A COMPARISON OF THREE EFFECT SIZE INDICES FOR COUNT-BASED OUTCOMES IN SINGLE-CASE DESIGN STUDIES

by

Pragya Shrestha

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Education

Summer 2023

A COMPARISON OF THREE EFFECT SIZE INDICES FOR COUNT-BASED OUTCOMES IN SINGLE-CASE DESIGN STUDIES

by

Pragya Shrestha

Approved: _____
Steven J. Amendum, Ph.D.
Interim Director, School of Education

Approved: _____
Gary T. Henry, Ph.D.
Dean of the College of Education and Human Development

Approved: _____
Louis F. Rossi, Ph.D.
Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Henry May, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Zachary K. Collier, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Laura Desimone, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

James E. Pustejovsky, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Henry May for his guidance and support in completing my PhD. It would not have been possible without his faith in me. I am also thankful for the guidance from my dissertation committee members: Dr. Zachary Collier, Dr. Laura Desimone, and Dr. James Pustejovsky.

I have been able to complete my doctoral program because of the support from my family and friends. Special thanks to my husband (Sunendra Joshi), my son (Prasun), parents, and siblings for their love, understanding, and motivation to help achieve my goals. Also, I would like to acknowledge my friend, Jing Yan for his help and suggestions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In Single-Case Designs (SCD), the outcome variable most commonly involves some form of count data. However, statistical analyses and associated effect size (ES) calculations for count outcomes have only recently been proposed. Three recently proposed ES methods for count data are: Nonlinear Bayesian effect size (Rindskopf, 2014), Log Response Ratio effect size (Pustejovsky, 2018), and Bayesian Rate Ratio effect size (Natesan Batley, Shukla Mehta, & Hitchcock, 2021). Although all three methods calculate ES for count outcome data and can be used with an ABAB design, they use either different statistical modeling or a different estimation framework (Bayesian or frequentist), they may assume the presence or absence of autocorrelation, which is frequently present in SCD data and it is yet to examine how the ES and standard error estimates from these three ES indices are affected by overdispersion, a common occurrence in count data. These fundamental differences call for a closer examination and comparison of the methods and estimates obtained. The proposed dissertation aims to investigate the interpretability and understandability of the estimates produced as proposed by May (2004), examine if the three ES indices can be converted to a common metric to facilitate comparison of the ES estimates, document the benefits and challenges while implementing each method, and examine the performance of these ES methods under positive autocorrelation and overdispersion using Monte Carlo Simulation. Schmidt (2007), a published SCD study that examined the effect of Class-wide Function-related Interventions Teams (CW-FIT) on reducing the disruptive behavior of three first grade students using an ABAB

design, was used to examine the interpretability and understandability of the estimates produced and whether the indices can be converted to a common metric. It consisted of 3 cases with 4 phases (ABAB) for each case. For the simulation study, 1000 datasets for each case were simulated using pre-specified data parameters (number of cases, number of data points within each phase of a case, and phase means) taken from Schmidt (2007) study and for various conditions of autocorrelation and overdispersion. A fully crossed factorial design with three autocorrelation (0.0, 0.2, 0.4) and four overdispersion (0.0001, 0.05, 0.1, 0.3) resulting in 12 simulation conditions for each case was used for the data generation purpose. All analyses were carried out in R software. Results indicate all three ES estimates are interpretable. LRR meets the understandability criteria, however both BRR and NLB require advance statistical knowledge to run the models. The three ES can be converted to a common metric because they are ratios of the mean count of the phases. Based on simulation, all the three methods produce almost unbiased estimates of the effect size under different data conditions, however the standard error is affected by autocorrelation and overdispersion. This dissertation can serve as a resource for other SCD researchers and applied practitioners to understand and interpret the different ES values from the LRR, NLB, and BRR methods and help them make better informed decision about which of the three ES indices to use in their own research study if there is presence of autocorrelation and overdispersion in their data.

Chapter 1

INTRODUCTION

1.1 Single-Case Designs

Single-Case Designs (SCDs) are a form of interrupted time-series design that provide a rigorous methodological evaluation of treatment effects (WWC, 2022a; Kratochwill et al., 2010). They are used to identify and study effective interventions (Biglan, Ary, & Wagenaar, 2000; Flay et al., 2005) and to document evidence-based practices (Byiers, Reichle, & Symons, 2012; Horner et al., 2005). This design has several alternate names, including single-case research design (Kazdin, 1982), single-subject research (Horner et al., 2005), n-of-1 trials (Gabler, Duan,Vohra, & Kravitz, 2011), and single subject design (Olive & Smith, 2005).

In SCDs, an individual case is the unit of intervention and unit of analysis where a case can be a single subject or a single entity, each case serves as its own experimental control, the outcome variable is repeatedly measured within and across different phases resulting in sequential response data, and the independent variable is actively manipulated by the researcher (Smith, 2012; WWC, 2022a) (See Figure 1 for an illustration of a SCD with two phases). SCDs are widely used in many fields including Education and Psychology, to investigate whether a causal relationship exists between the researcher-manipulated independent variable and outcome variable.

Traditionally, randomized controlled trials (RCTs) are considered the “gold standard” for documenting causality, and SCDs are sometimes perceived as being

inferior (Kratochwill et al., 2013). However, SCDs can provide a scientifically rigorous alternative to RCTs and other group designs to determine treatment effectiveness (Kratochwill & Levin, 2014) and allow investigation of a variety of research questions, outcomes, settings, cases, and independent variables (Kratochwill & Levin, 2010). Generally, SCDs serve as a design of choice when treatment needs to be tailored for an individual case as opposed to a group (Geirut, Morrisette, & Dickinson, 2015; Shadish, 2014), when incidence cases are low (Wilson, 2011), and as a pilot to demonstrate proof of concept before conducting a larger experiment (Shadish, 2014).

Proper designing and execution of SCD studies provides strong internal and external validity (Lobo, Moeyaert, Cunha, & Babik, 2017). Replication and/or randomization can improve the internal validity of SCDs (Kratochwill & Levin, 2010). Replication of results within a SCD study controls for the most common internal validity threats like maturation and history (Kratochwill et al., 2010; Cannon, Guardino, Antia, & Luckner, 2016). The external validity in SCD is improved by replication of the intervention effects across various cases, various settings, and/or various dependent variable measures (Horner et al., 2005). Kazdin (1982) notes that the ultimate test to ensure generalizability of findings is through replication.

The most basic form of SCD consists of a baseline phase (A) and a treatment phase (B). In the simplest AB design, there is only one interruption when the treatment is introduced (see Figure 1). The outcome variable is repeatedly measured over time across the two phases. The baseline is the business-as-usual phase without any intervention. Observations are gathered until a stable baseline is established. Generally, the baseline stability is determined by limited variability in the baseline

data and lack of a clear trend, especially in the direction that is expected in the treatment phase (Byiers, Reichle, & Symons, 2012). Observations following the intervention comprise the treatment phase. Finally, observations in the two phases are compared to demonstrate an intervention effect. It is important to note that an SCD with only one demonstration of an intervention effect is insufficient to infer evidence of causality (WWC, 2022a; Kratochwill et al., 2010). This is because it is difficult to rule out alternative explanations for the observed change in the outcome variable (Kratochwill & Levin, 2010) with just one demonstration. Nevertheless, an AB design is a basic building block of other SCD variants.

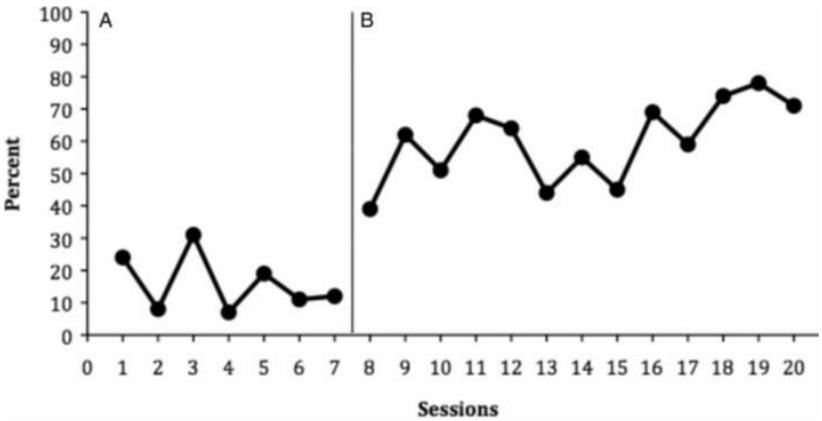

Figure 1: An example of AB design using hypothetical data.

Some of the SCD types are withdrawal or reversal (ABA or ABAB), changing criterion, multiple baseline/multiple probe designs (Byiers, Reichle, & Symons, 2012), and alternating treatments designs (Smith, 2012). Kratochwill & Levin, (2010) have classified the major SCD types into within series (e.g., AB, ABAB), between series

(e.g., alternating treatment design), and combined series (e.g., multiple baseline). In the within series designs, the outcome is measured within each condition and compared between or among the conditions. In the between series, the design allows the researcher to compare two or more conditions (e.g., two different interventions) in shorter period of time than the within series design. With combined series, both within series and between series comparisons are made to examine the treatment effect. Depending on the research question and other factors, one can choose the type of SCD to be used. For example, a multiple baseline design (MBD) is most appropriate when it is not possible for participants to return to the original baseline conditions after the treatment is introduced (Hersen & Barlow, 1976). See Figure 2 for an MBD illustration, and Figure 3 for an ABAB illustration. Once the functional relationship is established between the dependent and independent variables, the next step is to quantify the treatment effect.

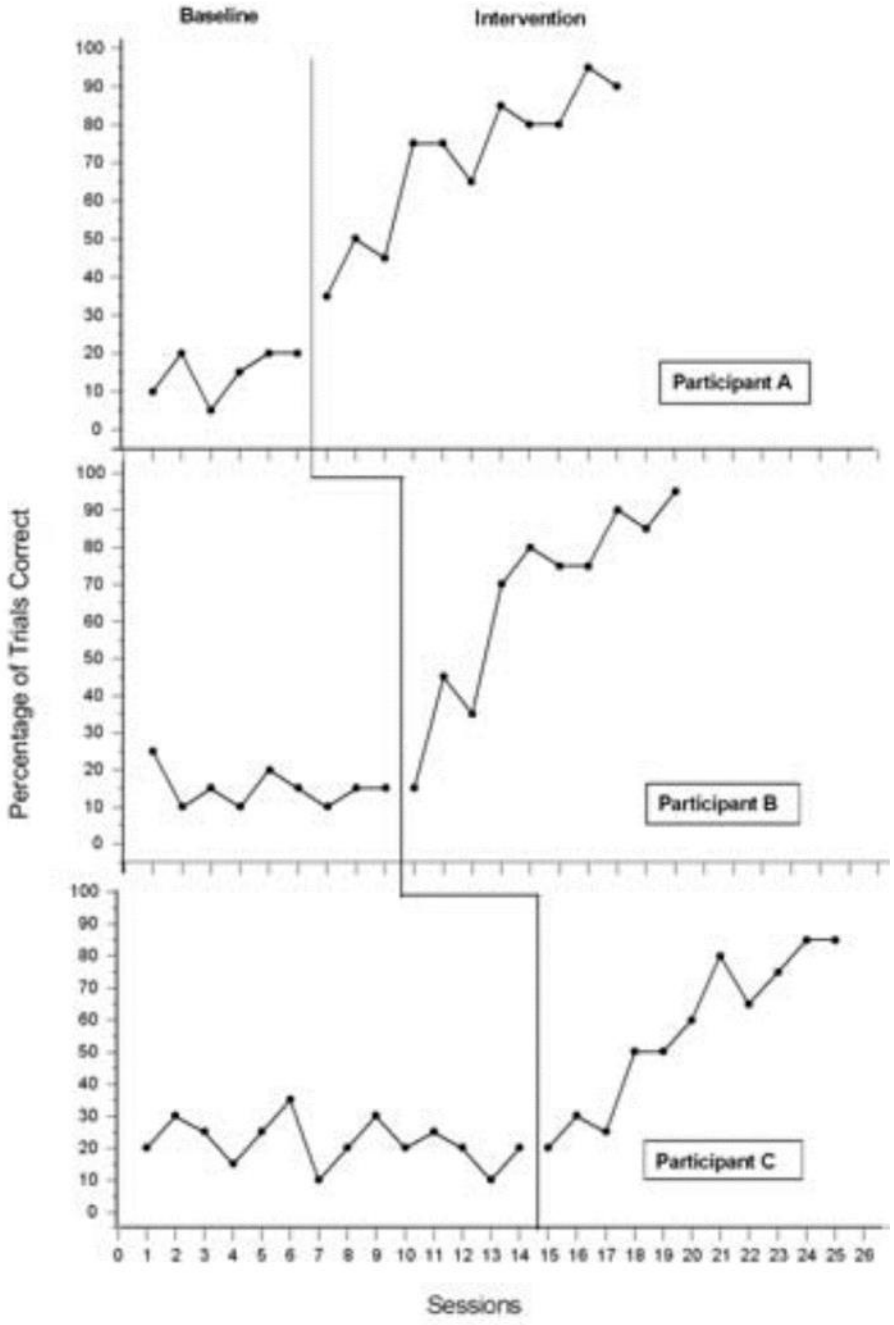

Figure 2: An example of Multiple Baseline Design.

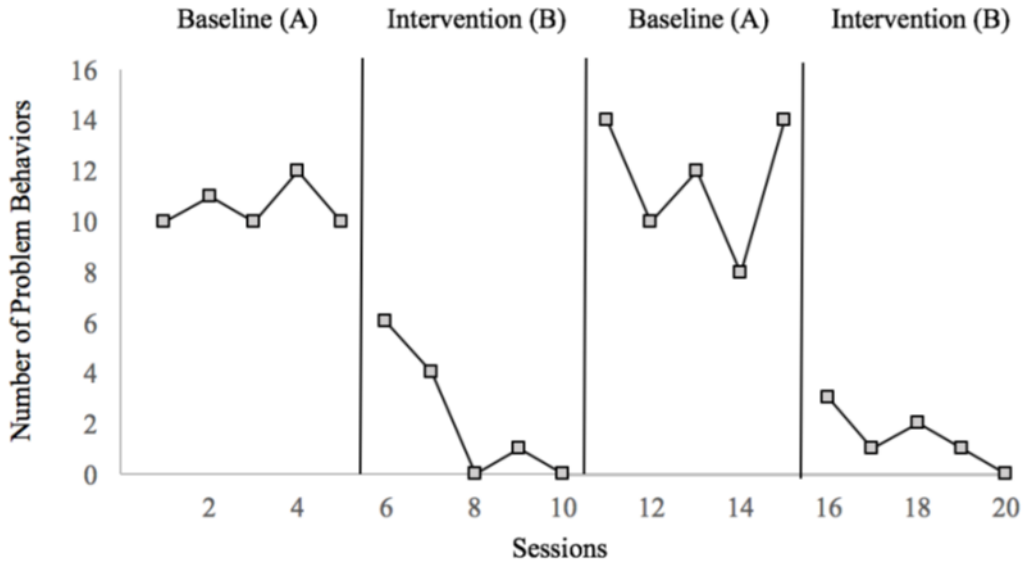

Figure 3: An example of ABAB design.

1.2 Why measure Effect sizes?

Effect sizes (ES) are quantitative indices of the strength of relationships among variables (Hedges, 2008). They provide meaningful expression to the magnitude of the treatment effect and are invariant to the choice of scale (Jenson, Clark, Kircher, & Kristjansson, 2007; Hedges, 2008). Effect sizes allow cross-comparisons of individuals, experimental conditions, populations, and studies (Gierut, Morrisette, & Dickinson, 2015). They are widely used in meta-analysis (MA). In a MA, findings from individual studies are statistically accumulated into a review summary. That is, results across studies are combined to a common metric known as effect size (Glass, 1978; Slavin, 1984). MA allows statistical cross-comparisons of different studies to

establish the efficacy of a treatment. These in turn inform evidence-based practices (Dollaghan, 2007; Law, Garrett, & Nye, 2004). MA is considered as an acceptable method to identify evidence-based practices (Therrien, Zaman, & Banda, 2011). For example, Institute for Education Sciences (IES) lists MA in the top tier of evidence for "What Works in Education" (Kame'enui, 2006).

Clearly, effect sizes are important because they quantify the treatment effect and provide opportunity for meta-analysis which in turn provides support for evidence-based practices.

