart and colleagues (2006) provide a useful overview of each of these methods. Crosbie's ITSACORR (Crosbie, 1993) was also used for several years, but has been found to have little relationship with other methods or visual analysis (Brossart, et al., 2006; Parker et al., 2011; Parker & Brossart, 2003). Crosbie officially retired the model after Huitema (2004) identified fatal flaws in its analysis (Southerly, 2006).

In general, the Allison-MT method (Allison & Gorman, 1993) is the newest of the regression techniques and addresses flaws that were present in previous models. It has repeatedly outperformed other regression-based techniques and for many it is currently the regression method chosen when comparing regression methods to other calculation techniques (Brossart et al. 2006; Campbell, 2004; Manolov & Solanas, 2008; Parker et al., 2005; Parker, et al., 2003).

The Allison-MT model predicts the effects of an intervention while controlling for trend present in the baseline phase, if a positive trend exists (Allison & Gorman, 1993). Controlling for baseline trend prevents one from mistakenly concluding that an intervention is effective when in actuality a pre-existing trend in baseline is being allowed to continue. However, this model does not control for trend in the intervention phase, as intervention-phase trend should be attributed to the in-place intervention. This is an improvement from other models (e.g., Center-MT) that controlled for trend in both baseline and intervention phases.

Brossart and colleagues (2006) evaluated the use of five effect size methods (listed above) and visual analysis with 35 experimenter-developed AB designs. Each of the five statistics was converted to an R^2 effect size. R^2 scores varied greatly across calculation methods, with mean effect sizes for very effective interventions (as judged by visual analysis) ranging from 0.03 (GORSUCH) to 0.90 (LTD). The Allison-MT method yielded mean R^2 scores from .65 for graphs judged to show no effects of intervention to .87 for graphs judged to show very effective interventions. The Allison-MT method was best matched to the judgments of visual analysis. The authors concluded that no previous effect size interpretation guidelines from group research were appropriate for use with single-case designs and that future research is needed in the area.

Parker and Brossart (2003) investigated the same five effect size methods as the previous study (as well as the mean-only methods of the Allison and Center models) and examined power, effect of autocorrelation, range of effect sizes, and intercorrelation between methods. This study used 50 constructed AB datasets representing a range of type and degree

of intervention effects. The authors found that the Allison-MT method had the strongest level of power, almost no effect of autocorrelated data (along with Center-MT), and was clustered with each of the other methods. Additionally, this method (along with LTD) produced the largest effect sizes.

Parker and colleagues (2005) completed a comparable study using 77 previously published AB graphs. The articles from which the graphs were chosen examined a range of dependent variables, including time off-task and other disruptive behavior, depressive symptoms, and delusional thoughts. None of the articles examined the WCPM dependent variable. Similar to previous findings, the Allison-MT model had strong levels of power, was least affected by autocorrelation, and was most intercorrelated with the other methods.

Percentage of non-overlap methods. The original Scruggs, Mastropieri, and Casto (1987) Percentage of Non-Overlapping Data (PND) procedure involves determining the percentage of data points in Phase B (usually the intervention phase) that exceed the most extreme data point in Phase A (usually the baseline phase). Scruggs and Mastropieri (1994) developed guidelines for interpreting PND scores, stating that if using a conservative interpretive framework, scores greater than 70% indicate effective interventions, between 50 and 70% indicate questionably effective interventions, and below 50% indicate that the intervention is ineffective. Scruggs and Mastropieri (2001) completed a review of single-case meta-analyses and found that of the 15 meta-analyses identified, two-thirds used the PND method to describe results, whereas only two used regression-based methods. It should be noted that none of these meta-analyses examined reading fluency interventions or the WCPM variable.

Although, there are supporters of the original PND method, based mainly on its intuitiveness and ease of calculation (Campbell, 2004), there are several inherent flaws to this method that have led some researchers to abandon it altogether. Allison and Gorman (1993) describe three situations in which the PND statistic will likely be inaccurate. First, if even one outlier is present in the baseline phase, the PND score may change significantly. Therefore, the original PND method may be a poor choice for conditions and behaviors that contain high variability, such as classrooms and social behavior (Jenson, Clark, Kircher, & Kristjansson, 2007), or with WCPM data which has been shown to have a fair amount of variability in scores (Christ & Silberglitt, 2007). Second, outliers in the treatment phase can lead to a small positive PND score, even if the treatment has had a generally detrimental effect. Finally, the PND score can lead to errors in judgment if there is a trend present in the data. Therefore, a positive PND score can be obtained if the treatment has no effect but only allows a pre-existing trend to continue. Additionally, a ceiling effect is present when using PND, meaning that it lacks discrimination ability between very successful interventions as it nears 100%. (Parker, Hagan-Burke, & Vannest, 2007).

Several new percentage of non-overlap methods have been introduced in recent years with goals of offering solutions to the problems present in the original PND method. The Percentage of All Non-Overlapping Data (PAND) method (Parker et al., 2007) takes into account all data points and counts the minimum number of measurements that need to be removed in order to obtain a series with no overlap. Because this method uses all data points, outliers do not carry as much weight as with the PND method. Additionally, PAND can be translated into Pearson's *Phi*, which allows for calculations of *p* values and confidence

intervals. An important limitation of the PAND method is that it does not offer solutions to trend or ceiling effects present in the PND method. In one study, Parker and colleagues (2007) utilized a convenience sample of 75 multiple baseline graphs (using only the first AB comparison of each graphs for analysis) to compare the PAND and PND methods. A variety of dependent variables were utilized in the articles from which the graphs were obtained. For the most successful interventions, PAND and PND yielded similar results. However, results varied greatly for less effective interventions, with the least effective interventions earning a PAND score of 50% (representing chance-level overlap between phases) and a PND score of 0 to 10%.

Ma's (2006) Percentage of Data Points Exceeding the Median (PEM) method calculates the percentage of treatment measurements greater than the baseline phase median score. This method corrects the ceiling effect present in the PND and PAND methods, but does not address the effect of trend in these methods. Ma compared the PEM and PND methods by judging them against original authors' visual analysis determinations from 61 previously published single-case design graphs focusing on self-control training. The PEM method was found to have a slightly stronger correlation with the original authors' judgment (.57) than the PND method (.49). Chen and Ma (2007) completed a study with similar methods and again found that the PEM method was more highly correlated (.68) with original authors' judgments than PND (.60) when examining interventions for disruptive behavior. However, neither of these studies utilized significance tests to determine whether the correlation associated with the PEM method was significantly higher.

Manolov and colleagues (2010) compared the performance of the PND, PAND, and PEM methods when using experimenter-developed data sets with varying levels of autocorrelation present. The PEM method was found to yield basically the same results regardless of the presence of autocorrelation. Autocorrelation was associated with higher effect size estimates for the PAND and PND methods, with the PND being the more affected of the two. Additionally, the PND and PEM method were found to better discriminate between effective and non-effective interventions than the PAND method, though the PND yielded smaller overall effect sizes than the PEM method. The PAND method outperformed the PEM method only when the baseline phase is considerably shorter than the treatment phase. In general, all three methods were affected by the presence of trend in the data, leading the authors to suggest the need for visual inspection of data prior to applying any of the three procedures.

Parker and colleagues (2011) recently developed a non-overlap method that accounts for data trend. The Tau-U method combines non-overlap between phases with trend from within the intervention phase, while controlling for trend in the baseline phase. Tau-U is derived from Kendall's Rank Correlation and the Mann-Whitney U-Test between groups. The authors tested the method on a sample of 176 published AB data series (obtained from a convenience sample of single-case research monitoring a variety of target behaviors) and found that the inclusion of trend in the intervention phase led to more modest results than simple non-overlap. Additionally, the authors found that controlling for baseline trend led to only a modest decrease in non-overlap score. The Tau-U score is not affected by the ceiling effect present in other non-overlap methods, and performs well in the presence of

autocorrelation. Of the 72 datasets for which lag-1 autocorrelation was removed, 75% showed only minor changes to Tau-U score, while 25% of the Tau-U scores changed substantially. The authors suggested future exploration of this method (Parker et al., 2011).

Standardized mean difference methods. Busk and Serlin (1992) introduced an effect size method for single-case designs that is based on the Glass (1976) model for group designs. This effect size method obtains the ratio of the difference between two sample means (usually the baseline and intervention) to a standard deviation measure. Busk and Serlin suggest that this method can be calculated using three approaches, each differing in the assumptions made about the data. This study will focus on the second approach, but it is useful to mention the assumptions made by the first and third models. The first model makes no assumptions concerning population distribution or equality of intermeasure variances and covariance, while the third approach assumes a normal distribution and equality of variances and intercorrelations across baseline and treatment phases (Busk & Serlin).

The second approach (referred to as the Busk-Serlin-Model 2 in this report) assumes equality of variances across baseline and treatment phases and involves dividing the mean difference of the baseline and intervention phases by the pooled standard deviation from both phases. This procedure was first recommended for use with single-case designs by White and colleagues (1989), and gives an output similar to Cohen's *d*.

Manolov & Solanas (2008) compared Busk and Serlin's Model 1 and 2 with PND and three regression-based methods: Gorsuch's (1983) trend analysis, White and colleagues (1989) LTD method, and Allison & Gorman's (1993) Allison-MT method, all discussed above. The authors used computer-generated AB designs with varying series lengths,

treatment effects, and levels of autocorrelation. All methods except for PND were converted to η^2 to aid in comparison among calculation techniques. The Gorsuch results were not discussed because they resulted in extremely low effect size scores (0.01 to 0.06). The authors state that PND and both the Busk and Serlin models were less affected by autocorrelation than the regression-based methods. PND and the Busk-Serlin Models 1 and 2 were also better able to differentiate between effective and non-effective interventions than the regression-techniques, although they were affected by general trend in the data not related to the treatment's introduction. In general, the η^2 values obtained by the Busk-Serlin Models 1 and 2 were much smaller than the values of the regression-based methods. The authors recommend the use of the Busk-Serlin models and PND when examining effect sizes in single-case designs.

Summary of Previous Research with Effect Sizes for Single-case Design Data

Compared to other regression-based methods, the Allison-MT (Allison & Gorman, 1993) method has been shown to most closely match with visual analysis judgments of treatment effects (Brossart, et al. 2006), have the highest level of power (Parker & Brossart, 2003; Parker, et al., 2005), be least affected by autocorrelation (Parker & Brossart, 2003; Parker, et al., 2005), and be most intercorrelated with the other methods (Parker & Brossart, 2003; Parker, et al., 2005).

Several non-overlap methods have been introduced to address the flaws present in the traditional PND method (Scruggs et al., 1987), including the PAND (Parker, et al. 2007) and PEM (Ma, 2006) methods. When the PND, PEM, and PAND methods were compared, the

PEM method was found to be least affected by autocorrelation and also to better discriminate between effective and non-effective interventions (Manolov, et al. 201). However, none of the three methods were immune to the effects of baseline trend in the data, which allows for positive scores to be obtained when no treatment effect is present. To address the problem of baseline trend, Parker and colleagues (2011) developed the Tau-U non-overlap method. Although preliminary research shows that the Tau-U method minimizes the effects of autocorrelation (Parker et al.), no studies have compared the Tau-U calculation method to other single-case effect size methods that have been discussed in the literature.

Finally, the Busk-Serlin-Model 2 for single-case designs (Busk & Serlin, 1992) is most similar to traditional effect size calculations for group designs. The Busk-Serlin-Model 2 (in addition to PND) was found to be less affected by autocorrelation and better able to discriminate between effective and non-effective interventions than the Allison-MT method (Manolov & Solanas, 2008).

Though several studies have examined single-case effect size methods, there are still important gaps in the literature that should be addressed. First, although one study compared the Allison-MT and the Busk-Serlin-Model 2 methods using simulated data (Manolov & Solanas, 2008), the study did not use data obtained from actual participants, and neither of the two methods have been compared to the new, non-overlap Tau-U method. As such, using Allison-MT, Busk-Serlin-Model 2, and Tau-U methods to analyze published data should inform the field by investigating the presence of significant differences between the effect sizes obtained from the three methods as well as examining “real world” data—similar to that which would be obtained and used by educators in applied settings. Second, each previously

published study has only used descriptive statistics (typically through box plots and descriptions of data ranges) or intercorrelation methods to describe effect size ranges. To extend the research that compares different effect size formulae, inferential statistics (e.g., ANOVAs) could be used to explore possible statistically significant differences between the scores obtained from the effect size formulae. Finally, no previous single-case effect size studies have specifically examined the WCPM dependent variable, which is an important limitation because this variable is frequently used in schools and as part of the RTI process (Burns & Gibbons, 2008). By examining the WCPM variable in the context of differing effect size methods, one can begin to make estimates of typical effect size ranges that are specific to the WCPM variable.

Purpose

The purpose of this study was to compare the most promising effect size calculation methods for single-case AB designs that utilized the WCPM variable, and in doing so attempt to address the aforementioned gaps in this area of research. The three methods that were compared included the ALLISON-MT method (from the regression-based models; Allison & Gorman, 1993), the new Tau-U method (from the non-overlap models; Parker, et al. 2011), and the Busk-Serlin-Model 2 (from standardized mean difference models; Busk & Serlin (1992). This study aimed to answer four main research questions: (a) Do the obtained effect sizes from the three methods differ significantly from each other? (b) Do effect size scores change significantly within a particular method when autocorrelation is removed from the data? (c) Do the obtained effect sizes from the three methods differ significantly from

each other when lag-1 autocorrelation is removed from the data? (d) What is the typical score range for each effect size method when evaluating AB designs that utilized the WCPM dependent variable?

Hypotheses

Do obtained effect sizes from the three methods differ significantly from each other? Although each of the three effect size estimates aim to describe the same phenomena, it was hypothesized that each of the three calculation techniques will differ significantly from the other. Although there are no published studies that test for differences between scores, based on theory, it was hypothesized that scores obtained using the Busk-Serlin-Model 2 method (converted to η^2) would be significantly larger than those obtained using the Allison-MT and Tau-U calculation techniques. This hypothesis is proposed because both of the latter methods integrate more conservative techniques to estimate effect sizes (i.e., they both control for baseline trend which, when positive, artificially inflates the effect size score). It was also hypothesized that the Allison-MT method will be significantly larger than the Tau-U method. This hypothesis was based mainly on the theory that non-parametric tests, such as Tau-U, make no assumptions about data being analyzed and are therefore more suited for small sample sizes, like the AB designs used in this study. Allison-MT is regression-based, and would therefore be predicted to artificially inflate scores from small sample sizes.

Do effect size scores change significantly when autocorrelation is removed from the data? To ensure that autocorrelation was removed in the same way for each effect size method, this study removed the effects of lag-1 autocorrelation from original scores. This is

different from previously published studies that removed autocorrelation from the residuals of the Allison-MT method (Parker & Brossart, 2003; Parker et al., 2005) and the Busk-Serlin-Model 2 method (Manolov & Solanas, 2008), as it can be calculated using a one-predictor regression. Because the Tau-U method is non-parametric and cannot be calculated using a regression-based method (therefore yielding no residuals from which to remove autocorrelation), it was decided to remove autocorrelation (i.e., “cleanse” the data) from original data and recalculate effect sizes using the cleansed original data. It was hypothesized that effect size scores obtained after lag-1 autocorrelation removal from the original data will be significantly smaller than those obtained before autocorrelation removal. This hypothesis was based on data illustrating that one-third or more of published datasets from single-case designs are autocorrelated to an undesirable degree (.20 to .25) (Parker, et al., 2011), implying that scores will decrease significantly when autocorrelation is removed.

Do obtained effect size scores from the three methods differ significantly from each other when lag-1 autocorrelation is removed from the data? In line with the hypotheses from the first research question, it was predicted that the Busk-Serlin-Model 2 method would produce significantly larger effect sizes than the Allison-MT and Tau-U methods. Also, it was predicted that the Allison-MT model would produce significantly larger effect size scores than the Tau-U method.

What are the typical effect size ranges for each effect size method when evaluating AB designs that utilized the WCPM dependent variable? Because there are no previous studies examining effect sizes using only WCPM data, this research question is exploratory.

Chapter 2

Method

Overview

This study began by collecting all reading intervention studies that measured reading fluency using the WCPM variable and were published between 1995 and 2010. Of the identified studies, those that followed a multiple-baseline design were identified and divided into several AB graphs. The data from the graphs that fit criteria for inclusion in this study were extracted and analyzed using each effect size method. The effect size estimates from the multiple methods were then compared to one another using a repeated measure ANOVA.

Selection of Published Data

The AB datasets that were used in this study were obtained from 66 graphs within 23 published articles from 15 different peer-reviewed journals. Each article was published between the years 1995 and 2010. Using the database systems of *PsychINFO* and *ERIC*, a computer search of the literature published between 1995 and 2010 was conducted using words and combinations of key words that would likely yield reading intervention research articles. The 26 words and word combinations included: reading aloud; oral reading; prosody; reading fluency; automaticity; reading speed; reading rate; reading practice; reading accuracy; repeated reading; fluency with modeling; paired reading; assisted reading; fluency with peer tutoring; fluency with peer assisted; fluency with peer assisted learning strategies; fluency with previewing; fluency with paired reading; fluency with passage preview; fluency with Read Naturally; fluency with Reader's Theatre; reading automaticity; fluency with fast

start; fluency with phrase drill; fluency with the fluency assessment system; fluency with everybody reads. Key words of known fluency-based reading programs were also included (e.g., Read Naturally, Great Leaps).