1.3 Effect Sizes for Single-Case Designs

A number of effect sizes are available for between-group designs. Some of the common ones are Cohen's *d*, Hedges' *g*, and Glass's delta, each of which are variants of a standardized mean difference. However, these existing between-subject effect sizes cannot be directly used for SCD studies. For example, the commonly used standardized mean difference ES does not account for autocorrelation. This might result in inaccurate standard errors and inflated Type I (Brossart, Parker, Olsen, & Mahadevan, 2006; Gage & Lewis, 2013). Autocorrelation is defined as the serial dependency of error terms (Shadish & Sullivan, 2011; Natesan, 2019). As the data are collected over time for a case, the within participant data are not independent. According to Busk & Serlin, (1992) the computation of the effect size for SCD must be specific to the single-subject design because there are differences in the underlying assumptions of within-subject versus between-groups comparison. Moreover, SCDs have small sample sizes and thus effect sizes for SCDs need to cater to these characteristics of SCDs as pointed out by Smith (2012). According to his paper, the

main challenges associated with SCD design are presence of autocorrelation in the data, interpreting effect sizes, and small sample size.

To this end, several methods have been developed for SCDs to evaluate the effectiveness of treatment and estimate effect sizes. Broadly, the ES in SCDs are non-overlap based, regression-based, and more recently based on Bayesian methods. Most of these proposed methods in SCDs are developed for continuous outcome data and they generally assume a linear model (Hedges, Pustejovsky, & Shadish, 2012; Van den Noortgate & Onghena, 2008) including the between-subject effect size for SCDs. However, published literature suggests some form of count data are the most common outcome in SCD studies. Shadish & Sullivan (2011) in their review article of 809 individual SCDs across 113 studies found that 92.9% had some form of count outcome data.

Count outcomes are basically the number of occurrences of an outcome (e.g., behavior or an event) in a fixed period of time; for example, the number of times a student demonstrates on-task behavior during an observation period. A count outcome can take only positive integer values because an occurrence of a behavior or event cannot occur a negative number of times (Coxe, West, & Aiken, 2009). If the outcome variable is not continuous but rather a binary or count variable, then a linear model is not appropriate. Generalized linear models, including logistic regression model and Poisson count models are more appropriate for such outcomes (Swaminathan, Rogers, & Horner, 2014). Thus, to analyze count outcome data in SCDs, appropriate statistical models and related effect size metrics are needed.

Only in recent years have methodologists proposed effect sizes specifically for count-based outcomes in SCDs (Rindskopf, 2014; Pustejovsky, 2018). Three of the

recently proposed ES for count outcome data are: the Nonlinear Bayesian (NLB) effect size (Rindskopf, 2014), the Bayesian Rate Ratio (BRR) effect size (Natesan Batley, Shukla Mehta, & Hitchcock, 2021), and the Log-Response Ratio (LRR) effect size (Pustejovsky, 2018). The NLB effect size method calculates both within-subject effect sizes and an effect size across all the subjects for count SCD data without accounting for autocorrelation. This method uses multilevel modeling (MLM) to analyze the data with Bayesian estimation of model parameters. Also, the method provides functional forms for both count and proportion data. The BRR effect size calculates a within-subject effect size for count data using a Poisson regression model. This method accounts for autocorrelation and also uses Bayesian estimation. The LRR effect size is based on a frequentist framework. It calculates the within-subject effect size assuming the SCD data are independent (i.e., without accounting for autocorrelation). The LRR can be used with outcomes that are measured on a ratio scale, and thus is appropriate for both count and proportion data. All the three methods assume absence of a time trend in the baseline and treatment phases (i.e., the outcomes for a given subject are stable, not trending upwards or downwards, within each of the phases). All three methods calculate ES for some form of count outcome data, but they either use different statistical modeling, different frameworks (Bayesian or frequentist), or they assume presence or absence of autocorrelation in the data. Both BRR and NLB methods use a Poisson model and assume the data are Poisson distributed. A key property of Poisson distribution is it has a single parameter representing both its mean and variance (i.e., it assumes both mean and variance to be equal). In reality, this might not always be the case. When the variance of count data is larger than the mean, it is known as overdispersion (Agresti, 2005). If the count

outcome data are overdispersed, this might be an issue when using the Poisson distribution if there is no separate parameter (for the variance) to describe variability. The effect of these probable SCD count data characteristics (autocorrelation and overdispersion) when they are recognized versus ignored is an important question to be answered. Thus, a closer examination and comparison of the methods and estimates obtained will be helpful for potential users of these methods; namely, researchers planning an SCD study.

1.4 Importance of Meaningful Statistics

For any research study, after the statistical analysis is conducted and the estimates are obtained, the next step is to present the results in a meaningful way. May (2004) provided a general approach to presenting statistical information meaningfully in the context of policy and evaluation research. These guidelines are applicable in any research field in the social sciences. To be able to meet these guidelines, answers to the following questions need to be sought: Do the audiences (consumers of the study results) need to know specific statistical concepts to understand the information presented; is the statistical information presented in familiar metrics/units; and are the magnitudes of the different estimates directly comparable within and across studies? In short, are the statistics understandable, interpretable, and comparable?

For the aforementioned criteria to be met, the first step is that the researcher using a particular method in their study consider the statistical underpinnings of the method in order to employ the method and communicate the findings in an understandable manner. However, given that SCD is used in diverse fields, a researcher or applied practitioner using a SCD may or may not understand relevant statistical methods. This may hinder the researchers in using methods that are

appropriate to answer their research questions but demand a higher level of statistical knowledge for their implementation. These days, many methods provide computer analysis codes (R programs are common) that simplify analytic processes. However, if the method uses metrics that are not easy to understand, then this might also limit SCD users' access to and use and interpretation of appropriate methods. Based on May's (2004) understandability criteria, if the statistics obtained from a method are relatively simple (e.g., averages and proportions), then they are more easily understood by wider audiences and one can say the results obtained from the method itself are understandable. In terms of interpretability, if the statistical method used provides some or all of the statistics in metrics that are familiar to the general audiences (e.g., rates, percentage change) then it can be said that the interpretability criteria is met. For this proposed dissertation, the comparability criterion is especially relevant because it speaks to the primary goal behind effect size calculation (i.e., making comparisons across studies, interventions, etc.), which aligns with the goal of this dissertation. Therefore, it's important to examine whether the magnitude of the estimates (comparing the numerical estimates and associated standard error (SE)) from the three different ES methods are directly comparable even though the details of their calculation are different (i.e., do they mean the same thing, and are they on the same scale?). Moreover, an evaluation of the estimates obtained from the NLB, BRR, and LRR effect size methods using the May (2004) guidelines puts emphasis on the need to present statistical results in a manner that is meaningful to different audiences. It helps to assess whether the statistics obtained from the three ES methods are already understandable, interpretable, and comparable, or whether additional computations must be carried out by the researcher to reformulate the statistic so that it is

meaningful. According to May (2004), the key to meeting these guidelines is "to rethink the way statistics are presented and present the results in ways that capitalize on commonly understood statistical metrics" (p. 528).

1.5 Purpose of the Proposed Study

Given that SCDs provide an empirical approach to inform evidence-based practices and are a scientifically rigorous alternative to randomized controlled trials and other group-based designs, a conceptual and empirical comparison of these newly proposed ES methods for count outcomes is warranted. Moreover, implementation of these different methods in existing SCD studies (i.e., using them in the same study) and examination of the parameter estimates obtained has yet to be conducted. Two of these methods (NLB and BRR) use complex statistical modeling and estimation processes, which might not be accessible to all the researchers who could use these methods in their own study. The main purpose of this dissertation is to help other SCD researchers and applied practitioners understand and interpret the different ES values from the LRR, NLB, and BRR methods, so that they can make better informed decisions about which one to use in their own research study. The reasons for focusing on these three effects sizes are: they are proposed for some form of count outcome data (count, proportion, rates) commonly used in SCDs, all the three methods assume the absence of a time trend (i.e., stable baseline or treatment phases), they can be used with AB or ABAB designs, and each effect size is calculated based on the mean or average levels of the phases. Thus, direct comparison of these effect sizes, or transformations thereof, is possible. Two of the methods do not account for autocorrelation while the BRR method does. Thus, a comparison of the ES under different levels of autocorrelation can illustrate the performance of the different ES

estimates. Thus, it is important to empirically examine how the estimates and standard errors behave when autocorrelation and overdispersion is present in the data. Moreover, examining the understandability and interpretability of the effect size estimates from these three methods are important because they contribute to making the statistical findings from SCDs more meaningful.

With this background, the purpose of this dissertation is to answer the following questions:

1. In what ways are the effect sizes proposed for count outcome data in a single-case study similar and different in terms of assumptions and interpretations?
    a. To what extent do the estimates from LRR, NLB, and BRR satisfy the understandability and interpretability criteria proposed by May (2004) for meaningful statistics?
    b. How can each effect size be transformed in order to enable direct comparisons of the estimates?
2. To what degree should researchers be aware of potential problems with using these effect sizes when overdispersion and/or autocorrelation are present in count outcome data in a single-case study?
    a. How different are the LRR, NLB, and BRR effect size and standard error estimates when used with overdispersed count outcome data in a single-case study?
    b. How robust are the ES estimates to lag-1 autocorrelation commonly seen in SCD outcomes data?

3. What are the benefits and challenges associated with implementing these ES indices with a real SCD dataset?

1.6 Study Implications

The most important implication of this study is it may help SCD users who are not very familiar or comfortable with the statistical information used or interpretation of the estimates from the three ES methods (LRR, BRR, and NLB) to be better equipped and more confident in implementing these methods and interpreting the estimates confidently. This dissertation will use data from a real SCD study, including details of how the ES is estimated using the three methods (including annotated R code), and it will provide illustrative interpretation of the estimates. This is important because applied researchers may not be able to understand the traditional methodological publications that present advanced statistical concepts and formulae underlying these methods, despite the fact that these methods can be described and interpreted in simple, intuitive, and less technical ways. The key to enabling use of these methods by applied researchers is to show the results from these methods with direct and simple meaning that requires minimal statistical expertise. This dissertation aims to achieve this. The details that will be provided in this dissertation may serve as a resource to applied SCD users to more easily understand the estimates from the methods and help them make better informed decisions about which one to use in their own research study.

This proposed dissertation will be the first study to compare results from the LRR, BRR, and NLB ES methods and to examine the effect of positive autocorrelation and overdispersion on the ES estimates and standard error from these methods. The findings from this dissertation might also serve as a resource for methodologists and applied researchers on the importance of making the estimates from their methods minimally complicated and intuitively understandable.

Chapter 2

LITERATURE REVIEW

This chapter provides information on two of the common single-case design types, some of the existing effect sizes in SCDs, details of the LRR, BRR, and NLB ES methods used in this dissertation, autocorrelation in SCDs, and overdispersion in count outcome. But first, a brief overview on existing single-case design standards for reviewing a SCD study is discussed.

2.1 Existing Single-Case Design Standards

Over the years, SCD experts, researchers, and organizations have developed standards and guidelines to analyze SCD studies. Some of the widely used standards are; the What Works Clearinghouse pilot standards for single-case designs (Kratochwill et al., 2010), those suggested by the National Reading Panel (NRP; National Institute of Child Health and Human Development, 2000), the single-case experimental design scale (Tate et al., 2008), the single-case reporting guideline in behavioral interventions (SCRIBE) (Tate et al., 2016), and the Council of Exceptional Children's Standards (CEC; Cook et al., 2014). The WWC standards for SCDs, now integrated into WWC procedures and standards version 5.0 (WWC, 2022a) are widely used in Education and related fields.

2.2 What Works Clearinghouse Standards for SCD Studies

WWC, in its procedures and standards handbook (WWC, 2022a) provides research rating standards to evaluate the evidence from SCDs to determine if they receive “Meet WWC Standards Without Reservations”, “Meets WWC Standards With Reservations”, or “Does Not Meet WWC Standards” research ratings. These standards examine the rigor and execution of a SCD study. A step-by-step guide for the review process of this standard is shown in Figure 4. Only if the criterion in a step is satisfied it can move to the next. Studies that satisfy all these steps (e.g., outcome measures standards, data availability, independent variable is systematically manipulated by the researcher, design assessment, limited risk of bias, etc.) are classified as studies that “Meet WWC Standards Without Reservations”.

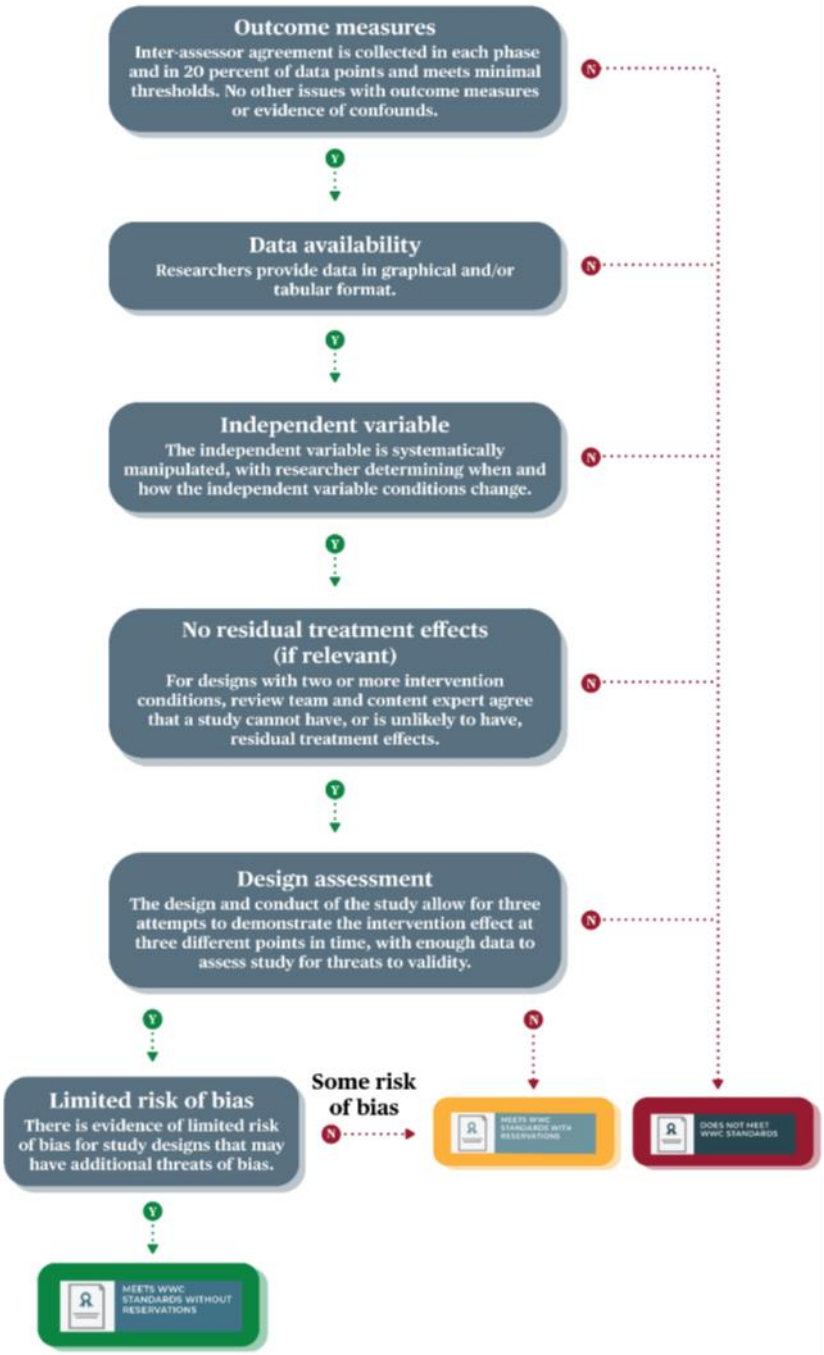

Figure 4: Single-case design review process for eligible study findings.

They also provide guidelines on the effectiveness rating, that is used to characterize evidence of the treatments' effect. Only studies that "Meet WWC Standards Without Reservations" or "Meets WWC Standards With Reservations" are examined for effectiveness rating of the interventions. The effectiveness rating tells us whether the treatment affected change in the outcome. WWC uses effect size and statistical significance to determine evidence of treatment's effectiveness.

As mentioned in the Introduction chapter, an effect size is a numerical index which quantifies the direction and magnitude of a causal relationship between a treatment and outcome (Pustejovsky, 2018). To compute statistical significance first standard errors must first be estimated. These standard errors capture the uncertainty in estimated effect sizes. Larger standard error suggests less precise estimates. For a study to be statistically significant, it must have a large effect size relative to the standard error (WWC, 2022a).

WWC handbook provides details on how to calculate the design-comparable effect size (Pustejovsky, Hedges, & Shadish, 2014) for SCDs, that is comparable with standardized mean difference effect sizes (Hedge's g) used in group designs. This effect size is most appropriate for continuous outcome. However, published literature shows that most of the outcome in SCDs are counts (Shadish & Sullivan, 2011; Rindskopf, 2014; Pustejovsky, 2018; Declercq et al., 2019). WWC document states that "it aims to report SCD study findings in a consistent way, using a common metric and accounting for differences across analyses that may affect their results". The present dissertation can contribute towards this endeavor in SCDs when the outcome variable involves count data by examining whether the three different count-based ES

can be converted to a common metric and thus compare the effect size estimates and precision from these three methods. In practice, it might not always be feasible to calculate the design-comparable effect size and a study might report its own effect size and statistical significance. Having said this, two of the common SCD types are now discussed.