Articles published between 1995 and 2006 were obtained as part of a larger literature review project that focused on reading fluency interventions. The inclusion criteria for the larger literature review project are included in Appendix B. The abstract of each article was reviewed to determine article eligibility. If some abstracts did not offer enough information, the full article was obtained and reviewed. Because search engines such as *PsycINFO* and *ERIC* sometimes do not locate all relevant articles, search procedures continued by reviewing the Introduction, Discussion, and References sections of the identified articles that meet the inclusion criteria, as well as the Reference sections of previously published literature reviews that have been completed in the area of reading fluency (Chard, Vaughn, & Tyler, 2002; Kuhn & Stahl, 2003; Meyer & Felton, 1999; Morgan & Sideridis, 2006; NICHD, 2000; Therrien, 2004). A total of 216 articles met the criteria for the larger project. The articles that utilized a Multiple Baseline Design across individuals were then selected, identifying a total of 34 articles.

Appropriate articles published between 2007 and 2010 were obtained by searching each of the search terms listed above along with the term, “baseline.” These searches were also completed in the *PsycInfo* and *ERIC* databases, using the same criteria described in Appendix B, while also stipulating the use of a multiple baseline design. This search identified 14 additional articles. The 48 identified articles (34 from the larger literature review project that searched between 1995 and 2006, and 14 from the search of 2007-2010)

were then examined to determine if the WCPM dependent variable was used within each Multiple Baseline design. This excluded 16 of the 48 articles; bringing the total number of articles to 32. The AB graphs that comprised each of the Multiple Baseline designs within each of the 32 articles were then examined to determine eligibility for this study using the graph inclusion criteria below.

Graph inclusion criteria. Using guidelines set forth by Parker and colleagues (2005), graphs were required to display a minimum number of 6 data points per A and B phase, and at least 14 data points across the two phases. Additionally, only graphs that were determined by the published author to show positive intervention effects were included. This inclusion criterion was decided upon because there were only 9 graphs that fit criteria for inclusion and also were described by authors as showing no effect or negative intervention effects, making the use of group comparison statistics inappropriate. A total of 66 graphs, within 23 journal articles met the criteria for use in this study. The journal articles with graphs included in this study are listed in the References section and denoted with an asterisk.

Data Extraction

Published graphs were saved as individual image (jpeg) files and uploaded using *GrabIt XP 10* data extraction software (DataTrend Softward, 2000). Using this program, individual datapoints were extracted from the graphs, resulting in a spreadsheet of the original data from each graph. Because the *GrabIt XP 10* software identifies values up to the 10th decimal place, values were rounded to the nearest whole number (e.g.,

94.0206399527626 was rounded to 94). Whole numbers were chosen for use because the original WCPM data used in the articles were presented in whole numbers, and this study aimed to match the data used in these previous studies exactly. Reliability of graph extraction was checked on approximately 50% of graphs by comparing the values obtained from extraction to visual analysis of each graph to make certain that the extracted values approximated those from the graphs. In all cases, the data extraction method was accurate.

Data Analysis

This study required six steps in data analysis: (a) calculating effect sizes for each data set using the three described effect size methods; (b) using a repeated-measures ANOVA to identify possible differences between the scores from the three effect size methods; (c) removing lag-1 autocorrelation from the original data and recalculating each effect size; (d) using a repeated-measures ANOVA to identify possible differences between scores from the same effect size method before and after autocorrelation is removed; (e) using repeated-measures ANOVAs to identify possible differences between the scores from the three effect sizes once autocorrelation was removed; and (f) illustrating each effect size range, before and after autocorrelation-removal, using box plots. Each of these steps of analysis is described below.

Effect Size Analyses

The three effect size methods used for analysis include, (a) the regression-based method, ALLISON-MT (Allison & Gorman, 1993), (b) the non-overlap with trend method,

Tau-U (Parker et al. 2011), and (c) the standardized mean difference method, Busk-Serlin-Model 2 (Busk & Serlin, 1992).

The *ALLISON-MT* (Allison & Gorman, 1993; Faith et al., 1996) requires a statistical package that can calculate a multiple regression model. The calculation method is as follows:

1. Calculate a simple linear regression using phase A data, with the time (trend) variable as a predictor: $Y_A = b_0 + B_1T_A + e$.
2. If the slope of the trend variable is negative, complete steps 3 and 4. If the slope of the trend variable is positive, jump to step 5 (to control for positive baseline slope).
3. Calculate a multiple linear regression using all data, with the treatment variable (X) and the time (trend) variable as predictors: $Y = b_0 + B_1X + B_2T + e$.
4. Obtain adjusted R^2 .
5. Compute residuals for baseline data.
6. Use the regression equation generated in step 1 to generate predicted Y values for phase B.
7. Compute residuals for treatment phase by subtracting the predicted values in step 4 from the actual Y values in the treatment phase.
8. Use the residual or “detrended” data from step 5 and 7 to complete step 9.
9. Calculate a multiple linear regression with the treatment variable (X) and time/trend of the treatment phase ($X*T$) as predictors : $residual(Y) = b_0 + b_1X + b_3XT + e$.
10. Obtain adjusted R^2 .

The *Tau-U* (Parker, et al., 2011) method requires a statistical package that can calculate Kendall's Rank Correlation. *StatsDirect Statistical Software*, version 2.7.8, was used for this study. The calculation method is as follows:

1. Make a column for Score and Phase (0 or 1).
2. Backwards-code phase A Scores (i.e., put them in order from last to first). Maintain the actual score values for phase B.
3. Conduct a KRC analysis and obtain value for *S*.
4. Calculate the # of pairs for the model, where $\#pairs = [n(n - 1)]/2$
5. Calculate Tau, where $Tau = S / \#pairs$.

The Busk-Serlin-Model 2 (Busk & Serlin, 1992) method can be calculated by hand using the following procedures:

1. Obtain the difference between the means of both phases: $\bar{y}_B - \bar{y}_A$.
2. Calculate the pooled standard deviation (SD) of both phases:

$$SD_{pooled} = \sqrt{[df_A(SD_A)^2 + df_B(SD_B)^2] / DF_{total}}$$

3. Obtain effect size by dividing the value obtained in Step 1 by the pooled SD.
4. Convert the effect size value (*x*) to η^2 using $\eta^2 = [x / [x^2 + \left[\frac{N^2 - 2N}{n_1 n_2} \right]^5]]^2$. This

formula was originally proposed as a conservative method for use when converting *d* to R^2 when small sample sizes are present (Aaron, Kromrey, & Ferron, 1998).

Statistical Differences between Effect Size Estimates

A repeated-measures analysis of variance (ANOVA) was used to identify differences between scores obtained from each effect size method. This represents a non-traditional use

of a repeated-measures ANOVA which typically includes measuring the same individual or case at two or more points in time. Instead, because each of the effect size measurements used in this study aim to describe the same thing (i.e., the effect of an intervention on WCPM) and use the same data, a repeated-measures ANOVA can be used to look for differences between scores obtained by the three methods.

Autocorrelation Removal

The effect of autocorrelation on each of the three effect-size methods was examined by cleansing all lag-1 autocorrelation from the original data using an ARIMA model through the *NCSS Statistical Software and Graphics* program (Hintze, 2000) and then recalculating each effect size using the methods described above. A repeated-measures ANOVA was then used to identify differences between the three methods after autocorrelation is removed. Another repeated-measures ANOVA was used to identify differences between effect-size scores from the same method before and after autocorrelation was removed.

Score Distributions

Box plots were used to illustrate the typical effect-size ranges of each method when analyzing “effective interventions.” This graphical technique is considered the method of choice for describing score distributions (Tukey, 1977) and is used in much of the current single-case effect size research (Parker, et al., 2005; Parker & Brossart, 2003). A typical box plot includes the top and bottom of the box representing the 75th and 25th percentile ranks, with the box divided by the 50th percentile (median).