2.3 Multiple Baseline and Treatment Reversal Designs

A multiple baseline design investigates the functional relationship of treatment variable on the same outcome variable across different cases (at least 3 cases) (see Figure 2). The treatment is introduced in a staggered manner for the cases. Each case's outcome is measured repeatedly over time in a baseline and in a treatment condition. Once the stability of the baseline is established, one case receives the intervention while the baseline is maintained for other cases. Only after an improvement is seen for the first case, an intervention is provided to the second case and so on. The idea here is that the change in the outcome is observed only after the treatment is introduced to a case. If there is change in the outcome of other cases although they were in baseline condition, then one can infer that the treatment is probably not causing the observed change in the outcome (Barger-Anderson, Domaracki, & Kearney-Vakulick, 2004).

Treatment reversal designs are also commonly known as ABAB design. This design consists of two baselines and two treatment phases and the phases are alternated (see Figure 3). There are three opportunities to demonstrate experimental control, first when the treatment is introduced (first AB phase), secondly, when the treatment is withdrawn and returned to the baseline condition (BA phase), and when the treatment is re-introduced (AB phase). ABAB designs can answer questions related to the effect of a single intervention (Byiers, Reichle, & Symons., 2012).

2.4 Overview of some of the Single-Case Design Effect Sizes

Traditionally, results from SCDs have been analyzed using Visual Analysis (VA). As the name suggests, VA involves viewing and inspecting the SCD time-series data points presented in graphs, and determining whether there is change in the outcomes across the phases based on visible data characteristics (Ledford, Lane, & Severini, 2017). The WWC panel (Kratochwill et al., 2010) has recommended using both visual analysis and statistical methodologies in analyzing SCD data. VA can determine whether there is a clear effect of the treatment, and the magnitude of the effect can be calculated using an effect size (Ledford, Lane, & Severini, 2017).

Before discussing the three ES methods for count data used in this dissertation, some of the other common SCD methods are discussed. Over the years, methodologists have developed various methods to analyze SCD data. In general, techniques based on nonoverlap of data points, frequentist methods, and Bayesian methods have been proposed. Each approach has its advantages and disadvantages, and there is no consensus on the most appropriate method to calculate the effect size in SCDs (Gage & Lewis, 2013).

2.4.1 Non-Overlap Indices

Non-overlap indices are common techniques used to provide an effect size estimate in SCDs. Generally, the non-overlap methods are based on observing the number of data points above (or below) the highest (or lowest) data point in the baseline phase (Gage & Lewis, 2013). Some of the common non-overlap methods are percentage of non-overlapping data (PND), percentage of all non-overlapping data (PAND), non-overlap of all pairs (NAP), and Tau-U. These non-overlap methods are quite popular because they are easy to calculate, interpret, augment visual analysis of

SCD data, and claim to be distribution-free (Parker, Vannest, & Davis, 2011). However, several limitations exist in using these methods for analyzing SCD data. Most of these methods use few data points, e.g., the PND method uses only one data point, most are insensitive to trends, and have insufficient statistical power to detect smaller effects (Parker, Vannest, & Davis, 2011). A study comparing four non-parametric overlap methods (PND, pairwise data overlap squared (PDO), percentage of data exceeding the median (PEM), and percentage of data exceeding a median trend (PEM-T)) suggested avoiding these methods altogether given their findings and limitations (Wolery, Busick, Reichow, & Barton, 2010).

2.4.2 Frequentist Methods

One of the earliest, regression-based methods was proposed by Center, Skiba, and Casey in 1986. They proposed a piece-wise regression model to calculate the effect size. The model could test change in both intercept and slope and assumed that the observations are independent (Center, Skiba, & Casey, 1986, Shadish & Rindskopf, 2007). Other developments using frequentist methods include using statistical modeling techniques such as multilevel modeling (Van den Noortgate & Onghena, 2003a; 2003b; 2007; 2008)) and randomization of SCD studies (Kratochwill & Levin, 2010). Each of these methods has its own strengths and challenges. For example, MLM can be used to fit three-level models where observations are nested within cases, and cases nested within studies (Van den Noortgate & Onghena, 2007) and provide both between-subject and within-subject effect sizes. However, challenges such as inflated type I error rates, autocorrelation, trend, small sample size, and short data streams can complicate the MLM methods even though these seem very promising (Muller et al., 2007). Hedges, Pustejovsky, & Shadish, (2012) proposed a

between-subject standardized effect size that is comparable to the between-groups standardized mean difference. Like many of the existing SCD effect sizes, this method also assumes the SCD outcome data to be interval-scaled and the model assumes normally distributed errors. However, if the outcome data is count or proportion, a model based on normally distributed errors might not be appropriate (Rindskopf, 2014).

2.5 Bayesian based Effect Size Indices

In recent years, Bayesian methods are being used to analyze SCD data and calculate effect sizes. Major advantages of Bayesian approach are it does not depend on large sample theory (Ansari & Jedidi, 2000; Swaminathan, Rogers, & Horner, 2014; Natesan & Hedges, 2017), produces more precise estimates of autocorrelation (Shadish, Rindskopf, Hedges, & Sullivan, 2013), and provides detailed information on the parameter estimates (Kruschke, 2015). The downside is that Bayesian methods have a steep learning curve (Natesan, 2019). Swaminathan, Rogers, & Horner (2014) used Bayesian analysis to provide a standardized mean difference effect size for a linear model that takes into account changes in intercepts and slopes in the presence of autocorrelation. This effect size is comparable to the between-groups standardized mean difference. Moeyaert, Rindskopf, Onghena, & Van den Noortgate (2017) compared Bayesian estimation of multilevel modeling and maximum likelihood estimation of SCD data. The Bayesian Unknown Change-point model proposed by Natesan & Hedges (2017), examined immediacy in SCD studies for interval-scaled outcome data. Natesan Batley, Nandakumar, Palka, & Shrestha (2020), compared two simulation-driven approaches, the Bayesian Unknown Change-point model and Simulation Modeling Analysis for SCD datasets showing “clear” immediacy,

“unclear” immediacy, and delayed effects. All of these methods assume a linear model.

However, many studies have found that count or proportion data are more common in SCDs (Shadish & Sullivan, 2011; Rindskopf, 2014; Pustejovsky, 2018; Declercq et al., 2019). Yet a linear model is not generally suitable to analyze count data; a Poisson model is a more appropriate generalized linear model for such data type (Swaminathan, Rogers, & Horner, 2014). If the outcome variable is count-based then the simplest appropriate distributions are the Poisson (for a count over a period of time) and the binomial (for a count out of a fixed number of trials) (Rindskopf, 2014). A study by Declercq et al, (2019) compared the performance of the linear mixed model (LMM) with generalized linear mixed model (GLMM) to analyze SCD count data. Compared to the LMM, the goodness of fit and power were better for the GLMM.

Thus, the focus of this dissertation is to examine and compare three recently proposed SCD effect sizes developed specifically for count outcome data and interpret the estimates.

2.6 Effect Sizes for Count Outcomes in Single-Case Designs

Three of the recently proposed SCD effect sizes for count outcome data: Log Response Ratio (LRR), Non-Linear Bayesian (NLB), and Bayesian Rate Ratio (BRR) are described in the following sections. NLB provides both within-subject and between-subject effect sizes, while LRR and BRR are both within-subject effect sizes. Both NLB and BRR use Bayesian estimation, and in the next section, the Bayesian framework is described. Following this, the mathematical models for LRR, NLB, and BRR are described.

2.6.1 Bayesian Framework

Bayesian methods are based on the Bayes theorem (a.k.a., Bayes rule) developed by Thomas Bayes (1701-1761). The Bayes rule provides the relationship between marginal and conditional probabilities. According to Bayes theorem, the probability of event A occurring, given that event B has already occurred is as follows:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \qquad (1)$$

If A is a parameter and B is the data, then the probability distribution of the parameter given the data (P(A|B)) is proportional to the probability distribution of observing the data given the parameter values, weighted by the probability distribution of the parameter (P(A)) (Swaminathan, Rogers, & Horners, 2014). In Equation 1, P(A|B) is the posterior distribution of the parameter, P(B|A) is the likelihood of the data given the model parameter; P(A) is the prior distribution, and P(B) is the normalizing constant. The prior distribution describes the prior knowledge about the parameter before collecting the data (e.g., based on results from prior studies or assumptions about anticipated estimates), and the posterior distribution combines the prior information with observed data to update our knowledge of the parameter (Agresti, 2015).

In the Bayesian approach the parameters are treated as random variables while in “frequentist” or classical approach, parameters are considered fixed (Swaminathan, Rogers, & Horners, 2014). In frequentist methods, generally a point estimate is calculated with an associated standard error. With Bayesian methods, posterior distributions of the parameters are estimated. Because the full distributions of parameters are estimated, there is a probability value associated with each possible value of the parameter (Natesan & Hedges, 2017). Thus, Bayesian estimation provides

more detailed information about the parameter estimate. Using the posterior distribution one can compute any summary statistic like the mean and mode (representing different versions of an expected value estimate) and visually examine the density plot of the distribution of parameters.

Similar to confidence intervals, high density intervals (HDI) can be calculated in a Bayesian framework. The definition of HDI is more straight forward and intuitive compared to confidence intervals (Kruschke and Liddell, 2018). A 95% HDI means there is a 95% probability that the true parameter value lies within that interval. All the values inside the HDI have higher credibility compared to values outside the interval (Kruschke, 2015). Also using Bayesian statistical significance testing, one can reject or accept the null hypothesis (Kruschke, 2013), whereas in the frequentist approach one can only reject the null hypothesis.

In Bayesian inference, formulas for Bayes' rule become intractably complicated when many variables and parameters are involved. To work around this, Markov Chain Monte Carlo (MCMC) techniques are used because they do not require closed-form versions of the equation for Bayes' theorem to produce a posterior. MCMC helps the user to approximate aspects of a posterior distribution that cannot be directly calculated. Thus, MCMC simulation methods are generally used for Bayesian computations to obtain samples from the posterior distributions of the parameters (Rindskopf, 2014). In MCMC, the Markov chain property generates random samples by a sequential process such that a new random sample is generated based only on the one sampled before it and not on any other samples. Monte Carlo establishes the properties of the estimates by examining random samples from the posterior distribution (Van Ravenzwaaij, Cassey, & Brown, 2018). Generating large numbers of

random samples allows us to approximate the true posterior distribution from which the samples are generated (Kruschke, 2015). Gibbs sampling introduced by Geman and Geman (1984) is a commonly used MCMC process in Bayesian estimation that generates samples from the posterior distribution. This sampling method follows an iterative procedure to repeatedly sample from the conditional distribution of one variable given all of the other variables. A generic Gibbs sampler is described in Appendix A. Next each of the three ES methods are described.

2.6.2 Log Response Ratio Effect Size

Pustejovsky (2018) proposed the Log Response Ratio effect size, a within-subject effect size for single-case designs. It is calculated as the natural log of the proportionate change from baseline to the treatment phase. In cases involving more than two phases (e.g., an ABAB design), LRR is calculated for each adjacent phase (e.g., two LRR estimates for an ABAB design, say LRR1 for A1B1 and LRR2 for A2B2), and then a composite LRR is calculated by taking the average of these two LRR estimates. LRR can be used for data measured on a ratio scale and for outcomes such as counts, rates, proportions, and percentages. The LRR assumes no time trend (i.e., the level of the outcome variable is stable within each phase, and autocorrelation is absent).

While using the LRR effect size for multi-level meta-analysis, robust variance estimation can provide valid assessments of uncertainty even when the sampling variances of effect size estimates are not accurate (Chen & Pustejovsky, 2022). However, robust variance estimation has not been used for individual SCD studies, and autocorrelation can lead to biased estimates of the sampling variance of the LRR effect size. Thus, one should be careful in interpreting the standard error while

calculating the LRR effect size for individual SCD studies (Pustejovsky, 2018). The details of estimating the LRR effect size are described below.

For a SCD with two phases (AB), let the baseline phase have "*m*" sessions and treatment phase have "*n*" sessions. The outcome variable in baseline are $Y_1, Y_2, \ldots, Y_m$ and in the treatment are $Y_1, Y_2, \ldots, Y_n$. The mean levels in baseline and treatment are μ_A and μ_B respectively. The LRR effect size parameter is as follows.

$$\Psi = \ln\left(\frac{\mu B}{\mu A}\right) \quad (2)$$

Equivalently,

$$\Psi = \ln(\mu B) - \ln(\mu A) \quad (3)$$

In terms of percentage change of the outcome from baseline to the treatment, the LRR can be expressed as,

$$\%change = 100\% \times [\exp(\Psi) - 1] \quad (4)$$

Where, $\exp(\Psi) = \left(\frac{\mu B}{\mu A}\right)$

Pustejovsky (2015) suggested using truncated sample means because it takes account of the possible scenario when the sample means might be equal to zero and true means are positive. The truncated means are calculated as follows.

$$\widetilde{y_A} = \max\left\{\frac{1}{2Dm}, \ \frac{1}{m}\sum_{i=1}^{m} Y_i^A\right\} \quad (5)$$

$$\widetilde{y_B} = \max\left\{\frac{1}{2Dn}, \ \frac{1}{n}\sum_{i=1}^{n} Y_i^B\right\} \quad (6)$$

In the above equations $\widetilde{y_A}$ and $\widetilde{y_B}$ represent baseline and treatment truncated means respectively. *D*, a constant that depends on the scale of the dependent variable. For count outcome variable, *D* equals 1 and for proportion outcome variable, *D* equals the number of intervals. Because of these values of *D*, the truncated mean is invariant to the changes of scale. Accounting for the small-sample bias, the bias-corrected LRR is calculated as,

$$R2 = ln(\widetilde{y_B}) + \frac{s_B^2}{2n\tilde{y}_B^2} - ln(\widetilde{y_A}) - \frac{s_A^2}{2m\tilde{y}_A^2} \qquad (7)$$

In the above equation, s_A and s_B are the sample standard deviations of the dependent variable from the baseline and treatment phases respectively. The corresponding estimate of sampling variance for bias-corrected LRR is given as

$$V^R = \frac{s_A^2}{m\tilde{y}_A^2} + \frac{s_B^2}{n\tilde{y}_B^2} \qquad (8)$$

and standard error is calculated as the square root of the sampling variance as shown

$$SE^R = \sqrt{V^R} \qquad (9)$$

An interpretation of the bias-corrected LRR effect size using a hypothetical example follows. Let us assume, effectiveness of a new treatment to reduce the number of disruptive behavior in a classroom is being studied using a single-case design. To determine the effect of this new treatment, 3 disruptive students from a third grade science classroom are selected and the study uses a MBD. The LRR estimate is discussed for only 1 student as an example. Let the phase length in the baseline equal 8 and in the treatment phase equal 10 for this student. Using the equations 5 and 6, let the estimated truncated means for baseline equal 10.32 (*sd* = 5.80) and 2.8 (*sd* = 1.08) for the treatment phase.

Using the formula for the bias-corrected LRR (R2) effect in equation 7, the estimated LRR effect size equals -1.3169 and the calculated percentage change equals -72.86%. The bias-corrected LRR estimate is negative suggesting decrease in the disruptive behavior of the student following the treatment. In fact, there is 72.86% reduction in the disruptive behavior of the student following the treatment.

2.6.3 Nonlinear Bayesian Effect Size

Rindskopf (2014) proposed using Bayesian methods for analyzing SCD data. In his paper, nonlinear Bayesian estimation for count outcomes and for continuous variables that require a nonlinear curve and/or with floor and ceiling effects are discussed. Since this dissertation is focused on count outcome for a fixed time, only the method for estimating ES of counts over a period of time (Poisson distributed data) namely, NLB will be used. However, the ES for count out of a certain fixed number of trials (binomial distributed data) namely, NLB_Bin is also discussed because the Rindskopf (2014) paper provides WINBUGS program codes only for NLB_Bin and I have referred to this code to build the code for Poisson distributed count data.

A multilevel approach is used to model the outcome of each case as a function of the phase. These models assume no time trend exists. Though these models can include the time of measurement as a predictor and predictors at level-2, models with only phase as predictor are described. This will aid in comparing the NLB effect size with other two effect sizes; BRR and LRR, which assume absence of a time trend.

First a general multilevel model with normally distributed outcome data is discussed. Let Y_{ij} be the outcome or response at time i, for person j. Dummy variable, X_{ij} denotes the phase for person j at time i; X_{ij} equals 0 in the baseline phase and equals 1 in the treatment phase. The residual term, r_{ij} is the difference between the observed value and expected value for person j at the ith time point.