Chapter 3

Results

Differences Between Effect Sizes

A repeated-measures ANOVA was completed to explore differences in effect size estimates obtained from the three effect size methods. Mean effect size scores for each calculation method are presented in the first row of Table 1. Effect size method was found to have a significant effect on the obtained effect size scores ($F(1.783, 115.87) = 6.76, p = .002, \eta^2 = .094$). Observed power equaled .887. Bonferroni post-hoc tests were completed to further investigate differences between the three effect size methods. The Allison-MT method ($M = .557$) produced significantly larger effect size scores ($p = .001$) than the Tau-U method ($M = .477$), as did the Busk-Serlin-Model 2 (converted to η^2 ; $M = .552, p = .010$). There was no significant difference between scores produced from the Allison-MT and Busk-Serlin-Model 2 methods ($p = .788$).

Differences Between Scores Produced by the Same Effect Size Method, Before and After Autocorrelation Removal

Three repeated-measures ANOVAs were completed to examine differences between effect size scores before and after autocorrelation removal using the same effect size calculation method. A significant effect of autocorrelation removal was found for each of the three methods used (Allison-MT: $F(1, 65) = 214.44, p < .001, \eta^2 = .767$; Tau-U: $F(1, 65) = 130.83, p < .001, \eta^2 = .668$; Busk-Serlin-Model 2 (converted to η^2): $F(1, 65) = 348.44, p < .001, \eta^2 = .841$), with original effect size scores being significantly larger than their corresponding effect size scores when autocorrelation was removed. Mean effect size scores

obtained from each effect size method both before and after lag-1 autocorrelation removal are presented on the first and second rows of Table 1.

Differences Between Effect Sizes After Lag-1 Autocorrelation Removal

Another repeated-measures ANOVA was completed to investigate differences between effect size scores after lag-1 autocorrelation was removed from the original data. Mean effect size scores for each calculation method are presented in the second row of Table 1. After lag-1 autocorrelation was removed from the data, effect size method was still found to have a significant effect on the obtained effect size scores ($F(1.362, 88.51) = 9.006, p = .001, \eta^2 = .122$). Observed power equaled .913. Bonferroni post-hoc tests revealed that the Allison-MT method ($M = .109$) resulted in scores significantly larger than the scores obtained from the Busk-Serlin-Model 2 (converted to η^2 ; $M = .059, p = .016$). The Tau-U method ($M = .157$) also resulted in scores significantly larger than the Busk-Serlin-Model 2 (converted to η^2 ; $p < .001$). There was no significant difference between scores produced from the Allison-MT and Tau-U methods ($p = .115$).

Descriptive Statistics

Box-plots of the data obtained from each effect size method before and after autocorrelation removal are included in Figure 1. The bottom and top of each box plot represents the first and third quartiles (25th and 75th percentiles) respectively, with the box divided by the 50th percentile (median). The range of scores from the first to third quartiles is called the inter-quartile range (IRQ). The boxplots show variability among the three analytic methods in median value and in score variability (IRQ). Before autocorrelation removal, the

median effect size score was .601 for Busk-Serlin-Model 2, .467 for Tau-U, and .643 for Allison-MT. Distribution variability also varied between models, with IRQ values equaling .450 for Busk-Serlin-Model 2 (with the middle 50 percent of scores ranging from 0.31 to 0.76), .312 for Tau-U (with the middle 50 percent of scores ranging from 0.32 to 0.63), and .525 for Allison-MT (with the middle 50 percent of scores ranging from 0.27 to 0.79). Overall, median scores and IRQ values were lower after lag-1 autocorrelation removal. After autocorrelation removal, the median effect size scores were .045 for Busk-Serlin-Model 2, .144 for Tau-U, and .091 for Allison-MT. IRQ values after autocorrelation removal were .063 for Busk-Serlin-Model 2 (with the middle 50 percent of scores ranging from 0.014 to 0.076), .213 for Tau-U (with the middle 50 percent of scores ranging from 0.043 to 0.256), and .265 for Allison-MT (with the middle 50 percent of scores ranging from -0.045 to 0.221). After autocorrelation removal, each of the distributions was approximately able to fit effect size guidelines set forth by Cohen for group designs, which specifies .01 to .14 for weak to strong effects for η^2 scores (which would be appropriate for Busk-Serlin Model 2), and .01 to .25 for weak to strong effects for R^2 scores (which would be appropriate for Allison-MT and Tau-U; Ellis, 2010, p.42). The Busk-Serlin Model 2 scores are slightly lower than expected if using the Cohen classifications for categorical variables as a guide.

Chapter 4

Discussion

The purpose of this study was to compare three competing methods in the analysis of single-case AB design data that utilized the WCPM variable. Three calculation techniques were compared, including the ALLISON-MT technique (from the regression-based models;

Allison & Gorman, 1993), the new Tau-U technique (from the non-overlap models; Parker, et al. 2011), and the Busk-Serlin-Model 2 (from standardized mean difference models; Busk & Serlin (1992). Data from a total of sixty-six graphs were analyzed and comparisons between and within effect size estimates were calculated both before and after the removal of lag-1 autocorrelation from the original data. Descriptive statistics of effect size ranges and median scores were also examined.

Before discussing overall findings of the study, it is important to reiterate the importance of using effect sizes for single-case designs to supplement traditional visual analysis methods. First, interrater reliability estimates of visual analysis are reported to be as low as .58 (Brossart, et al. 2006), suggesting that practitioners and researchers would likely disagree about visual interpretations at least some of the time. Second, effect sizes provide a way to illustrate relative effectiveness of intervention techniques. In this study, all graphs were deemed to show “effective” interventions, but effect size estimates ranged widely across graphs. Finally, there is evidence that many educators who make instructional decisions have little training and applied experience with single-case designs and interpreting graphed data (Begeny & Martens 2006); therefore, providing an objective way of measuring improvement would be useful and likely support educators’ instructional decision-making.

Differences Between Effect Size Scores

The first research question examined in this study addressed whether effect size estimates would differ significantly from each other when analyzing the same sets of data. As predicted, both the Busk-Serlin-Model 2 and Allison-MT methods produced significantly

larger effect sizes than the Tau-U method. This prediction was made because Tau-U is a more conservative calculation method for effect sizes than the other two. It is arguably more conservative than the Busk-Serlin-Model 2 because it takes into account non-overlap of all pairs and baseline trend; and because Tau-U is a non-parametric method, its scores would theoretically be less affected by small sample sizes compared to the Allison-MT method (a parametric, regression-based approach). However, there was no significant difference found between the Busk-Serlin-Model 2 and Allison-MT methods, when it was expected that the Busk-Serlin-Model 2 scores would be higher. This was a surprising finding because, theoretically, the Allison-MT method is a more conservative approach to the Busk-Serlin-Model 2 because it first controls for baseline trend, and then computes a regression with two predictors (phase and phase 2 trend). This finding calls into question the importance of controlling for baseline trend when calculating effect sizes and suggests that at least with the WCPM variable, baseline trend may not artificially increase effect size scores to the degree previously predicted by researchers. However, it is also important to remember that these graphs were all identified by original authors as showing positive intervention effects. This suggests that baseline trend control may be more necessary when trying to distinguish between a positive effect and no effect.

The examination of box plots in Figure 1 allows for the examination of ranges of effect size scores before lag-1 autocorrelation removal. Before lag-1 autocorrelation removal the median effect size score was .60 for Busk-Serlin-Model 2, .47 for Tau-U, and .64 for Allison-MT. Overall effect size scores ranged from .08 to .94 for Busk-Serlin Model 2, -.07 to .85 for Tau-U, and .02 to .95 for Allison-MT. These large effect size ranges indicate that

there is much variability in data that previous authors have determined through visual analysis to depict effective interventions, and also illustrates the value of effect sizes in providing a magnitude of effect estimate. By using effect sizes, interventions that were previously grouped together as “effective,” can now be classified in order of their relative effectiveness.

Differences Between Scores Produced by the Same Effect Size Method, Before and After Autocorrelation Removal

The second question addressed by this study was the degree to which original effect size calculations differed from those calculated after the removal of lag-1 autocorrelation. Because of research reporting high levels of autocorrelation in single-case data (Matyas & Greenwood, 1996; Parker, et al., 2005; Sharpley & Alavosius, 1988; Suen & Ary, 1989), and because autocorrelation arguably increases effect size scores (Brossart et al., 2006; Scheff, 1959), it was hypothesized that original scores would be significantly larger than scores obtained from the same effect size calculation after the removal of lag-1 autocorrelation. As predicted, all original scores were significantly larger than their corresponding scores calculated after the removal of autocorrelation. Difference in median scores from before and after autocorrelation removal equaled 0.56 for Busk-Serlin-Model 2, 0.32 for Tau-U, and 0.55 for Allison-MT, illustrating quite a change of scores with the removal of autocorrelation. This is not surprising, given previously published papers suggesting that one-third or more of published single-case design datasets are autocorrelated to an undesirable degree ($>.20$ to $.25$) (Matyas & Greenwood, 1996; Parker, et al., 2005; Sharpley & Alavosius, 1988; Suen & Ary, 1989).