The level-1 model is

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \qquad (10)$$

In the above equation, β_{0j} is the expected response/outcome for person j in the baseline phase. $\beta_{0j} + \beta_{1j}$ is the expected response/outcome for person j in the treatment phase. β_{1j} is the expected treatment effect for person j.

The level-2 model is

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (11)$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad (12)$$

where γ_{00} and γ_{10} are the fixed effects. γ_{00} is the average baseline level across all the cases and γ_{10} is the average treatment level across all the cases. The random effects are u_{0j} and u_{1j}. The unexplained variation among the cases are represented by the u_{0j} for baseline and u_{1j} for treatment effects. The corresponding variances are τ_{00} and τ_{11} respectively and covariance is τ_{01} or τ_{10}. The level-2 model is a simple model without any predictors such that the difference among cases is allowed but the model does not seek to explain the variation between the cases.

Because, the count data are not normally distributed, generalized linear models (GLiM) instead of general linear models are used. GLiM uses a link function (e.g., log, logit, etc.) to transform a potentially non-linear relationship between the outcome and predictors to a linear relationship (Coxe, West, & Aiken, 2009). The generalized hierarchical linear model for proportion data and count data is discussed next.

2.6.4 Nonlinear Binomial Model (a.k.a., Logistic Model)

Let the dependent variable, Y_{ij} is proportion data i.e., count out of a total. Y_{ij} follows a binomial distribution with number of trials, n_{ij}, probability of response/success on a given trial, π_{ij}, and mean $n_{ij}\ \pi_{ij}$

$$Y_{ij} \sim Bin(n_{ij}, \pi_{ij})$$

The level-1 model for proportion data is given as,

$$ln(\frac{\pi_{ij}}{1-\pi_{ij}}) = \beta_{0j} + \beta_{1j}X_{ij} \quad (13)$$

Where, the intercept, β_{0j} is the log-odds of the outcome variable during baseline phase; slope, β_{1j} is the change in the log-odds of the outcome variable between the baseline and the treatment phase i.e. the effect of the treatment, Say, phase $B = \beta_{0j} + \beta_{1j}$ is the log-odds of outcome variable during treatment phase. The level-2 model is the same as discussed for normally distributed outcome data. These parameters are estimated using Bayesian estimation. Rindskopf (2014) used WINBUGS (Lunn et al., 2000) for Bayesian estimation. A WINBUGS code that describes the model is included in Appendix B.

In Bayesian framework, using the information from the observed data, priors are updated to get the posterior. An example of uninformative prior specification for the parameters is as follows.

$$\beta_{0j} \sim norm(\mu_0, prec_0) \quad (14)$$

$$\beta_{1j} \sim norm(\mu_1, prec_1) \quad (15)$$

where precision (*prec*) = 1/variance.

Equivalently, standard deviation = 1/sqrt (precision)

Both the intercept and slope follow normal distribution with means μ_0 and μ_1, and precisions $prec_0$ *and* $prec_1$ respectively. The hyperpriors (i.e, prior on a prior) for means and precision is as shown

$$\mu_0 \sim norm(0,0.01) \quad (16)$$

$$\mu_1 \sim norm(0,0.01) \quad (17)$$

$$prec_0 \sim gamma(0.01,0.01) \quad (18)$$

$$prec_1 \sim gamma(0.01,0.01) \quad (19)$$

The means μ_0 and μ_1 follow normal distribution with means 0 and precision equal to 0.01. The precision follows gamma distribution with both shape and rate

parameters equal to 0.01. The posterior distribution of the parameters are approximated using an MCMC algorithm such as the Gibbs sampler. Though Rindskopf (2014) used WINBUGS, all the programming for the current dissertation is in R (R Core Team, 2021). To obtain parameter estimates in R, Just Another Gibbs Sampler (JAGS) (Plummer, 2003) will be used. JAGS automatically builds MCMC samplers for hierarchical models (Plummer, 2003, 2012). In R, the data and defined model is run using JAGS to obtain samples from the posterior distribution for the phase means, treatment effect, and standard deviations. Since the modeling is done in the logit (i.e., log-odds) scale, the parameters are transformed back to proportion as shown. The transformation of baseline mean and treatment mean respectively are as follows,

$$phase.A = \frac{exp(\mu 0)}{(1+\exp(\mu 0))} \tag{20}$$

$$phase.B = \frac{exp(phaseB)}{(1+\exp(phaseB))} \tag{21}$$

The nonlinear binomial (NLB_Bino) effect size is calculated as

$$phases.AB = phase.A - phase.B \tag{22}$$

Using the above transformations, posterior distribution of the parameters in their original form are also estimated. E.g., if the mean *phase.A* equals 0.88 then it means the expected probability of the outcome variable for a typical case in the baseline phase is 0.88. If the mean *phase.B* equals 0.11, then the expected probability of the outcome variable for a typical case in the treatment phase is 0.11. If *phases.AB* equals 0.77 then the expected reduction in the outcome variable for a typical case following treatment is 77%. The non-linear binomial model provides both across subjects and within-subject effect sizes and for each parameter provides several

summary outputs (e.g., mean, median, sd, 95% HDI). The program code is presented in Appendix B.

2.6.5 Nonlinear Poisson Model

Let the dependent variable Y_{ij} is count data (count over a period of time) for person j at time point i. Then, Y_{ij} follows a Poisson distribution with parameter η_{ij}. Using the generalized linear model with log link function, the level-1 model is now expressed in terms of the natural logarithm of the expected counts, η_{ij}.

$$ln(\eta_{ij}) = \beta_{0j} + \beta_{1j}X_{ij} \qquad (23)$$

where η_{ij} is the expected value of Y_{ij}.

In the above equations, β_{0j} is the average log count in the baseline phase, Phase B = $\beta_{0j} + \beta_{1j}$ is the average log count in the treatment phase. β_{1j} is the difference in average log count of the outcome between baseline and treatment phases. The level-2 model is similar to equations 11 and 12. γ_{00} and γ_{10} are the fixed effects. γ_{00} is the average baseline level across all the cases and γ_{10} is the average treatment effect across all the cases. The random effects are u_{0j} and u_{1j}, which are assumed to be distributed as a multivariate normal and independent between cases. The unexplained variation among the cases are represented by the variance of u_{0j} for baseline and u_{1j} for treatment effects. The corresponding variances are τ_{00} and τ_{11} respectively and covariance is τ_{01} or τ_{10}. These parameters are estimated using Bayesian estimation.

Similar to the Nonlinear Binomial model, the priors for the parameters can be specified and JAGS can be used to approximate the posterior distributions. Likewise, for the ease of interpretation, parameters are transformed back into the original scale using exponentiation. The average counts for a case j in baseline phase, treatment phase, and effect size are transformed as follows.

$$p.Aj = \exp(\beta_{0j}) \tag{24}$$

$$p.Bj = \exp(\beta_{0j} + \beta_{1j}) \tag{25}$$

The Nonlinear Poisson (NLB) effect size is calculated as,

$$p.AjBj = \exp[(\beta_{0j} + \beta_{1j}) - \beta_{0j}] = \frac{\exp(\beta_{0j})\exp(\beta_{1j})}{\exp(\beta_{0j})} = \exp(\beta_{1j}) \tag{26}$$

In the above equations, *p.Aj* equals the average baseline count for case *j*, *p.Bj* equals the average treatment count for case *j*, and *p.AjBj* equals the average treatment effect for case *j*. E.g., for case *j* in a given SCD study, if the average baseline mean *p.Aj* equals 10 and *p.Bj* equals 5 it means for case *j*, average count of outcome variable in the baseline phase and in the treatment phase is 10 and 5 respectively. Then, *p.AjBj* equals 0.5 and is interpreted as the average reduction in the outcome variable following treatment phase for case *j* is 50%. Nonlinear Poisson ES provides both between-subjects (the average across all the cases) and within-subject (for each case) effect sizes and for each parameter provides several summary outputs (e.g., mean, median, sd, 95% HDI). The WINBUGS code for the nonlinear Poisson model is in Appendix C.

2.6.6 Bayesian Rate Ratio Effect Size

Natesan Batley, Shukla Mehta, & Hitchcock (2021) proposed a within-subject BRR effect size for count data in SCDs. The mathematical details of this model are discussed in the following paragraphs.

Let $y_{p1}, y_{p2}, \ldots\ldots, y_{pt}$ denote the observed count outcomes where phase *p* equals 1 if the observations are in the baseline phase and equals 2 if in the treatment phase; *t* = 1,2,… t_b, t_{b+1},…,T The observed value at the first time point (y_{p1}) in phase

p follows Poisson distribution with mean $\hat{y}_{p1}$, where $\hat{y}_{p1}$ denotes the probability of obtaining a given response on the given model.

The observed values in the subsequent time series are distributed as shown in equation 27.

$$y_{pt}|H_{pt-1}, \Theta \sim Po\left(\hat{y}_{pt|(pt-1)}\right) \quad (27)$$

In equation 27, H_{pt-1} denotes the past history, Θ is the vector of parameters to be estimated, and Po refers to Poisson distribution. The observed values at time t in Phase p is Poisson-distributed with mean $\hat{y}_{pt|(pt-1)}$, where $\hat{y}_{pt|(pt-1)}$ is the probability of the predicted value of the current data point given the previous data point. The generalized linear model and the autocorrelation of the residual term is expressed as follows:

$$\hat{y}_{pt} = \begin{cases} \exp\left(\beta_{01} + e_{pt-1}\right), if\ t \leq t_b \\ \exp\left(\beta_{02} + e_{pt-1}\right), otherwise \end{cases} \quad (28)$$

and,

$$e_{pt-1} = \rho e_{pt-2} + \varepsilon \quad (29)$$

In equation 28, $\hat{y}_{pt}$ is the probability of the predicted value of the outcome variable at time t in phase p; β_{01} and β_{02} are the intercepts of baseline and treatment phases respectively; e_{pt} is the error at Time t in phase p; ρ is the autocorrelation coefficient, and ε is the independently distributed error. e is the white noise created by a combination of random error, ε and autocorrelation between adjacent time-points, ρ. Their standard error is calculated as shown

$$\sigma_e = \frac{\sigma_\varepsilon}{\sqrt{1-\rho^2}} \quad (30)$$

The intercepts for the baseline and treatment phase is modeled as shown below where there are $1, 2, \ldots, t_b$ and $t_{b+1}, \ldots, t_n$ time points in the baseline and treatment phases respectively.

$$\beta_{0p} = \begin{cases} \beta_{01}, if\ t \leq t_b \\ \beta_{02}, otherwise \end{cases} \quad (31)$$

To obtain the posterior, an example of the uninformative prior specification for the parameters is discussed. Let the intercepts be drawn from normal distributions with hyperpriors. The log phase means μ_{0p} are normally distributed with means 0 and variance 100. The standard deviations for each phase follow gamma distributions with both shape and rate parameter equal to 1. Autocorrelation is drawn from uniform distribution.

$$\beta_{0p} \sim norm(\mu_{0p}, \sigma_p^2) \quad (32)$$

$$\mu_{0p} \sim norm(0, 100); p = 1, 2 \quad (33)$$

$$\sigma_p \sim gamma(1, 1) \quad (34)$$

$$\rho \sim uniform(-1, 1) \quad (35)$$

The BRR effect size estimate is obtained from the posterior distribution of the rate ratio of the mean of the distribution from which the baseline and treatment intercepts are drawn as shown in Equation 36.

$$\mu_{ratio} = \frac{e^{\mu_2}}{e^{\mu_1}} \quad (36)$$

The rate ratio is interpreted as an increase or reduction in the outcome variable in the treatment compared to the baseline.

2.7 Common notation across LRR, NLB, and BRR effect sizes

All the three ES methods as shown in the equations above calculate effect size as the rate ratio of average treatment mean to the average baseline mean in either log count metric or in the raw count metric. Thus, the effect size from the three methods LRR, NLB, and NLB can be written in similar notation as follows:

$$\text{LRR ES} = \ln\left(\frac{\text{TrtMeanCount}}{\text{BaselineMeanCount}}\right) \quad (37)$$

$$\text{NLB ES} = \exp\left(\frac{\beta_{1j}}{\beta_{0j}}\right) = \frac{\text{TrtMeanCount}}{\text{BaselineMeanCount}} \quad (38)$$

$$\text{BRR ES} = \frac{e^{\mu_2}}{e^{\mu_1}} = \frac{\text{TrtMeanCount}}{\text{BaselineMeanCount}} \quad (39)$$

In the above equations, the numerator are the treatment mean counts and denominator are baseline mean counts. This shows that all the three methods estimate the ES as a rate ratio of either log counts or a ratio of exponentiated counts in the treatment and baseline phases and are potentially comparable after a mathematical transformation.

2.8 Autocorrelation in SCDs

Shadish, Rindskopf, Hedges, & Sullivan (2013) state that in the last 20 years, one reason for increased endeavors in developing statistical analysis for SCD studies, was to account for autocorrelation present in the data. Failing to account for autocorrelation can result in biased estimates and inflated type I error. In short time series, the estimates of autocorrelation tend to be negatively biased (Shadish, Rindskopf, Hedges, & Sullivan, 2013). Large autocorrelations can reduce within-phase variation in the data, resulting in inflated estimates of intervention effects (Sigurdsson & Austin, 2006). Presence of significant positive autocorrelation can lead to underestimation of type I error while significant negative autocorrelation can overestimate type I errors (Jenson, Clark, Kircher, and Kristjansson, 2007). Although many recommend taking autocorrelation into account while analyzing SCD data, others suggest the concern in unwarranted (Huitema & McKean, 1998; Matyas & Greenwood, 1996). This is mainly because the evidence of autocorrelation in short time series data like SCD is typically not conclusive. This is because accurate

estimation of autocorrelation in SCDs is challenging given the small number of data points typically available in SCDs (Shadish, Kyse, & Rindskopf, 2013).

One of the questions this dissertation aims to understand is how do different levels of known autocorrelation affect the estimates (and associated standard errors) from the three SCD ES methods for count data. Among the NLB, BRR, and LRR ES methods, only BRR takes autocorrelation into account. If simulated autocorrelated count data are used with the three ES methods, the results can provide insight on the quality of the estimates obtained with a known value of autocorrelation.

Harrington & Velicer (2015) examined the extent to which interrupted time series analysis is applied in SCDs. In their review of 25 articles, 46.01% had low lag-1 autocorrelation (0.00 to 0.50), 41.10% had moderate autocorrelations (0.51 to 0.75), 4.91% had high autocorrelation (greater than 0.75), and 7.98% had negative autocorrelation (-0.32 to -0.05) for the time series data. An extensive review by Shadish & Sullivan, (2011) of the characteristics of 809 SCD across 113 studies from 21 journals published in 2008, found most of the single-case designs had autocorrelation significantly higher than zero, while alternating treatment (AT) and changing criterion (CC) designs had an average autocorrelation significantly less than zero. Out of the 809 SCD only 10.6% were either AT or CC. Moreover, studies that recognize autocorrelation tend to focus on positive autocorrelation values (Swan, Pustejovsky, & Beretvas, 2020; Ferron, Farmer, & Owens, 2010). Hence, in the proposed dissertation only positive autocorrelation values will be used to simulate autocorrelated count data.

2.9 Overdispersion in count data

Generally, overdispersion occurs when we model the count outcome data with either a binomial or Poisson distribution and the variance of the outcome is larger than what we anticipate for the choice of response distribution (Montgomery, Peck, & Vining, 2006). Since this dissertation is focused only on counts during a fixed time period, the concern is about the overdispersion that might occur when outcome data is modeled with Poisson distribution. In the Poisson distribution, mean and variance are related and are equal to the parameter λ (lambda).

If the response variable, Y_i, $i=0,1,2,..$ are counts observed over a fixed period of time then a reasonable model for this data is Poisson distribution with the probability mass function as,

$$f(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!} \; ; \qquad y = 0,1, ... \tag{40}$$

and both mean and variance are equal to parameter λ.

However, in real scenarios, it is seldom the case where both the mean and variance are equal (aka equidispersion). Routinely, the variance is larger than the mean. Because the Poisson distribution has only one parameter to account for both mean and variance, this may induce bias. A Poisson model used for overdispersed data can result in underestimated standard errors of the parameter estimates (Payne, Gebregziabher, Hardin, Ramakrishnan, & Egede, 2018). Thus, if overdispersion is not accounted for, then standard error estimates might be too small, test statistics for the parameter estimates will be too big, significance will be overestimated (i.e., Type I errors are inflated), and confidence limits will be very small (Coxe, West, & Aiken, 2009).

Both NLB and BRR use a Poisson model (multilevel Poisson model and Poisson regression at single level respectively) to estimate the treatment effects. Thus, it is important to also examine how the presence of overdispersion affects the estimates, especially the standard errors of the effect size from these methods.