An examination of box plots in Figure 1 show that after lag-1 autocorrelation removal, the median effect size score was 0.05 for Busk-Serlin Model 2, 0.14 for Tau-U, and 0.09 for Allison-MT. With the removal of autocorrelation from the original data, score ranges appeared more commensurate with the score levels that Cohen (1988) used to describe effect sizes for group designs. For R^2 values (which is appropriate nomenclature for Allison-MT and Tau-U scores because of the presence of categorical and continuous prediction variables), Cohen suggested that a large effect = .25, a medium effect = .09, and a small effect = .01. For η^2 values (appropriate nomenclature for the Busk-Serlin Model 2 scores because only a categorical variable is present), Cohen suggested that a large effect = .14, a medium effect = .06, and a small effect = .01. The median effect size scores found in this study after the removal of lag-1 autocorrelation approximately fit within the respective guidelines set forth by Cohen, with Busk-Serlin scores falling slightly lower than would be expected. This finding suggests that single-case design effect sizes using the WCPM variable may be interpretable according to effect size categories suggested for group designs, as long as lag-1 autocorrelation is first removed from the data. Although more research is needed to substantiate this interpretation, this finding has important implications. First, it suggests increased ability to interpret single-case effect sizes with commonly used categorizations, which addresses others' criticisms (e.g., Parker et al., 2005; Brossart, et al., 2006) that interpreting effect sizes is difficult due to seemingly inflated values and no consistent way to categorize such effect sizes. Second, being able to interpret single-case design effect sizes using categories commonly used with between group designs may also allow better interpretability of intervention effects that come from different research methods (e.g.,

between group versus single-case). That is, even though effect size estimates for single-case designs brought researchers a step closer to having a complementary method of analysis to group designs (compared to only using visual analysis with single-case designs), researchers still struggled to compare single-case effect size estimates to those from group designs because single-case effect sizes tended to be much larger, and as shown in his study, the reason for this may have been due to autocorrelated data.

Differences Between Effect Size Scores After Lag-1 Autocorrelation Removal

Interestingly, when lag-1 autocorrelation was removed from the data, comparisons among the three effect sizes revealed different relationships than without autocorrelation removal. Consistent with hypotheses related to the first research question, it was also hypothesized that Allison-MT and Tau-U would both be significantly smaller than the Busk-Serlin-Model 2 method after autocorrelation removal because the Busk-Serlin-Model 2 is the least conservative of the three approaches. However, both the Allison-MT and Tau-U approaches were found to be significantly larger than the Busk-Serlin-Model. This finding may suggest that of the three different formulae, the Busk-Serlin-Model 2 is most affected by autocorrelated data, which results in vastly different scores before and after autocorrelation removal. This idea is illustrated by the box plot representations of score ranges in Figure 1, which show that the Busk-Serlin-Model 2 median is lower and the IRQ is much smaller than those of the other two methods after lag-1 autocorrelation removal, whereas median score and IRQ was very similar to Allison-MT before removal of lag-1 autocorrelation removal. Additionally, although it was hypothesized that scores obtained from the Allison-MT method

would be significantly larger than those obtained from the Tau-U method after autocorrelation removal, no significant difference was found between the two sets of scores.

Limitations and Future Directions

One important limitation of this study is that it only investigated the use of effect size methods with AB single-case designs. As mentioned above, AB designs are the weakest of the single-case methodologies, mainly because the investigator is unable to determine if a change in behavior is a result of other environmental factors than the independent variable. Although the use of AB designs is relevant for practitioners, there are almost no AB designs published in the research literature (which is why multiple baseline designs were divided into AB designs for this study). Future studies should focus on utilizing effect sizes with other and potentially stronger single-case designs. This type of research would be important because with alternating treatment and changing criterion designs, effect size calculations may need to vary compared to those used with AB and extended AB designs, and there is also the possibility that effect sizes with other designs may be more difficult to calculate and/or interpret.

Additionally, a simultaneous advantage and disadvantage of this study involves the choice of a target dependent variable. No other studies have examined a single target variable when investigating effect size scores, but instead pooled all AB designs (taken from multiple baseline studies) published within a particular range of time. Examining a specific target variable (WCPM) is advantageous because one is able to develop an understanding of ranges of scores typical for that variable. This is the first step to then developing guidelines for

interpreting magnitude of response and determining a “successful” response to intervention for that particular variable. However, by restricting this study to only the WCPM variable, findings cannot be generalized to other variables, and additional research is needed to examine typical effect size ranges for other target variables (e.g., time on task, digits calculated correctly per minute).

Another limitation of this study is that the obtained results did not allow for a clear interpretation of the “best” effect size to use with single-case data. Although, after the removal of lag-1 autocorrelation from the original data, Tau-U and Allison-MT appear to be most interpretable according to commonly used categorizations for group designs, future research is needed before advocating that any specific effect size formulae be used instead of others. The process of determining the best method is difficult because there is no way to determine if the “true” or “accurate” effect size score has ever been obtained. Within this study, typical ranges of effect size scores obtained from cleansed WCPM data were identified. A next step might be to investigate if each effect size method rank orders the datasets in the same way. This would allow for not only the comparison of mean effect size scores but also to determine if the methods are able to categorize data similarly. Another possible research question will be to ask experts and non-experts to rate graphs showing small, medium, or large effects, and then examine the effect size ranges for each of the categories, while also examining if there were differences between expert and non-expert reviewers.

Overall Conclusions and Implications

This study aimed to compare the three leading single-case effect size analysis methods and determine whether one technique would appear as the “best choice” for analyzing AB designs utilizing WCPM data. When evaluating the data as a whole, all effect size scores decreased significantly after the removal of lag-1 autocorrelation. Because of research reporting high levels of autocorrelation in single-case data (Matyas & Greenwood, 1996; Parker, et al., 2005; Sharpley & Alavosius, 1988; Suen & Ary, 1989), and because autocorrelation increases effect size scores (Brossart et al., 2006; Scheff, 1959), it is arguably the best practice to remove lag-1 autocorrelation before calculating effect size scores. By doing so, the researcher makes an attempt to control potential confounds in the data and, at least when using the WCPM dependent variable, effect size scores may appear more interpretable because they are consistent with Cohen’s (1988) categorical guidelines for interpreting R^2 and η^2 .

When effect size scores were examined after the removal of lag-1 effect size scores, no significant difference was found between the scores obtained from the Allison-MT and Tau-U methods, suggesting that they may represent two different approaches for reaching the same result (as long as autocorrelation is first removed). Future research should continue to compare the two methods, and if they continue to produce similar results, the single-case research community may eventually use the methods interchangeably or in combination with one another. This would fall in line with single-case design standards set forth by the Institute of Educational Sciences, which state that it is best practice to use both a regression-based and non-parametric-based method when analyzing single-case design data (Kratochwill, et al.,

2010). The Busk-Serlin-Model 2 method appears to be most affected by autocorrelation, and changes in effect size scores after removing autocorrelation was arguably too conservative. Based on this and the fact that both the Allison-MT and Tau-U methods incorporate a method of controlling for baseline trend that the Busk-Serlin Model 2 does not, it appears that both the Allison-MT and Tau-U methods provide more accurate and reliable ways of calculating effect sizes for single-case research that utilizes WCPM as the dependent variable.

With respect to implications for educational practice, the use of single-case effect sizes within schools can arguably provide a way to increase the accuracy and effectiveness of the RTI process. Because high-stakes, special education decisions are more frequently being made based on student support team interpretation of a student's response to an intervention, school personnel should be held to an ethical standard of interpreting data based on sound analysis methods. Moderate levels of inter-rater reliability of visual analysis paired with many teachers' unfamiliarity with using and interpreting single-case designs (Begeny & Martens, 2006) may increase the risk of inaccurate judgments about student response or non-response and ultimately the possibility that a student may not receive the appropriate educational support (e.g., the student may inappropriately receive more or less intensive services). The use of effect sizes within the RTI process may therefore provide a more objective, interpretable approach to measuring student improvement. This study is particularly relevant for practitioners because of the use of AB designs. Although AB designs are methodologically the weakest of the single-case design options, many schools may not have the option (or it may be unethical) to employ more sophisticated designs; therefore, AB

designs will likely remain prevalent as part of the RTI process. The findings of this study offer preliminary support for the use of the Tau-U and Allison-MT methods after the removal of lag-1 autocorrelation from the original data. A logical next step would be to understand how to increase practitioners' use of these methods within their schools. The more effect size methods are utilized and published, the better practitioners should be able to interpret a student's response to intervention. Barth and colleagues (2008) illustrated the need and importance of specific guidelines for RTI by examining data collected from 339 first grade students involved in the RTI process across six schools. The researchers found low agreement among educators on methods used to determine response to intervention and also showed that different methods tend to identify different students as inadequate responders. An important future study would involve calculating effect size scores from a bank of existing RTI data and examining differences in scores between students identified as "responders" and "non-responders."