Using Monte Carlo simulation, the effect of overdispersion can be examined. The effect of both autocorrelation and overdispersion, which can be common characteristics of SCD count data is yet to be examined for the three ES methods, Thus, one of the key research question this study aims to answer is how autocorrelation and overdispersion influence the effect size and standard error estimates.

Chapter 3

METHODS

This dissertation aims to answer the research questions using Monte Carlo simulation, with data generation parameters guided by data from a published SCD study. Using data from a published study with an ABAB reversal design, the understandability and interpretability of the effect size estimates from the LRR, NLB, and BRR models will be examined. Analyzing the same simulated datasets, the estimates from the three methods will be compared, and the benefits and challenges associated with implementing these three models will be discussed. By using Monte Carlo simulation, the influence of autocorrelation and overdispersion on the effect size estimates and standard errors from the LRR, NLB, and BRR models can be explored. This chapter describes the data, and proposed methodology for this dissertation.

The rationale for choosing a specific SCD study for the dissertation along with its details, the Monte Carlo simulation conditions, and analysis plan are described next.

3.1 Dataset

Schmidt (2007) provides the initial dataset for this dissertation. This study used an ABAB design, it had 3 cases with at least 3 data points per phase, and the outcome was measured in frequency counts for a fixed time period (the number of disruptive behaviors demonstrated in each phase).

3.1.1 Rationale for choosing Schmidt (2007) study

The initial plan for this dissertation was to use one or more datasets from studies that satisfy the WWC criteria for single-case studies. However, a review of the 50 studies that were listed in the WWC website (WWC, 2022b) that meet WWC pilot design standards with or with reservations, only 8 studies used an ABAB design. Furthermore, none of these 8 studies were deemed appropriate for this dissertation because either they had different outcomes for different students, outcomes were not measured as counts, or they had different mandatory training for some students during data collection from the four phases and no training for the other student.

Many methods papers in the SCD literature have used the dataset from Lambert et al (2006) to demonstrate how their proposed method works with a real SCD study (Rindskopf, 2014; Natesan Batley, Shukla Mehta, & Hitchcock, 2021; Swaminathan, Rogers, & Horner, 2014; Hedges, Pustejovsky, & Shadish, 2012). The Lambert study examined the effect of using response cards on nine fourth grade students from two classrooms in reducing their disruptive behaviors during math lessons. Though Lambert et al (2006) was also reviewed as a potential dataset for this dissertation, the outcome variable (disruptive behavior) in the Lambert et al (2006) study was measured as an observed count based on a fixed number of trials per day (10 trials) while this dissertation focuses on count for a fixed time. Moreover, this study included nine students, which is far more than a typical SCD study.

The Schmidt (2007) dataset represents a more typical SCD study with 3 cases. Shadish & Sullivan (2011) in their review found the median and mode number of

cases in a SCD study was 3. Because Schmidt (2007) represents a typical study with three cases and at least 3 data points per phase, it serves as a good starting point to answer my research question on how the different ES statistics perform with a typical SCD study with ABAB design.

3.1.2 Schmidt (2007) study details

Schmidt (2007) examined the effect of the Class-wide Function-related Interventions Teams (CW-FIT) on the on-task behavior of eleven first grade students in an urban elementary charter school with a large number of English Language Learner (ELL) students, on-task and disruptive behaviors for three target students within the classroom, and effects of the intervention on the class-teacher's behavior. The CW-FIT, is a preventative approach that is implemented in classrooms where behavior problems are more likely to occur, or to address these behaviors if they already exists in classrooms. For the purpose of this dissertation, only the disruptive behaviors outcome data for three target students will be used because this outcome is measured as frequency counts. The pseudonyms of the three target students were Lilly (female), Albert (male), and Faith (female), all six-year-old ELL students who had recently moved to US from Africa and were identified as having ongoing problem behavior. These target students: Lilly, Albert, and Faith are also referred to as Case1, Case 2, and Case 3 respectively.

The effects of the CW-FIT intervention on disruptive behavior (verbal disruptive, physical disruptive, and general disruptive behaviors all combined for total disruptions) were examined using an ABAB design. The baseline consisted of a school wide system of positive behavior supports (SWPBS), which was established as school-

wide behavioral expectations and were reinforced through the school-wide reward system (pre-determined hole-punches to students' reward cards to redeem award, e.g., candy). The CW-FIT intervention consisted of three primary components: (a) teaching three appropriate classroom behaviors (1. "How to get the teacher's attention", 2." Follow directions the first time", and 3. "Ignore In-appropriate behavior"), (b) using a group contingency, and (c) differentially reinforcing those behaviors through a class-wide program using teams and awarding points. The second and third component consisted of dividing the class into groups of 3 or 4 students and each member of the group had to display the skills taught in the first component to earn points and if they earned the preset points then they would receive an award (e.g., stickers). For target students, they served as their own group.

The intervention was given to whole classroom (during the final hour of the day, with maximum 40 minutes to implement the CW-FIT each day). However, the disruptive behaviors were measured only for the three target students. The disruptive behavior was measured as frequency counts and students had at least three data points per phase (See Figure 5). Using the Multiple Option Observation System for Experimental Studies (MOOSES), the disruptive behavior was recorded. Each MOOSES observation occurred for 10 minutes and allowed collection of frequency of disruptive behavior. In the 10 minutes observation session, the upper limit can be 600 demonstrations of disruptive behaviors if the student is engaged in disruptive behavior every second in the 10 minutes. Baseline data were collected for 7 days. Behavioral skill instruction was provided approximately for 2 weeks during which there was no data collection. In the intervention phase CW-FIT game was introduced and was in place for 18 school days. There was 3-day reversal period and then the intervention

was reintroduced for 6 days. The target students did not have the same number of observations due to absences.

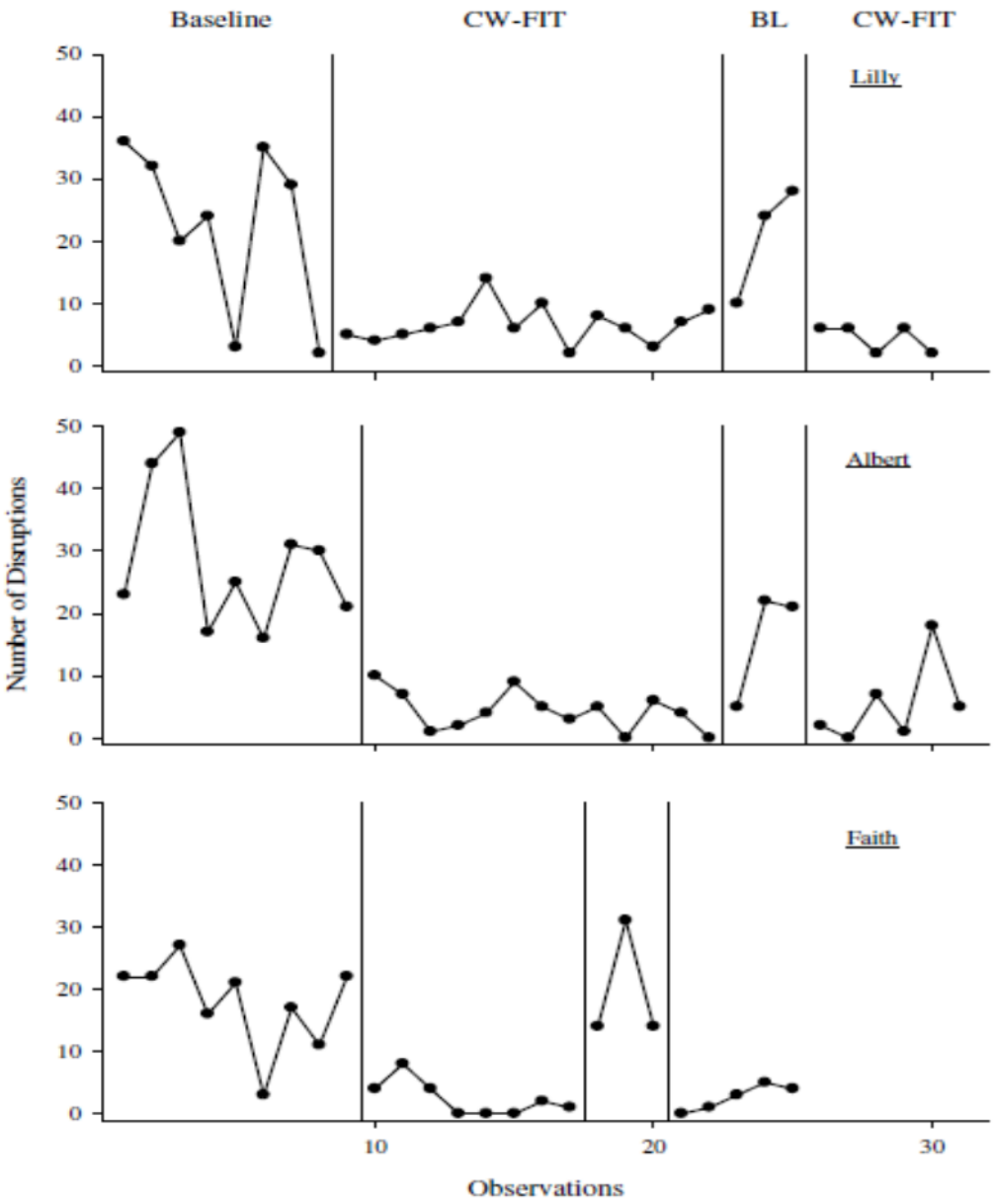

Figure 5: Number of disruptive behaviors for Schmidt (2007) (Lilly, Albert, and Faith per 10-min observation).

The study found CW-FIT intervention resulted in reduction of disruptive behavior. Specifically, for Lilly, during baselines the mean of 21-23 disruptions was reported which changed to mean of 7.4 disruptions in the intervention phases. Albert's data showed an average of 28 disruptions during the first baseline which reduced to a mean of 4 in the intervention phase. For Faith, in the baseline conditions, disruptive behavior averaged 16 to 18 which dropped to averages of 2 to 3 during treatment phases. Lilly showed the highest overall decrease in the disruptive behaviors during the intervention. The study also reported the inter observer agreement which was 86% (range 63% to 100%) for disruptive behavior.

3.2 Data Extraction

The Schmidt (2007) data was taken from the "SingleCaseES", R package that was created by Pustejovsky, Chen, & Swan (2021). This R package provides numerous R functions for calculating effect size indices for single-case designs and includes datasets from a number of real SCD studies including the Schmidt (2007) study. The Schmidt (2007) dataset in the "SingleCaseES" package consists of both the on-task and disruptive behavior of the three target students. For this dissertation only the disruptive behavior data was used.

3.3 Data Analysis

The details of the analysis of the disruptive behavior data from the Schmidt (2007) study are now discussed.

3.3.1 Estimate LRR effect size

For the Schmidt (2007) disruptive behavior data of the three target students, bias-corrected LRR ES was estimated using the "SingleCaseES" R package (Pustejovsky, Chen, & Swan, 2021). To get the estimates, observations from each adjacent baseline and treatment phases for each target student were supplied to the R program. The program provides the LRR ES estimates for adjacent pair of AB phases, its standard error, 95% confidence interval, and the percentage change in the level for each of the 2 adjacent phases being compared. Since the outcome of interest is decrease in the number of disruptive behaviors, thus LRR-d option in the "SingleCaseES" package was used. The estimates obtained for each target student or case were interpreted in two related formats. The first one, is as indicated in the Pustejovsky (2018), and a second interpretation used transformation to convert the LRR effect size estimates to a common metric with the NLB and BRR effect size estimates. This dissertation did not combine the two LRR effect size estimates for first baseline and first treatment phases (say A1B1) and for second baseline and second treatment phases (say A2B2) phases (Note: Instead of ABAB, for ease of explanation, A1B1A2B2 is used) into a single estimate (as shown in Pustejovsky (2018)) to be able to compare with the NLB and BRR estimates.

3.3.2 Estimate BRR effect size

The BRR effect size for the Schmidt (2007) dataset was estimated using the BRR program available on the GitHub site (Prathiba-stat, 2020). The BRR program files were downloaded and saved on the local files. During analysis, it was noticed that there is a difference in the BRR model as stated in the Natesan Batley, Shukla Mehta, & Hitchcock (2021) paper and the R program code available on the GitHub site.

Specifically, the BRR model in the paper used the parameter-driven model while the GitHub program used an observation driven-model. In simple words, in an observation-driven model, the autocorrelation is induced due to the dependence of current observation on the past observation, while in parameter-driven model, there is an unknown underlying mechanism responsible for autocorrelated time-series data (Davis, Dunsmuir, & Wang, 1999). This dissertation uses the model based (observation-driven model) on the GitHub program. The BRR model used in the paper (Natesan Batley, Shukla Mehta, & Hitchcock, 2021) is presented in the literature review chapter of this dissertation, and the details of the BRR model used in this dissertation are as follows.

3.3.2.1 Bayesian Rate Ratio Effect size model (observation-driven model)

Let $y_{p1,}\ y_{p2}, \dots \dots, y_{pt}$ denote the observed count outcomes where phase p equals 1 if the observations are in the baseline phase and equals 2 if in the treatment phase; t = 1,2,… t_b, t_{b+1},…,T, where t_b is the last observation in the baseline phase. The observed value at the first time point (y_{p1}) in phase p follows Poisson distribution with parameter $\hat{\lambda}_{p1}$ where $\hat{\lambda}_{p1}$ denotes the expected number of counts of the outcome variable.

The observed values in the subsequent time-series in a given phase follows a Poisson distribution with 1-lag autocorrelation,

$$y_{pt} \sim Po(\hat{\lambda}_{pt}) \tag{41}$$

where,

$$\hat{\lambda}_{pt} = \begin{cases} \exp(\beta_{01}) + \rho\left(y_{p(t-1)} - \lambda_{p(t-1)}\right), & if\ t \leq t_b \\ \exp(\beta_{02}) + \rho\left(y_{p(t-1)} - \lambda_{p(t-1)}\right), & otherwise \end{cases} \tag{42}$$

In equation above, $\hat{\lambda}_{pt}$ is the expected value of the outcome variable of the Poisson distribution for phase *p* at time *t*; β_{01} and β_{02} are the intercepts of baseline and treatment phases respectively; ρ is the autocorrelation coefficient, $y_{p(t-1)}$ is the observed outcome at phase *p* and time *t-1*, $\lambda_{p(t-1)}$ is the expected count of the outcome variable at phase *p* and time *t-1*.

The intercepts for the baseline and treatment phase are modeled as shown and there are $1, 2, \ldots, t_b$ and $t_{b+1}, \ldots, t_n$ time points in the baseline and treatment phases respectively.

$$\beta_{0p} = \begin{cases} \beta_{01}, if\ t \leq t_b \\ \beta_{02}, otherwise \end{cases} \quad (43)$$

To obtain the posterior distribution of the phase means, autocorrelation, precision, and treatment effect, either weakly informative priors or vague priors were used. The intercepts were drawn from normal distributions with hyper priors (prior on a prior). The log phase means μ_{0p} were normally distributed with mean 0 and variance 1. The standard deviations for each phase followed gamma distributions with both shape and rate parameter equal to 1. Autocorrelation was drawn from uniform distribution.

$$\beta_{0p} \sim norm(\mu_{0p}, \sigma_p^2) \quad (44)$$

$$\mu_{0p} \sim norm(0, 1); p = 1, 2 \quad (45)$$

$$\sigma_p \sim gamma(1, 1) \quad (46)$$

$$\rho \sim uniform(-1, 1) \quad (47)$$

The BRR effect size estimate was obtained from the posterior distribution of the rate ratio of the mean of the distribution from which the baseline and treatment intercepts were drawn as shown in Equation 37.

$$\mu_{ratio} = \frac{e^{\mu_2}}{e^{\mu_1}} \quad (48)$$

The rate ratio is interpreted as an increase or reduction in the outcome variable in the treatment compared to the baseline. This model takes observations from two phases at a time and estimates the parameters.

To estimate the posterior parameters of interest, specifically the phase means and effect size, this dissertation used weakly informative prior. The priors provided in the BRR and NLB papers and GitHub site were used as reference while estimating the model parameters for Schmidt (2007) study. The NLB method also used the same prior for the common parameters to facilitate comparison of the estimates.

Just Another Gibbs Sampler (JAGS) was used to fit the BRR model and approximate the posterior distributions of the phase means, autocorrelation, treatment effects, and variances. Under the R environment, either rjags or runjags (Denwood, 2016) is used to call JAGS (Kruschke, 2015). Specifically, the autorunjags option of the runjags R package was used. This primarily ensures convergence of the four parallel chains used and provides plausible values from the posterior distribution of the parameters. Runjags runs parallel chains and the model estimates are iterated until convergence is reached. The first 2,000 iterations were burned-in to avoid the influence of the initial values on the estimates. The initial values of the chains were set using random number generations available in the runjags package. Convergence was checked using Gelman and Rubin statistic (also called potential scale reduction factor (PSRF)), (Brooks & Gelman, 1998). For each parameter, 12,000 samples were generated with thinning of every third sample. Thinning was used so that it would remove MCMC autocorrelation and provide samples that are representative of the true underlying posterior distribution, while also reducing time in post-chain processing.

Using the JAGS, the posterior distribution was approximated for the study parameters: phase means, precision, effect sizes for each of the adjacent phase (A1B1, B1A2, and A2B2), and autocorrelation. These ES estimates are interpreted as indicated in the BRR paper and if necessary, a second interpretation, that is easily understood and in a common metric is provided. The ES point estimates from the LRR, NLB, and BRR are transformed into a common metric, and interpreted.

3.3.3 Estimate NLB effect size

The NLB model for an AB design with Poisson count outcome data is presented in equations 23 to 26. This will be extended for an ABAB design and is as follows:

3.3.3.1 Nonlinear Bayesian Model for ABAB design:

Let the dependent variable Y_{ij} is count data (count over a period of time) for person j at time point i. Then, Y_{ij} follows a Poisson distribution with parameter η_{ij}. Using the generalized linear model with log link function, the level-1 model is now expressed in terms of the natural logarithm of the expected counts, η_{ij}.

$$ln(\eta_{ij}) = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} \qquad (49)$$

where η_{ij} is the expected value of Y_{ij}.

In the above equations, β_{0j} is the average log count in the baseline phase (intercept), β_{1j} is the average log count of the treatment effect from first baseline (A1) to first treatment (B1) phase, β_{2j} is the average log count of the treatment effect from the first treatment (B1) phase to second baseline phase (A2) i.e., removal of the treatment, and β_{3j} is the average log count of the treatment effect from second baseline phase (A2) to second treatment phase (B2) i.e., re-introduction of treatment.

X_{1ij}, X_{2ij}, and X_{3ij} are corresponding phase dummy variables. Phase B1 = $\beta_{0j} + \beta_{1j}$, is the average log count in the first treatment phase (B1), phase A2 = $\beta_{0j} + \beta_{1j} + \beta_{2j}$, is the average log count in the second baseline phase (A2) and phase B2 = $\beta_{0j} + \beta_{1j} + \beta_{2j} + \beta_{3j}$, is the average log count in the second treatment phase (B2) . These parameters are estimated using Bayesian estimation via MCMC. The level-2 model is similar to equations 11 and 12. γ_{00}, γ_{10}, γ_{20}, and γ_{30} are the fixed effects. γ_{00} is average baseline level across all the cases, γ_{10} is the average treatment effect of A1B1 phases across all the cases, γ_{20} is the average treatment effect of B1A2 phases across all the cases, and γ_{30} is the average treatment effect of A2B2 phases across all the cases. The random effects are u_{0j}, u_{1j}, u_{2j} and u_{3j}, which are assumed to be distributed as a multivariate normal and independent between cases. The unexplained variation among the cases are represented by the variance of u_{0j} for baseline, u_{1j} for treatment effects of A1B1 phases, u_{2j} for treatment effects of B1A2 phases, and u_{3j} for treatment effects of A2B2 phases. The corresponding variances are τ_{00}, τ_{11}, τ_{22}, and τ_{33} respectively. The covariances are τ_{01} or τ_{10}, τ_{02} or τ_{20}, τ_{03} or τ_{30}, τ_{12} or τ_{21}, τ_{13} or τ_{31}, and τ_{23} or τ_{32}.

Similar to the Nonlinear Binomial model, the priors for the parameters can be specified and JAGS can be used to approximate the posterior distributions. Likewise, for the ease of interpretation, parameters are transformed back into the original scale using exponentiation. The average counts in baseline and treatment phases, and effect size for case *j* are transformed as follows.

$$p.A1j = \exp(\beta_{0j}) \tag{50}$$

$$p.B1j = \exp(\beta_{0j}) \exp(\beta_{1j}) \tag{51}$$

$$p.A2j = \exp(\beta_{0j}) \exp(\beta_{1j}) \exp(\beta_{2j}) \tag{52}$$

$$p.B2j = \exp(\beta_{0j}) \exp(\beta_{1j}) \exp(\beta_{2j}) \exp(\beta_{3j}) \tag{53}$$

The Nonlinear Poisson (NLB) effect size for adjacent phases is calculated as,

$$p.A1jB1j = \exp(\beta_{1j}) \quad (54)$$

$$p.B1jA2j = \exp(\beta_{2j}) \quad (55)$$

$$p.A2jB2j = \exp(\beta_{3j}) \quad (56)$$

For using the NLB model with ABAB design, to analyze the disruptive behavior data for the three target students, effects coding discussed in Shadish, Kyse, & Rindskopf (2013) was used. Step coding (using numbers 0 and 1), resembles steps pattern. In this coding, intercept represents the outcome during the first baseline phase, and there are three effects. The first effect measures the change (treatment effect) from A1 to B1 which is coded using the first dummy variable, X_{1ij} that equals 0 during A1 phase and equals to 1 for phases B1, A2, and B2. The second effect measures the change from B1 to A2, coded using second dummy variable, $X_{2ij,}$ which equals 0 for phases A1 and B1, and equals to 1 for phases A2 and B2. The third effect measures change from A2 to B2, coded using third dummy variable X_{3ij}, which equals 0 for phases A1, B1, and A2, and equals to 1 for phase B2. Thus, the first, second, and third effect respectively measure the treatment effects from A1 to B1, B1 to A2, and A2 to B2 respectively. However, only effect sizes from A1B1 and A2B2 are used because this will allow the two AB phases ES to be compared with both LRR and BRR ES estimates.

To fit the NLB model to the Schmidt (2007) data, the same priors used in BRR for phase means and precision (equations, 44 to 46) were used. Similarly, JAGS was used to approximate the posterior distribution of the parameters. The same conditions of 2,000 samples burn-in, 4 parallel chains, 12,000 samples per parameter with thinning of the third sample used in BRR was implemented for NLB. For all

parameters convergence was examined using Gelman and Rubin statistic. The NLB model uses partial pooling approach and thus both the within subject and across subject NLB estimates provided by the model will be interpreted as indicated in the paper and if necessary, second interpretation is provided, that is easily understood and in a common metric.

3.3.4 Assessing Understandability and Interpretability

Methods for assessing the understandability and interpretability of the three effect sizes are now discussed.

To answer the second research question (1a), to what extent do the estimates from LRR, NLB, and BRR satisfy the understandability and interpretability criteria proposed by May (2004) for meaningful statistics, first the effect size estimates from each of the method for the Schmidt (2007) study were interpreted as indicated in their respective papers. If the estimates are presented in familiar metrics/units and if one can easily understand the information presented, then it satisfies the understandability and interpretability criteria. For example, in the LRR method, the ES estimates are provided in the log count metric and in percentage change in the mean from baseline to treatment phase. Though the LRR ES in log count is not easy to interpret, presenting the ES estimates equivalently in terms of percentage change is easily understandable and interpretable. This is because percentage change is commonly used and is intuitive to understand, though a high value might suggest a rather exorbitant treatment effect.

To transform the results given in log count and interpret in original counts, a simple solution is to transform (i.e., via exponentiation) the log count estimate. This example presents an instance of how an estimate can be converted to a common metric

so that the result can be easily understood. A similar approach, if necessary, was used for the estimates obtained from NLB and BRR.

A key goal of this dissertation is to help potential SCD users (researchers and applied practitioners) in understanding and interpreting the ES and other estimates from the three methods so that they can make informed decision on which ES method to be used for their respective studies. Thus, efforts were made to describe the estimates in terms that are as understandable and interpretable as possible. This may help researchers who do not have strong statistical expertise to first clearly understand and interpret the results themselves, and they may in turn be able to present findings to wider audiences in a simplified manner as suggested by May (2004).

3.3.5 Common Metric

To be able to compare the ES estimates (their numerical values and corresponding standard error) from the three methods, the estimates should be in a common metric. LRR provides the ES estimates in log count, BRR produces the ES estimates as a ratio of expected counts, and NLB provides the ES estimate in both log count and as a ratio of expected counts. Since all the three estimates involve ratios of counts (equations 37 - 39), thus it might be feasible to transform the ES estimates to a common metric. The transformation is presented in the results chapter.

3.3.6 Simulation Design

Using Monte Carlo simulations, the behavior of the three models under various conditions of phase lengths, phase means, autocorrelation, and overdispersion for a SCD with ABAB design and 3 cases are examined. The simulation conditions are informed by the SCD literature and details are presented in the subsections.

3.3.6.1 Sample size and phase means

Simulated datasets were generated to replicate a typical SCD study with an ABAB design. For this purpose, information from the same Schmidt (2007) study was used to set the number of cases, the number of data points in each phase, and the phase means for each phase and case. Thus, for the simulation of SCD data, 3 cases with number of data points in each phase and phase means in each condition equal to Schmidt (2007) study was used. The phase means and sample size for the 3 cases are presented in Table 1.

Table 1: Summary Statistics of Schmidt (2007) study.

Students	Phase 1		Phase 2		Phase 3		Phase 4	
	N	Mean	N	Mean	N	Mean	N	Mean
Lilly	8	23	14	8	3	21	5	5
Albert	9	29	13	5	3	17	6	6
Faith	9	19	8	4	3	20	5	4

3.3.6.2 Autocorrelation

The parameter values of the autocorrelation were selected based on the SCD literature wherein prior studies used autocorrelation values between 0.0 and 0.4. Ferron, Farmer, & Owens (2010), used simulation to examine the accuracy of confidence-intervals of individual treatment effects obtained from multilevel modeling of multiple-baseline data. They used the following autocorrelation values (0.0, 0.1, 0.2, 0.3, 0.4). Similarly, Swan, Pustejovsky, & Beretvas (2020) used autocorrelation (0.0, 0.2, 0.4) in their study that examined the effect of various response design algorithms on the baseline data pattern. Thus, I will be using the three levels of autocorrelation (0.0, 0.2, 0.4) in my simulation.

3.3.6.3 Overdispersion

The Poisson distribution is not suitable for generating overdispersed count data. An alternative is to use the negative binomial distribution. This distribution can be used with count outcomes that have overdispersion as it relaxes the mean-variance relationship. Using the "ecological" parametrization of negative binomial distribution, μ is the mean number of counts in a sample and *k* is the overdispersion parameter that measures the amount of clustering or heterogeneity in the data (Bolker, 2008). The variance of the negative binomial distribution is $\mu + \frac{\mu^2}{k}$. As *k* increases the variance approaches the mean and the negative binomial distribution approaches the Poisson distribution. In this dissertation, the negative binomial distribution was used to generate overdispersed and autocorrelated count data using the "nbinom" function (with two parameters mean = mu and size = 1/*k*) in R.

Autocorrelation was modeled via the mean parameter of the negative binomial distribution. To obtain autocorrelated and overdispersed data, the equation for mean mu, and size (reciprocal of overdispersion) that was used is as follows. Let Y_1, Y_2,…, Y_n be the n count data to be generated for a phase then, the first data point using rnbinom function of R is generated as,

Y_1 = rnbinom (n=1, mu = average_rate, size = 1/*k*)

where average_rate is the sample mean for that phase.

For generating the subsequent data points, the following was used:

Y_i = rnbinom (n =1, mu = ((1-ρ) × average_rate + ρ × $Y_{[i-1]}$, size = 1/*k*)

where i = 2,3,..,n; ρ = autocorrelation; *k* = overdispersion parameter.

To confirm the overdispersion values for the data generation process, an exercise was undertaken to see how different overdispersion values affected the range of data generated. As mentioned in the study details of the Schmidt (2007) study, the

upper limit of the number of disruptive behaviors for any case can be 600. Overdispersed data was generated for two phases with mean values of 5 and 20 (similar to phase means in Schmidt study), sample size of 6 data points in each phase, 3 autocorrelation values (0.0, 0.2, 0.4) and different overdispersion values (ranging from 0.0001 to 0.9). As the overdispersion and autocorrelation grew large, there was considerable instability in the simulated data and some data points exceeded 600. Thus, it was decided to use four overdispersion values (0.0001, 0.05, 0.1, 0.3) so that there was minimal chance of generating any data points near or above 600.

3.3.6.4 Simulation Conditions

A fully crossed factorial design with 3 autocorrelation (0.0, 0.2, 0.4) and 4 overdispersion (0.0001, 0.05, 0.1, 0.3) levels resulted in 12 simulation conditions for each of the 3 cases. The sample size and mean of A1B1A2B2 phases for the 3 cases are presented in Table 1. Using the negative binomial distribution, 5,000 simulated SCD datasets for each of the 3 cases were generated. This resulted in a total of 180,000 ($5000 \times 3 \times 12$) Monte Carlo datasets.

Initially, the analytic plan included use of all 5,000 replications to obtain effect sizes and other estimates for each of the 3 methods. While the LRR method took minimal computing time to generate the estimates for all 5,000 datasets, both the NLB and BRR took substantial computing time to run just one replication. For example, without thinning, and using four parallel chains with 20,000 MCMC iterations, it took a Windows PC with a dual-core 2.5GHz CPU around 80 seconds to estimate the NLB model and around 40 seconds to estimate the BRR for one replication. Extrapolating these times suggests that to run these models for 180,000 datasets would take 4,000 hours for NLB and 2,000 hours for BRR.

To reduce the computational time, only 1,000 Monte Carlo replications were used for each model and condition, with MCMC thinning to select only every 3rd posterior parameter value sampled from the MCMC chain. This alone would reduce computing time by 80%, but the total time required would still be in excess of 1,000 hours.

To obtain results within just a few days, analyses were carried out using the cluster computing and parallel processing capabilities of the Delaware Advanced Research Workforce and Innovation Network (DARWIN) high performance computing system (HPC) at the University of Delaware.[1] The DARWIN HPC has 105 compute nodes with a total of 6,672 cores, 22 GPUs, 100TB of memory, and 1.2PB of storage.

To run the NLB and BRR R programs for the simulated data in DARWIN, cluster computing was used. Cluster computing is similar to distributed computing where multiple computers work in tandem to complete different computing tasks. Before the final programs were run in the DARWIN HPC, they were tested for few iterations on both Windows and Mac operating systems. Only after they worked successfully in Windows and Mac, were final programs run on the DARWIN HPC.

---

[1] This research was supported in part through the use of DARWIN computing system: DARWIN – A Resource for Computational and Data-intensive Research at the University of Delaware and in the Delaware Region, which is supported by NSF under Grant Number: 1919839, Rudolf Eigenmann, Benjamin E. Bagozzi, Arthi Jayaraman, William Totten, and Cathy H. Wu, University of Delaware, 2021, URL:https://udspace.udel.edu/handle/19716/29071

For BRR, it took approximately 5 hours under each of the 12 conditions to estimate each model and successfully generate the estimates for each case using 1,000 Monte Carlo replications. For each of the 12 conditions of NLB except 1 condition, it took approximately 7 hours to estimate each model and successfully generate the estimates for 1,000 Monte Carlo replications. For the one condition with autocorrelation 0.2 and dispersion 0.3, the program terminated on both the Mac and DARWIN prematurely, but without obvious error. Fortunately, the NLB program for this condition did run successfully under the Windows operating system. This may be due to some issues in running JAGS in Windows versus Unix/Linux operating systems used by the Mac and the DARWIN HPC.

3.3.7 Performance Measures

The main aim of conducting a simulation study is to examine how the effect size and standard error estimates from LRR, BRR, and NLB methods perform under various autocorrelation and overdispersion values for a typical ABAB SCD study with 3 cases. The performance of each method was assessed in terms of unbiasedness and efficiency. For this, average bias, root mean square error (RMSE), coverage rates of 95% confidence interval/credible intervals, and range of the upper and lower 95% confidence/credible intervals were compared.

3.3.7.1 Bias and RMSE

The bias of an estimator is the difference between the estimated value of the estimator and true value of the parameter being estimated. If an estimator is unbiased, then on average the estimated value will be equal to the true value. The bias of the effect size estimates for each of the methods was calculated by averaging across all the

three cases and 1000 iterations from each of the 3 cases. The formula used for average bias is as follows:

$$Bias = \frac{\sum_{h=1}^{c}\sum_{i=1}^{n_{sim}}(\theta_{hi}^{est} - \theta_{hi})}{n_{sim} * c} \quad (57)$$

where, n_{sim} is the number of replications; c is the number of cases within a study; θ_{hi} is the true parameter value as a ratio of raw counts of the parameter for the hth case from the ith simulation study; and θ_{hi}^{est} is the estimated value of the parameter for the hth case from the ith simulation study. The bias is calculated for effect size in both count scale and log scale. The formula for the bias in log scale is as follows:

$$Bias(\log scale) = \frac{\sum_{h=1}^{c}\sum_{i=1}^{n_{sim}}(\ln(\theta_{hi}^{est}) - \ln(\theta_{hi}))}{n_{sim} * c} \quad (58)$$

In the above formula, $ln(\theta_{hi})$ is the natural log of the true parameter value as a ratio of raw counts of the parameter for the hth case from the ith simulation study; and $ln(\theta_{hi}^{est})$ is the natural log of the estimated value of the parameter for the hth case from the ith simulation study.