References

*Articles that included graphs used for effect size calculations

*Alber-Morgan, S. R., Ramp, E. M., Anderson, L. L., & Martin, C. M. (2007). Effects of repeated readings, error correction and performance feedback on the fluency and comprehension of middle school students with behavior problems. *The Journal of Special Education, 41*, 17-30.

- *Allen-DeBoer, R. A., Malmgren, K. W., Glass, M. (2006). Reading instruction for youth with emotional and behavioral disorders in a juvenile correctional facility. *Behavioral Disorders, 32*, 18-28.
- Allison, D. B. & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behavior Research and Therapy, 31*, 621-631.
- Alberto, P., & Troutman, A. (1999). *Applied behavior analysis for teacher* (5th ed.). Columbus, OH: Merrill.
- American Psychological Association. (2010). Publication manual of the American psychological association. Washington, D.C.: American Psychological Association.
- Ardoin, S. P., Suldo, S. M., Witt, J., Aldrich, S., & McDonald, E. (2005). Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly, 20*, 1-22.
- *Arthaud, T. J., & Ranin, J. L. (1996). Effects of differential feedback up the oral reading fluency of secondary students with learning disabilities. *Diagnostique, 21*, 41-57.
- Bain, S. K., & Garlock, J. W. (1992). Cross-validation of criterion-related validity for CBM reading passages. *Diagnostique, 17*, 202-208
- Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change*. Boston, MA: Pearson Education, Inc.
- Barth, A. E., Stuebing, K. K., Anthony, J. L., Denton, C. A., Mathes, P. G., Fletcher, J. M., & Francis, D. J. (2008). Agreement among response to intervention criteria for identifying responder status. *Learning and Individual Differences, 18*, 296-307.

- *Barton-Arwood, S. M., Wehby, J. H., & Falk, K. B. (2005). Reading instruction for elementary-age students with emotional and behavioral disorders: Academic and behavioral outcomes. *Exceptional Children, 72*, 7-27.
- Begeny, J. C., & Martens, B. K., (2006). Assessing pre-service teachers' training in empirically- validated behavioral instruction practices, *School Psychology Quarterly, 21*, 262-285.
- Bliss, S. L., Skinner, C. H., Hautau, B., & Carroll, E. E. (2008). Articles published in four school psychology journals from 2000 to 2005: An analysis of experimental/intervention research. *Psychology in the Schools, 45*, 483-499.
- Bonner, M., & Barnett, D. W. (2004). Intervention-based school psychology services: Training for child-level accountability; preparing for program-level accountability. *Journal of School Psychology, 42*, 23-43.
- *Bray, M. A., Kehle, T. J., Spackman, V. S., & Hintz, J. M. (1998). An intervention program to increase reading fluency. *Special Services in the Schools, 14*, 105-125.
- Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*, 531-563.
- Burns, M., & Gibbons, K. (2008). *Implementing Response to Intervention Procedures in Elementary and Secondary Schools: Procedures to Assure Scientific-Based Practice*. Routledge: New York, New York.

- Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meta-analysis for single-case consultations outcomes: Applications to research and practice. *Journal of School Psychology, 33*, 269-285.
- Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill, R. Thomas, and J. R. Levin (Eds.). *Single-case research design and analysis: New directions for psychology and education*. (pp. 187-212). England: Lawrence Earlbaum Associates.
- Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*, 234-246.
- Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19*, 387-400.
- Chard, D. J., Vaughn, S., & Tyler, B. J. (2002). A synthesis of research on effective interventions for building reading fluency with elementary students with learning disabilities. *Journal of Learning Disabilities, 35*, 386-406.
- Chen, C. W., & Ma, H. H. (2007). Effects of treatment on disruptive behaviors: A quantitative synthesis of single-subject researches using the PEM approach. *The Behavior Analyst Today, 8*, 380-397.
- Christ, T., & Silbergliitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review, 36*, 130-146.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

- Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61*, 966-974.
- Deno, S. L., & Mirkin, P. K. (1977). *Data-Based Program Modification: A Manual*. Minneapolis: Leadership Training Institute/Special Education.
- DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intra-subject data. *Journal of Applied Behavior Analysis, 12*, 573-579.
- *Drakeford, W. (2002). The impact of an intensive program to increase the literacy skills of youth confined to juvenile corrections. *Journal of Correctional Education, 53*, 139-144.
- *Dufrene, B. A., Henington, C., & Townsend, A. E. (2006). Peer tutoring for reading fluency: Student implementation and effects on reading fluency. *Journal of Evidence-Based Practices for Schools, 7*, 118-137.
- *Dufrene, B. A., Reisener, C. D., Olmi, D. J., Zoder-Martell, K., McNutt, M. R., & Horn, D. R. (2010). Peer tutoring for reading fluency as a feasible and effective alternative in response to intervention systems. *Journal of Behavioral Education, 19*, 239-256.
- Edgington, E. S. (1984). Statistics and single-case designs. *Progress in Behavior Modification, 16*, 83-119.
- Faith, M. S., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single-case research. In R.D. Franklin, D.B. Allison, and B.S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245-277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Furlong, M. J., & Wampold, B. (1981). Visual analysis of single-subject studies by school psychologist. *Psychology in the Schools, 18*, 80-86.

- Furlong, M. J., & Wampold, B. (1982). Intervention effects and relative variation as dimensions in experts' use of visual inference. *Journal of Applied Behavior Analysis, 15*, 415-421.
- Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation, *Psychological Bulletin, 70*, 245-251.
- *Gilbert, L. M., Williams, R. L., & McLaughlin, T. F. (1996). Use of assisted reading to increase correct reading rates and decrease error rates with students with learning disabilities. *Journal of Applied Behavior Analysis, 29*, 255-257.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.
- Good, R. H., Simmons, D. C., & Kameenui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*, 257-288.
- Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 5*, 141-154.
- Graney, S. B., & Shinn, M. R. (2005). Effects of reading curriculum-based measurement (R-CBM) teacher feedback in general education classrooms. *School Psychology Review, 34*, 184-201.
- Harbst, K. B., Ottenbachers, K. J., & Harris, S. R. (1991). Interrater reliability of therapists' judgments of graphed data. *Physical Therapy, 71*, 107-115.

- *Harris, P. J., Oakes, W. P., Lane, K. L., & Rutherford, R. B. (2009). Improving the early literacy skills of students at risk for internalizing or externalizing behaviors with limited reading skills. *Behavioral Disorders, 34*, 72-90.
- Hintze, J. M., & Shapiro, E. S. (1997). Curriculum-based measurement and literature-based reading: Is curriculum-based measurement meeting the needs of changing reading curricula? *Journal of School Psychology, 35*, 351-375.
- Hintze, J. M., Shapiro, E. S., & Lutz, J. G. (1994). The effects of curriculum on the sensitivity of curriculum-based measurement in reading. *Journal of Special Education, 28*, 189-202.
- Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34*, 372-386.
- *Hook, C. L., & Dupaul, G. J. (1999). Parent tutoring for students with attention-deficit/hyperactivity disorder: Effects on reading performance at home and school. *School Psychology Review, 28*, 60-75.
- Hosp, M. H., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review, 34*, 9-26.
- Houle, T. T. (2009). Statistical analyses for single-case experimental designs. In D. H. Barlow, M. K. Nock, and M. Hersen (Eds.), *Single case experimental designs: Strategies for studying behavior change* (pp. 271-305). Boston, MA: Pearson Education, Inc.

- Howe, K. B., & Shinn, M. M. (2002). *Standard reading assessment passages (RAPs) for use in general outcome measurement: A manual describing development and technical factors*. Retrieved January 2010, from www.aimsweb.com.
- H.R. 1--107th Congress: No Child Left Behind Act of 2001. (2001). In *GovTrack.us* (database of federal legislation). Retrieved July 27, 2010, from <http://www.govtrack.us/congress/bill.xpd?bill=h107-1>
- Individuals with Disabilities Education Improvement Act of 2004 (IDEIA), Pub. L. No. 105-17, 20 U.S.C. §§ 1400 et seq.
- Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44*, 483-493.
- Kazdin, A. E. (2010). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.
- Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.
- Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology, 95*, 3-21.