The mean square error (MSE) is a function of both variance and bias of an estimator (Morris, White, & Crowther, 2019), and is generally used as it provides a balance between bias and efficiency (Carsey & Harden, 2014). The RMSE is calculated as a square root of the MSE as shown. While comparing the methods, the estimator with the lowest RMSE is considered as performing better. The RMSE is calculated for the effect size in both counts and log scale.

$$RMSE = sqrt\left(\frac{\sum_{h=1}^{c}\sum_{i=1}^{n_{sim}}(\theta_{hi}^{est} - \theta_{hi})}{n_{sim} * c}\right) \quad (59)$$

$$RMSE(log\ scale) = sqrt\left(\frac{\sum_{h=1}^{c}\sum_{i=1}^{n_{sim}}(\ln(\theta_{hi}^{est}) - \ln(\theta_{hi}))}{n_{sim} * c}\right) \quad (60)$$

3.3.7.2 Coverage rates

Coverage rate is calculated as the proportion of 95% credible intervals (HDIs) that contain the true parameter value (Natesan & Hedges, 2017). Coverage rates are used for examining the efficiency of the standard error of the parameter estimate. Coverage rates were calculated for 95% confidence interval of the effect size estimates from the three methods and for 95% credible interval of the autocorrelation estimate from the BRR method.

3.3.7.3 Range of 97.5th and 2.5th percentiles

Generally empirical standard error and model standard error are calculated to evaluate the efficiency of the standard error of the estimates (Morris, White, & Crowther, 2019). In this dissertation, LRR provides ES estimates in log ratio and both BRR and NLB provides ES estimates in count ratio. Thus, the standard error of the ES estimates cannot be directly compared because they are on different scales. An alternative is to transform the lower and upper limits of 95% confidence intervals (CI) to a common scale (i.e., a raw count ratio) and examine how the average CI range differs across the 12 conditions for each method and across the methods. The results can indicate which methods have more or less precision in the estimates produced.

3.3.8 Examining Benefits and Challenges

The third research question of this dissertation aims to examine the benefits and challenges associated with implementing these three ES indices in a real SCD dataset. To answer this research question, the benefits and challenges that were faced in estimating the effect sizes for Schmidt (2007) study will be listed and discussed. Some of the key considerations were related to the ease of using the methods (e.g., in

terms of availability of programs to be used, the amount of information provided for the interpretation of the estimates by key references). Likewise, some challenges include the required knowledge of complex statistical modelling technique like MLM and Bayesian estimation, which can be challenging to SCD practitioners who might not be very familiar with advanced statistics.

Chapter 4

RESULTS

4.1 Schmidt (2007) estimates

Using LRR, BRR, and NLB methods, Schmidt (2007) disruptive behavior data for three cases were analyzed and parameters estimated. The summary statistics of the disruptive behavior of the three cases is presented in Table 2. Next, the results from each model are discussed. Among the three models, first the results from running the LRR count models on the Schmidt (2007) dataset are presented.

Table 2: Summary Statistics of disruptive behavior of three cases from Schmidt (2007)

Cases	Phase A1			Phase B1			Phase A2			Phase B2		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
Case1 (Lilly)	8	23.4	13.1	14	7.5	2.9	3	21.3	9.1	5	5.2	2.5
Case2 (Albert)	9	29.3	11.5	13	5.2	3.2	3	16.7	9.3	6	6.2	6.4
Case3 (Faith)	9	18.7	7.5	8	3.8	3.0	3	20.0	9.5	5	4.0	2.4

4.1.1 LRR estimates

The parameter estimates for the three cases using the LRR model are presented in Table 3. The “SingleCaseES” package provides the bias-corrected LRR-d ES estimate along with the standard error, 95% CI, and the LRR-d ES in the percentage change metric. The true autocorrelation for Schmidt (2007) dataset is not available. Thus, as noted in the Pustejovsky (2018) paper the SE for the LRR estimates for a single SCD study might not be valid in presence of positive autocorrelation, and thus will not be interpreted.

Table 3: LRR-d parameter estimates for disruptive behavior data of three cases from Schmidt (2007)

Cases	Phases	Lower 95	Upper 95	LRR ES	SE	% Change	% Change Lower 95	% Change Upper 95
Case 1 (Lilly)	A1B1	-1.5888	-0.7136	-1.1512	0.2233	-68.3739	-79.5824	-51.0123
	A2B2	-2.0574	-0.7802	-1.4188	0.3258	-75.8004	-87.2219	-54.1701
Case 2 (Albert)	A1B1	-2.1536	-1.3126	-1.7331	0.2145	-82.3266	-88.3935	-73.0885
	A2B2	-2.0016	0.0907	-0.9554	0.5338	-61.5343	-86.4875	9.4994
Case 3 (Faith)	A1B1	-2.1820	-0.9633	-1.5727	0.3109	-79.2511	-88.7188	-61.8379
	A2B2	-2.3711	-0.8487	-1.6099	0.3884	-80.0083	-90.6617	-57.2011

Again, for ease of explanation, the ABAB design is written as A1B1A2B2 for the first baseline, first treatment, second baseline, and second treatment phases respectively. For the first case, the bias-corrected LRR-d estimate (SE) of the first introduction of the treatment (comparing phases A1B1) is -1.1512 (0.2232). Thus, the average treatment effect in log count is -1.1512, which is not easily interpretable as it

is in log units. However, the LRR-d ES is also produced in percentage change metric which is much more intuitive and understandable and interpretable. The result indicates there is 68% reduction in disruptive behavior following the first introduction of the treatment for the first case. Similarly, the disruptive behaviors reduced by almost 76% when the treatment was re-introduced in the B2 phase.

For the second case, there was reduction of 82% and 62% in the disruptive behaviors following the first introduction of treatment and second (re-introduction) introduction of the treatment respectively. For the third case, the percentage change in the disruptive behavior was 79% and 80% for phases A1B1 and A2B2 respectively. The percentage decrease in disruptive behaviors among the three cases ranged from 68% to 82% and 62% to 80% for A1B1 and A2B2 respectively. Pustejovsky (2018) recommends combining the two LRR estimates for A1B1 and A2B2 into a single summary effect size. However, for the purpose of comparing LRR estimates with the BRR and NLB effect size estimates, the two LRR estimates are not combined.

4.1.2 BRR estimates

The BRR results for the three cases are presented in Tables 4 to 6. For each case, the BRR model used JAGS to draw 16,000 samples from the posterior distribution of the parameters (i.e., via 4 parallel chains, 12,000 iterations in each chain, and thinning of 3). The PSRF was below 1.05 for all the parameters estimated suggesting convergence. In the BRR output, for each of the parameters estimated, the mean, standard error, median (50th percentile), and 95% credible intervals are provided.

Table 4: BRR parameter estimates for disruptive behavior of first case (Lilly) from Schmidt (2007).

Case 1 Parameter	Lower95	Median	Upper95	Mean	SD
Baseline I to Intervention I					
rhoA1B1	-0.5395	-0.2486	0.0587	-0.2461	0.1544
muA1[1]	3.0293	3.1485	3.2613	3.1465	0.0587
muB1[2]	1.8517	2.0080	2.1569	2.0055	0.0770
sigmaA1[1]	0.3526	1.1993	4.4582	1.7668	2.6461
sigmaB1[2]	0.3657	1.2024	4.4829	1.7872	3.2339
precA1[1]	0.0000	0.6953	3.0459	1.0046	1.0073
precB1[2]	0.0000	0.6917	3.0179	1.0008	1.0035
rate_ratio A1B1	0.2779	0.3195	0.3759	0.3206	0.0267
Intervention I to Baseline II					
rhoB1A21	-0.7117	-0.0970	0.3796	-0.1314	0.2749
MuB1[1]1	1.8101	2.0071	2.1763	2.0017	0.0909
muA2[2]1	2.7570	3.0068	3.2494	3.0024	0.1240
sigmaB1[1]1	0.3644	1.1980	4.5399	1.8139	4.2350
sigmaA2[2]1	0.3613	1.2053	4.4675	1.7625	2.6295
precB1[1]1	0.0000	0.6967	3.0405	1.0086	1.0154
precA2[2]1	0.0000	0.6883	2.9954	1.0003	1.0017
rate_ratioB1A21	2.1692	2.7215	3.5212	2.7464	0.3642
Baseline II to Intervention II					
rhoA2B22	-0.9999	-0.4352	0.2859	-0.3962	0.3960
muA2[1]2	2.7829	3.0006	3.2218	2.9960	0.1141
muB2[2]2	1.2779	1.6450	1.9026	1.6246	0.1553
sigmaA2[1]2	0.3574	1.2040	4.3701	1.7729	3.6095
sigmaB2[2]2	0.3595	1.2084	4.4613	1.7590	2.3304
precA2[1]2	0.0000	0.6899	2.9708	0.9943	0.9920
precB2[2]2	0.0001	0.6849	2.9907	0.9933	1.0028
rate_ratioA2B22	0.1898	0.2573	0.3398	0.2573	0.0402

Table 5: BRR parameter estimates for disruptive behavior of second case (Albert) from Schmidt (2007)

Case 2 Parameter	Lower95	Median	Upper95	Mean	SD
Baseline I to Intervention I					
rhoA1B1	-0.2165	0.1362	0.4392	0.1243	0.1684
muA1[1]	3.2151	3.3580	3.4926	3.3555	0.0706
muB1[2]	1.3391	1.6190	1.8738	1.6158	0.1363
sigmaA1[1]	0.3536	1.2035	4.4032	1.7807	5.3976
sigmaB1[2]	0.3854	1.2051	4.4723	1.7705	2.9408
precA1[1]	0.0000	0.6904	3.0057	0.9998	1.0109
precB1[2]	0.0000	0.6886	2.9693	0.9934	0.9908
rhoA1B1	0.1395	0.1757	0.2282	0.1771	0.0241
Intervention I to Baseline II					
rhoB1A21	-0.5079	-0.0064	0.4301	-0.0153	0.2482
MuB1[1]1	1.3734	1.6289	1.8650	1.6233	0.1238
muA2[2]1	2.4001	2.7434	3.0322	2.7280	0.1743
sigmaB1[1]1	0.3571	1.2001	4.4583	1.7620	2.6381
sigmaA2[2]1	0.3683	1.2060	4.4512	1.7545	2.3087
precB1[1]1	0.0000	0.6943	3.0429	1.0054	1.0075
precA2[2]1	0.0001	0.6875	2.9728	0.9946	0.9952
rate_ratioB1A21	2.2286	3.0536	4.1994	3.0818	0.5319
Baseline II to Intervention II					
rhoA2B22	-0.8820	-0.2132	0.0974	-0.2610	-0.8820
muA2[1]2	2.5345	2.7684	3.0137	2.7683	2.5345
muB2[2]2	1.5365	1.8134	2.0647	1.8054	1.5365
sigmaA2[1]2	0.3420	1.2055	4.3828	1.7589	0.3420
sigmaB2[2]2	0.3697	1.1994	4.5158	1.7848	0.3697
precA2[1]2	0.0000	0.6881	3.0105	1.0022	0.0000
precB2[2]2	0.0000	0.6952	2.9584	0.9914	0.0000
rate_ratioA2B22	0.2903	0.3840	0.5120	0.3865	0.2903

Table 6. BRR parameter estimates for disruptive behavior of third case (Faith) from Schmidt (2007).

Case 3 Parameter	Lower95	Median	Upper95	Mean	SD
Baseline I to Intervention I					
rhoA1B1	-0.1362	0.1090	0.3416	0.1047	0.1230
muA1[1]	2.7519	2.9170	3.0844	2.9168	0.0849
muB1[2]	0.7859	1.2322	1.6332	1.2187	0.2173
sigmaA1[1]	0.3708	1.1990	4.5227	1.7958	5.4566
sigmaB1[2]	0.3612	1.1993	4.3887	1.7484	3.0317
precA1[1]	0.0000	0.6956	3.0313	0.9989	0.9982
precB1[2]	0.0000	0.6953	3.0041	1.0059	0.9968
rate_ratio A1B1	0.1274	0.1853	0.2664	0.1875	0.0380
Intervention I to Baseline II					
rhoB1A21	-0.1786	0.2992	0.9025	0.3229	0.3041
MuB1[1]1	0.5502	1.1727	1.6346	1.1383	0.2823
muA2[2]1	2.1899	2.8218	3.2149	2.7648	0.2727
sigmaB1[1]1	0.3650	1.2045	4.4031	1.7473	2.6589
sigmaA2[2]1	0.3793	1.2079	4.5261	1.7910	3.2236
precB1[1]1	0.0000	0.6893	2.9862	0.9949	0.9951
precA2[2]1	0.0000	0.6853	2.9702	0.9937	0.9938
rate_ratioB1A21	3.0647	5.1251	8.8367	5.3052	1.5244
Baseline II to Intervention II					
rhoA2B22	-0.9999	-0.1772	0.6964	-0.1761	0.4714
muA2[1]2	2.4388	2.9578	3.2191	2.9082	0.2093
muB2[2]2	0.6080	1.3380	1.7539	1.2755	0.3084
sigmaA2[1]2	0.3838	1.2066	4.5951	1.7860	2.8223
sigmaB2[2]2	0.3682	1.2028	4.3490	1.7778	4.2732
precA2[1]2	0.0000	0.6869	2.9932	1.0009	1.0031
precB2[2]2	0.0000	0.6913	2.9959	0.9999	0.9954
rate_ratioA2B22	0.1317	0.1976	0.2987	0.2010	0.0430

The BRR model produces three different effect sizes; for first introduction of the treatment, removal of the treatment, and re-introduction of the treatment. First, the results from the first case is discussed. For phases, A1 and B1, the posterior mean (SD) of the model intercepts are 3.1465 (0.0587) and 2.0055 (0.0770) respectively. These represent the mean of the log count of disruptive behavior, which is not very intuitive. To obtain the mean levels in original count (not provided by the model), one needs to exponentiate to obtain exp (3.1465) = 23.2545 for phase A1 and exp (2.0055) = 7.4298 for phase B1. This suggests, the estimated mean counts of disruptive behavior is less in phase B1 compared to phase A1.

The mean of the BRR effect size (i.e., as a ratio of raw counts) for the first case for A1B1 is 0.3206 (0.0267). This estimate is interpreted as the count of disruptive behavior in first treatment phase is 0.32 times the count in the first baseline phase. An alternative interpretation (similar to LRR) is, there is 68% reduction (i.e., 1.00 - 0.32) in disruptive behavior from A1 to B1 for first case due to the treatment.

The corresponding 95% high density interval for the ES is 0.2779 to 0.3759 times or in other words a reduction of 63% to 73% in disruptive behaviors in the B1 versus A1 phase. The posterior mean of the standard error for phase A1 (sigmaA1) was 1.7668 and for phase B1(sigmaB1) was 1.7872. The mean of the BRR effect size after removing the first treatment (B1A2 phases) is 2.7464 suggesting that the disruptive behaviors for the first case was on average 2.7464 times higher after removing the treatment. Similar interpretations can be made for results from phases B1A2 and A2B2.

For the second case, the posterior means of the intercepts in log counts for first pair of adjacent phases (A1B1) are 3.3555 and 1.6158 respectively. Equivalently, the

posterior mean of the intercepts in counts for phase A1 and B1 are 28.7 (exp (3.3555)) and 5.0 (exp (1.6158)) respectively. The posterior mean of the autocorrelation for phases A1B1 is 0.1243 with standard error of 0.1684. This suggests the estimated autocorrelation is most likely small and positive, although the precision of this estimate is somewhat low. The mean of the BRR effect size estimate for A1B1 is 0.1771 (0.0241), indicating the mean disruptive behavior in B1 phase was 0.1771 times or 17% of what was in the A1 phase. Alternatively, this suggests an 82% decrease in the disruptive behavior of the second case following introduction of first treatment. In similar lines, there is 61% decrease in the disruptive behavior when the treatment was re-introduced in phase B2 compared to the disruptive behaviors in phase A2. In the scenario of removal of the treatment (A2 phase), the posterior mean of the disruptive behaviors of the second case increased by 3.1 times after the treatment was removed. The disruptive behavior for the second case showed more of a reduction when the treatment was first introduced compared to the second introduction of the treatment.

For the third case, the mean of the intercepts while comparing A1B1 phases are exp (2.9168) and exp (1.2187) respectively, for B1A2 phases are exp (1.1383) and exp (2.7648) respectively, and for A2B2 phases are exp (2.9082) and exp (1.2755) respectively. The 95% credible interval for the posterior mean of the autocorrelation ranged from -0. 9999 to 0.9025. The mean of the BRR effect size estimate for phases A1B1 and A2B2 was 0.1875 and 0.2010 respectively, indicating 81% and 80% decrease respectively in the disruptive behavior following the CW-FIT treatment. When the treatment was removed, the counts of disruptive behavior increased by 5.3052 times in the A2 phase.