*Lane, K. L., Little, M. A., Redding-Rhodes, J., Phillips, A., & Welsh, M. T. (2007).

Outcomes of a teacher-led reading intervention for elementary students at risk for behavioural disorders. *Exceptional Children*, 74, 47-70.

Ma, H. (2006). An alternative method for quantitative synthesis of single-subject researches:

Percentage of data points exceeding the median. *Behavior Modification*, 30, 598-617.

Manolov, R., & Solanas, A. (2008). Comparing N=1 effect size indices in presence of

autocorrelation. *Behavior Modification*, 32, 860-875.

Manolov, R., Solanas, A., & Leiva, D. (2010). Comparing “visual” effect size indices for

single-case designs. *Methodology*, 6, 49-58.

Marston, D. B. (1989). A curriculum-based measurement approach to assessing academic

performance: What it is and why do it. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford Press.

Matyas, T. A., & Greenwood, K. M. (1996). Serial dependency in single-case time series. In

R. D. Franklin, D. B. Allison, and B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215-233). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to

predict performance on state assessments in reading. *School Psychology Review*, 33, 193-203.

Meyer, M. S., & Felton, R. H. (1999). Repeated reading to enhance fluency: Old approaches

and new directions. *Annals of Dyslexia*, 49, 283-306.

Morgan, D. L., & Morgan, R. K. (2009). *Single-case research methods for the behavioral*

and health sciences. Los Angeles, CA: Sage.

- Morgan, P. L., & Sideridis, P. D. (2006). Contrasting the effectiveness of fluency interventions for students with or at risk for learning disabilities: A multilevel random coefficient modeling meta-analysis, *Learning Disabilities Research & Practice, 21*, 191-210.
- National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.
- *Nelson, J. S., Alber, S. R., & Gordy, A. (2004). Effects of systematic error correction and repeated readings on the reading accuracy and proficiency of second graders with disabilities. *Education and Treatment of Children, 27*, 186-198.
- *Noell, G. H., Freeland, J. T., & Witt, J. C. (2001). Using brief assessments to identify effective interventions for individual students. *Journal of School Psychology, 39*, 335-355.
- *Noell, G. H., Gansel, K. A., & Witt, J. C. (1998). Effects of contingent reward and instruction on oral reading performance and differing levels of passage difficulty. *Journal of Applied Behavior Analysis, 31*, 659-663.
- *Oddo, M., Barnett, D. W., Hawkins, R. O., & Musti-Rao, S. (2010). Reciprocal peer tutoring and repeated reading: Increasing practicality using student groups. *Psychology in the Schools, 47*, 842-858.
- Ottensbacher, K. J. (1990). Visual analysis of single-subject data: An empirical analysis. *Mental Retardation, 28*, 283-290.

- Ottenbacher, K. J., & Cusick, A. (1991). An empirical investigation of inter-rater agreement for single-subject data using graphs with and without trendlines. *Journal of the Association for Persons with Severe Handicaps, 16*, 48-55.
- Parker, R. I., & Brossart, D. F. (2003). Evaluation single-case research data: A comparison of seven statistical methods, *Behavior Therapy, 34*, 189-211.
- Parker, R. I., Brossart, D. F., Callicott, K. J., Long, J. R., Garcia de Alba, R., Baugh, F. G., et al. (2005). Effect sizes in single-case research: How large is large? *School Psychology Review, 34*, 116-132.
- Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*, 418-443.
- Parker, R. I., & Hagan-Burke, S. (2007). Useful effect size interpretations for single case research. *Behavior Therapy, 38*, 95-105.
- Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education, 40*, 194-204.
- 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*, 284-299.
- Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement on words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment, 23*, 326-338.
- *Resetar, J. L., Noell, G. H., & Pellegrin, A. L. (2006). Teaching parents to use research-supported systematic strategies to tutor their children in reading. *School Psychology Quarterly, 21*, 241-261.

- Richards, S.B., Taylor, R., Ramasamy, R., & Richards, R.Y. (1998). *Single subject research: Application in Educational and Clinical Settings*. Florence, KY: Wadsworth Publishing.
- Riley-Tillman, T. C., & Burns, M. K. (2009). *Evaluating educational interventions: Single-case designs for measuring response to intervention*. New York, NY: The Guilford Press.
- *Rinaldi, L., Sells, D., & McLaughlin, T. F. (1997). The effects of reading racetracks on the sight word acquisition and fluency of elementary students. *Journal of Behavioral Education, 7*, 219-233.
- Rosnow, R., & Rosenthal, R., (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.
- Scheff, H. (1959). *The analysis of variance*. New York: John Wiley.
- Scruggs, T. E., & Mastropieri, M. A. (2001). How to summarize single-participant research: Ideas and applications. *Exceptionality, 9*, 227–244.
- Scruggs, T. E., & Mastropieri, M. A. (1994). The utility of the PND statistic: a reply to Allison and Gorman. *Behaviour Research and Therapy, 32*, 879-883.
- Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remediate and Special Education, 8*, 24-33.

- Sharpley, C. F., & Alavosius, M. P. (1988). Autocorrelation in behavioral data: An alternative perspective. *Behavioral Assessment, 10*, 243-251.
- Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*, 407-419.
- *Staubitz, J. E., Cartledge, G., Yurick, A., & Ya-Yu, L. (2005). Repeated reading for students with emotional or behavioral disorders: Peer-and trainer-mediated instruction. *Behavioral Disorders, 31*, 51-64.
- *Steventon, C. E., & Fredrick, L. D. (2003). The effects of repeated readings on student performance in the corrective reading program. *Journal of Direct Instruction, 3*, 17-27.
- *Swain, K. D., & Allinder, R. M. (1996). The effects of repeated Reading on two types of CBM: Computer maze and oral reading with second-grade students with learning disabilities. *Diagnostique, 21*, 51-66.
- Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Lawrence Erlbaum.
- Therrien, W. J. (2004). Fluency and comprehension gains as a result of repeated reading: A meta-analysis. *Remedial and Special Education, 25*, 252-261.
- *Valleley, R. J., & Shriver, M. D. (2003). An examination of the effects of repeated readings with secondary students. *Journal of Behavioral Education, 12*, 55-76.
- White, O.R., & Haring, N. G. (1980). *Exceptional teaching* (2nd ed.). Columbus, OH: Merrill.

- White, O. R., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment, 11*, 281-296.
- Wolery, M., & Harris, S. R. (1982). Interpreting results of single-subject research designs. *Physical Therapy, 62*, 445-452.
- Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.
- *Yurick, A., Robinson, P. D., Cartledge, G., Ya-yu, L., & Evans, T. L. (2006). Using peer-mediated repeated readings as a fluency-building activity for urban learners. *Education and Treatment of Children, 29*, 469-506.

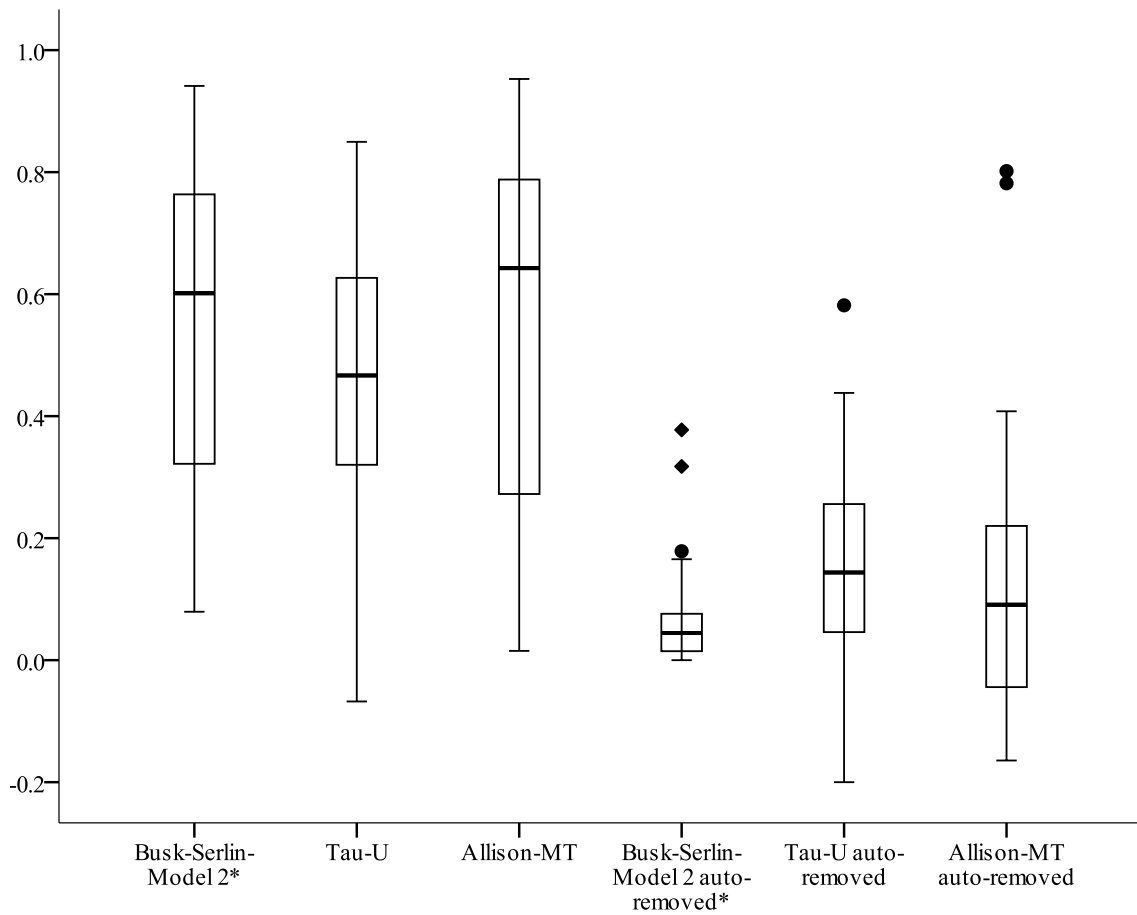


Figure 1. Box-Plot Display of Effect Size Ranges

Notes. Circles represent suspected outlier scores that fell outside the [(1.5)(Inter-Quartile Range)] range. Diamonds represent outlier scores that fell outside the [(3)(Inter-Quartile Range)] range. *The Busk-Serlin-Model 2 scores were converted to η^2 scores.

Table 1. Means of effect size estimates before and after autocorrelation removal (standard deviation)

	Allison-MT ¹	Tau-U ²	Busk-Serlin-Model 2 ^{3*}
Before Autocorrelation Removal	.557 ^{1>2} (.279)	.477 (.195)	.552 ^{3>2} (.243)
After Autocorrelation Removal	.109 ^{1>3} (.186)	.157 ^{2>3} (.195)	.059 (.069)

Notes. Bolded values reflect a statistically significant difference between the mean effect sizes of the same effect size formula. That is, for all three effect size formulae, the mean effect size before autocorrelation removal was greater ($p < .05$) than the mean effect size after autocorrelation removal. Additionally, each of the effect size methods are assigned a number in superscript, with greater-than signs used to indicate significant differences between scores obtained from each method within each category (before autocorrelation removal and after autocorrelation removal).

*Busk-Serlin-Model 2 scores have been converted to η^2 scores for ease of comparison with the other effect size scores.

Appendices

Appendix A

Overview of Six Most Common Single-Case Designs

AB Designs

- An AB design involves repeated measurement of behavior throughout a baseline (A) and intervention (B) phases of experimentation (Barlow, Nock, & Hersen, 2009).
- The A phase involves documentation of the natural rate of the chosen behavior, while the rate of behavior with an intervention in place is documented in phase B.
- Intervention effectiveness can be tentatively assumed if there is a change in the level of behavior from the A to B phase. However, practitioners should be cautious when using AB designs because it is difficult to distinguish between changes occurring as a result of intervention or changes that would have occurred regardless of phase change (Barlow, Nock, & Hersen).

Withdrawal and Extended ABA Designs

- During a Withdrawal design (also commonly called an ABA design), the typical AB design is completed, followed by the initial baseline phase being reinstated.
- Intervention effectiveness is determined examining the degree to which an individual's behavior returns to original baseline levels when the intervention is removed (Morgan & Morgan, 2009).
- To further increase experimental control, additional A and B phases can be introduced past the additional ABA phases, resulting in an extended ABA design (e.g., ABABAB).

- Although ABA designs successfully address the internal validity flaws of AB designs, there are sometimes ethical reasons that would prevent a practitioner from using these designs. For example, if a client shows a decrease in depressive symptoms with the onset of an intervention, it would sometimes be unethical for the clinician to then remove the intervention in an effort to demonstrate experimental control (Barlow, Nock, & Hersen, 2009).

Alternating Treatments Designs

- During an Alternating Treatments design, two or more conditions are alternated randomly across time (e.g., ABBABABAA or BACCBABC) (Morgan & Morgan, 2009).
- As with the above designs, alternating treatments designs are useful when comparing the effects of an intervention and baseline condition, but can also be used to compare the effects of two or more interventions.
- Intervention effectiveness is determined by the degree to which behavior changes consistently across phases.
- This design provides for even stronger levels of experimental control than previously discussed designs because there are multiple opportunities to show changes from one phase to another and because the phases are randomly chosen.
- The same ethical considerations that affect ABA and extended-ABA designs must be remembered when using alternating treatments designs as well.

- One important consideration regarding alternating treatments designs is the risk that effects of one intervention will carry-over to the next intervention's phase, or the carry-over effect (Barlow, Nock, & Hersen, 2009).
 - Basically, if one intervention is truly improving the student's skill level, one would expect that performance would increase during another intervention's phase as well.

Multiple Baseline Designs

- A multiple baseline design involves several AB designs, either using multiple participants, multiple behaviors within a single participant, or multiple settings within a single participant (Morgan & Morgan, 2009). The unique component of multiple baseline designs is that each AB design has a baseline of varying lengths.
- Intervention effectiveness is determined if there is a change from the baseline to intervention phase for all participants, regardless of length of baseline condition (Morgan & Morgan).
- These designs can be useful when it may be undesirable or unethical to withdrawal an intervention already in place and to avoid potential intervention carry-over effects.
- This design may be especially useful for practitioners interested in applied research who want to monitor the effects of an intervention on multiple students.

Changing Criterion Designs

- A changing criterion design involves evaluating the effects of an intervention on the systematic increase or decrease of the target behavior (Richards, et al. 1999).

- A Changing Criterion design begins with a baseline phase. Then incremental goals, or criterion levels, are decided upon, with the last criterion being the final goal for the target behavior. After the first criterion is met for two consecutive sessions (guidelines by Alberto & Troutman, 1999), then the next criterion becomes the goal, and so on.
- Intervention effectiveness is determined if criteria are continually met across time, with the ultimate goal being the final criterion (Richards, et al. 1999).
- This type of design can be especially useful when the terminal goal takes a relatively long time to reach (Alberto & Troutman, 1999) and when it is unethical to withdraw an intervention (Richards, et al. 1999).

Appendix B

Article Inclusion Criteria from Reading Fluency Literature Review Project

Each article included in the literature review project met the following criteria:

1. Published in an academic, peer reviewed journal.
2. Published in English.
3. At least one outcome variable measured reading rate (i.e., reading speed and accuracy) and/or prosody for some form of print (i.e., letters, nonsense words/phonemes, words in isolation, sentences, and/or connected text). *Articles that did not provide at least some quantitative form of reading fluency measurement (e.g., raw or standard score from a reading assessment, a subjective but quantifiable rating of reading ability) were excluded.*
4. An *intervention* is described in the report that:
 - A. provides instructional, practice, and/or motivational (e.g., performance feedback, reinforcement) strategies for participants with the goal of improving participants' reading ability;
 - B. utilizes some form of print for instructional material (i.e., letters, nonsense words/phonemes, words in isolation, sentences, and/or connected text); and
 - C. is *not* implemented with participants who have reading difficulties that strictly manifest from a sensory deficit (e.g., vision), an acquired brain injury, or some form of medical illness.

Based on the inclusion criterion described in #4, the following types of articles were excluded:

- A. Studies that only evaluated measures of reading fluency, but did not evaluate reading fluency as a result of an intervention (e.g., studies simply evaluating the relationship between oral reading fluency and other measures of reading).
- B. Studies that evaluated instruction in a *naturalistic* context (i.e., nothing was changed or added to participants' reading instruction), even if reading fluency was evaluated. In other words, "intervention" is considered to be some change in individuals' typical instructional routine, as deemed appropriate or potentially beneficial by the authors of the study.
- C. Studies that strictly utilize medication or an apparatus (e.g., color overlays) to improve reading performance, and thereby do not integrate the components described in 4A and/or 4B.

Lastly, to avoid duplicating methodological information (e.g., the same participants receive the same procedures and measures), studies that reported "follow-up" data that were

previously published are not included in the database of studies. With this same rationale, if an includable study reported more than one experiment within the same publication, we only coded “Experiment 1” of that study.