Overall, the largest reduction in disruptive behaviors following the treatment was for the third case, and largest increase in disruptive behaviors after removing the treatment was for the third case. The autocorrelation estimates for all the three cases are not very precise given the wide confidence interval.

4.1.3 NLB estimates

The NLB parameter estimates for each of the three cases and aggregated across all cases for Schmidt (2007) is shown in Table 7a and 7b. JAGS was used to draw 16,000 samples (with 4 parallel chains,12,000 values from each chain, and thinning of 3) from the posterior distribution of the parameters. The convergence of the parallel chains was examined using trace plots and PSRF, and for all the 51 parameter estimates, it was less than 1.05 suggesting convergence. For each of the parameters, the NLB model provides mean, standard error, median (50th percentile), and 95% credible intervals from the posterior distribution and shown in Tables 7a and 7b. The priors used for the intercepts and precision were presented previously in the methods chapter.

The NLB model produces parameter estimates in both log counts and raw counts. In the tables, all the variables beginning with “exp” are in original counts, variables sig0, sig1, sig3 are the standard deviation in the intercept, first, and second treatment effects across three cases respectively, and the remaining variables are in log counts. Similar to LRR and BRR, A1B1A2B2 is used for ease of explanation.

Table 7a: NLB parameter estimates of disruptive behavior for three cases for A1B1 phases from Schmidt (2007)

Parameter	Lower 95	Median	Upper 95	Mean	SD
Baseline I to Intervention I					
PhaseA1_mu_case[1]	3.0064	3.1475	3.2905	3.1471	0.0721
PhaseA1_mu_case[2]	3.2457	3.3711	3.4869	3.3703	0.0617
PhaseA1_mu_case[3]	2.7736	2.9194	3.0701	2.9185	0.0764
PhaseB1_mu_case[1]	1.8161	2.0086	2.1954	2.0076	0.0969
PhaseB1_mu_case[2]	1.4038	1.6491	1.8735	1.6462	0.1200
PhaseB1_mu_case[3]	1.0139	1.3575	1.6952	1.3525	0.1741
TrteffA1B1_case[1]	-1.3727	-1.1392	-0.9026	-1.1395	0.1199
TrteffA1B1_case[2]	-1.9843	-1.7221	-1.4549	-1.7241	0.1349
TrteffA1B1_case[3]	-1.9372	-1.5620	-1.1976	-1.5660	0.1886
exp_PhaseA1_mu_case[1]	20.0096	23.2770	26.6234	23.3290	1.6792
exp_PhaseA1_mu_case[2]	25.6805	29.1091	32.6834	29.1437	1.7978
exp_PhaseA1_mu_case[3]	15.9046	18.5293	21.4160	18.5674	1.4181
exp_PhaseB1_mu_case[1]	6.1058	7.4528	8.9337	7.4802	0.7230
exp_PhaseB1_mu_case[2]	4.0109	5.2025	6.4400	5.2246	0.6246
exp_PhaseB1_mu_case[3]	2.6396	3.8864	5.2791	3.9256	0.6802
exp_TrteffA1B1_case[1]	0.2485	0.3201	0.3995	0.3223	0.0387
exp_TrteffA1B1_case[2]	0.1333	0.1787	0.2282	0.1800	0.0243
exp_TrteffA1B1_case[3]	0.1368	0.2097	0.2917	0.2126	0.0400
PhaseA1_muA1_allcases	0.2287	2.2667	3.4859	2.0942	0.8760
PhaseB1_muB1_allcases	-1.1765	1.0656	2.8433	0.9511	1.0156
TrteffA1B1_allcases	-2.1345	-1.1943	-0.0858	-1.1431	0.5201
expPhaseA1_mu_allcases	0.2146	9.6479	24.9087	10.9379	7.5342
expPhaseB1_mu_allcases	0.0229	2.9024	11.2579	4.0648	4.7177
exp_TrtA1B1_allcases	0.0872	0.3029	0.8214	0.3732	0.2868
sig0_allcases	0.3572	1.0841	3.0064	1.3433	0.9016
sig1_allcases	0.3646	0.8298	1.8257	0.9488	0.4740

Table 7b: NLB parameter estimates of disruptive behavior for three cases for A2B2 phases from Schmidt data (2007)

Parameter	Lower 95	Median	Upper 95	Mean	SD
Baseline II to Intervention II					
PhaseA2_mu_case[1]	2.7976	3.0473	3.2827	3.0456	0.1233
PhaseA2_mu_case[2]	2.5308	2.8090	3.0718	2.8055	0.1386
PhaseA2_mu_case[3]	2.7036	2.9610	3.2095	2.9586	0.1297
PhaseB2_mu_case[1]	1.2718	1.6532	2.0273	1.6486	0.1929
PhaseB2_mu_case[2]	1.4771	1.8061	2.1163	1.8018	0.1627
PhaseB2_mu_case[3]	0.9766	1.4066	1.8174	1.3995	0.2154
TrteffA2B2_case[1]	-1.8562	-1.3930	-0.9716	-1.3970	0.2263
TrteffA2B2_case[2]	-1.4199	-1.0030	-0.5979	-1.0037	0.2103
TrteffA2B2_case[3]	-2.0604	-1.5545	-1.0875	-1.5591	0.2484
exp_PhaseA2_mu_case[1]	16.0904	21.0593	26.2830	21.1832	2.6034
exp_PhaseA2_mu_case[2]	12.2454	16.5926	21.2183	16.6940	2.3031
exp_PhaseA2_mu_case[3]	14.6258	19.3180	24.4032	19.4330	2.5140
exp_PhaseB2_mu_case[1]	3.3566	5.2235	7.2765	5.2962	1.0136
exp_PhaseB2_mu_case[2]	4.2312	6.0865	8.0902	6.1403	0.9919
exp_PhaseB2_mu_case[3]	2.4993	4.0819	5.9214	4.1467	0.8835
exp_TrteffA2B2_case[1]	0.1506	0.2483	0.3695	0.2537	0.0574
exp_TrteffA2B2_case[2]	0.2330	0.3668	0.5361	0.3747	0.0794
exp_TrteffA2B2_case[3]	0.1185	0.2113	0.3237	0.2168	0.0540
PhaseA2_mu_allcases	-0.4202	2.0259	4.0618	1.9371	1.1341
PhaseB2_mu_allcases	-1.6023	0.9795	3.3427	0.9123	1.2500
TrteffA2B2_allcases	-2.0467	-1.0692	0.0160	-1.0248	0.5206
expPhaseA2_mu_allcases	0.0381	7.5827	38.0824	12.2502	16.5345
expPhaseB2_mu_allcases	0.0061	2.6630	17.4141	5.1304	9.1370
exp_TrtA2B2_allcases	0.0886	0.3433	0.9103	0.4190	0.3217
sig3_allcases	0.3641	0.8268	1.7844	0.9418	0.4622

In the tables, PhaseA1_mu_case[1] is the level of disruptive behavior in log counts during the first baseline (Phase A1) for first case. The mean (used as a point estimate) for PhaseA1_mu_case[1] can be interpreted as the average log count of disruptive behaviors during the first baseline phase for first case, and is 3.1471. Instead of interpreting in log counts, exp_PhaseA1_mu_case[1] provides direct interpretation i.e., the mean counts of disruptive behavior in the first baseline phase for first case is 23.3290 with a standard error of 1.6792. Similarly, the mean number of disruptive behaviors in the first treatment phase for the first case is 7.4802 (0.7230), suggesting a decline in disruptive behaviors from first baseline to first treatment phase. For each case, the variables exp_PhaseB1_mu, exp_PhaseA2_mu, and exp_PhaseB2_mu provide the point estimates and credible intervals of the counts of disruptive behavior in the first treatment, second baseline, and second treatment respectively. The identifier for each case is given in the square bracket [ ] following the name of the variable (e.g., [2] corresponds to second case, and [3] corresponds to third case) in column 1 of Table 7a and 7b.

The estimates of the effect of first and second introduction of the CW-FIT intervention are given by variables TrteffA1B1 and TrteffA2B2 respectively. The mean of the first introduction of treatment effect for first case (exp_TrteffA1B1_case[1]) is 0.3223 (0.0387). This suggests there is a 68% reduction in the disruptive behavior of the first case when the treatment is first introduced. The corresponding median (0.3201) is nearly equal to the mean. The 95% credible interval indicates there is 95% chance that the true treatment effect (first introduction of treatment) for first case is within the interval (0.2485, 0.3995). Similarly, for the first student, when the treatment was re-introduced second time (A2B2), there was 75%

reduction in the disruptive behavior with 95% credible interval of (0.1506, 0.3695). The phase means and treatment effects can similarly be interpreted for the other two cases. The largest reductions in the disruptive behavior after the first introduction of the treatment was for the first case and after re-introduction of the treatment, it was for the second case.

The NLB also provides estimates across all the three cases. Across all the three cases, the mean counts of disruptive cases along with standard error in phases A1, B1, A2, and B2 are 10.9379 (7.5342), 4.0648 (4.7177), 12.2502 (16.5345), and 5.1304 (9.1370) respectively. Though the standard error was high for all the phases, it was highest for phase A2. The average treatment effect (exp_TrtA1B1) after first introduction of the treatment across all the three cases was 0.3732 (0.2868), indicating there is 63% reduction in the disruptive behavior across the three cases. After the treatment was re-introduced the second time, there was 59% reduction in the disruptive behavior across the three cases. The standard deviation in the treatment effect across the three cases during first and second introduction of the treatment was 0.95 and 0.94 respectively.

The program produces estimates in both the log and original count thus enabling an easier understanding of the estimates produced. Also, the Bayesian approach allows one to estimate the full sampling distribution of each parameter as the posterior distribution for each parameter of interest.

4.2 Comparisons of Effect Sizes Using a Common Metric

Both NLB and BRR estimate a ratio of average count outcomes (level) in treatment phase to baseline phase (of adjacent AB phases). LRR is a natural log of ratio of average level of count outcomes in treatment and baseline phases. Because all

three estimates involve ratios of counts, it is possible to convert the ES estimates to a common metric. Since NLB produces both the log and normal count parameter estimates, both BRR and NLB can be transformed to NLB with the following formulas:

$$\text{NLB ES (count metric)} = \exp(\text{LRR ES}) \quad (61)$$

and

$$\text{NLB ES (count metric)} = \text{BRR ES} \quad (62)$$

This gives a general conversion and using these formulae it can be seen that the outputs from the models (as rescaled ES estimates) are approximately equal. However, these equations cannot be directly applied to the standard error estimates since the transformation is non-linear. Nor can the standard error of LRR be directly compared with BRR and NLB, because they are on different scales. Between BRR and NLB, for all the three cases, the SE of the effect size estimate is least for BRR and thus the 95% CI is narrow for BRR compared to both NLB. Alternatively, the equations above can be applied to the lower and upper limits of 95% credible intervals from all the three methods in order to make comparisons of precision of the ES estimates across the models.

It is important to note that, Pustejovsky (2018) expressed caution regarding the interpretation of SE of LRR obtained from a single SCD study. This is not an issue when LRR is used for meta-analysis as it uses robust variance estimation. However, since the Schmidt (2007) data is based on a single estimate for each case, the SE from LRR might not be a good estimate of the true variability of the sampling distribution.

Using the above formulas, the LRR ES and corresponding 95% CI in log units are transfomed to a ratio of raw counts. The mean and 95% CI of the ES estimates

from the three methods are presented in Figures 6 to 11. It can be clearly seen for all the three cases across both A1B1 and A2B2 comparison, the ES point estimates are almost equal from the three methods. The 95% credible intervals are narrower for BRR compared to NLB and LRR . For Case 2, A2B2 phases, the 95% CI for the LRR is (0.1351 – 1.0950), which includes 1 and suggests that there is no statistically significant difference in the mean level of disruptive behaviors for the second case across the A2 and B2 phases. However, this is despite the LRR ES (Table 3) showing that there is almost a 62% reduction in the disruptive behavior following the treatment. This finding suggests that the LRR SE estimates for one study are imprecise, and perhaps too conservative. However, the results also suggest that the point estimates from the LRR method are similar to the point estimates from the other two methods, after converting to a common metric. While the conversion formula can be used to compare the ES estimates from the three methods, these formulaes are not sufficient to calculate the NLB or BRR estimates without first estimating the NLB and BRR models via MCMC and then converting results to the same metric and comparing the estimates.

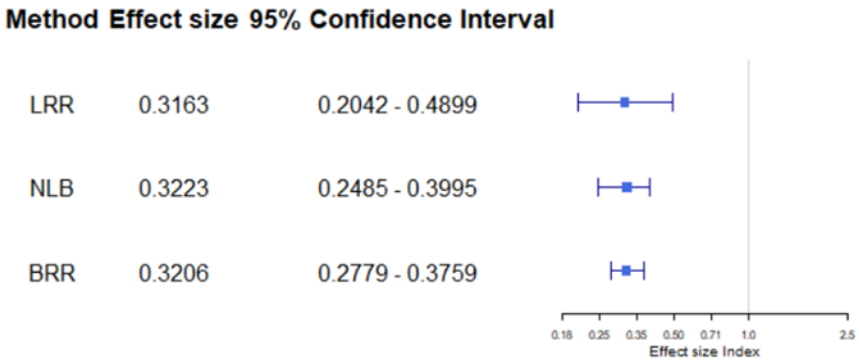

Figure 6: Baseline I to Intervention I (A1B1) Effect size estimates for first case

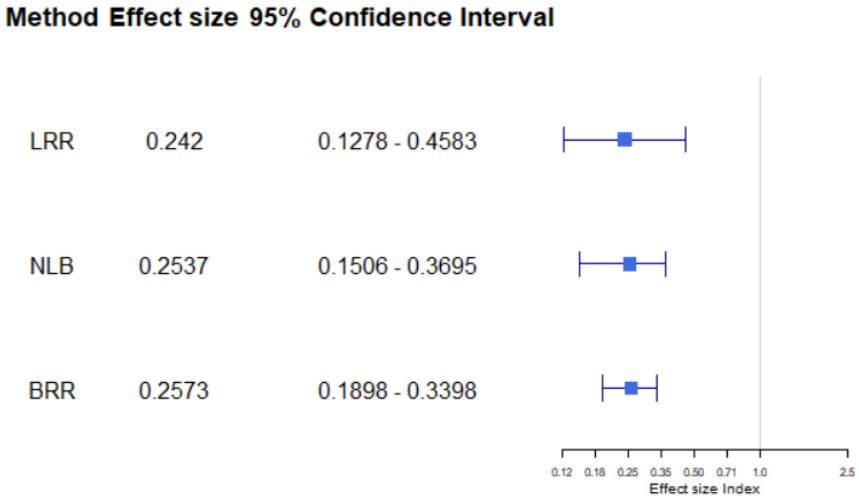

Figure 7: Baseline II to Intervention II (A2B2) Effect size estimates for first case

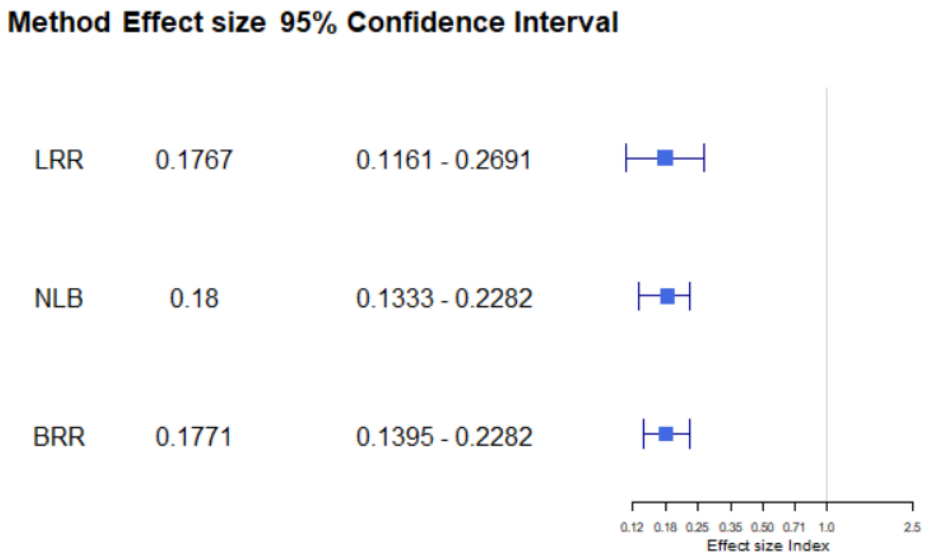

Figure 8: Baseline I to Intervention I (A1B1) Effect size estimates for second case

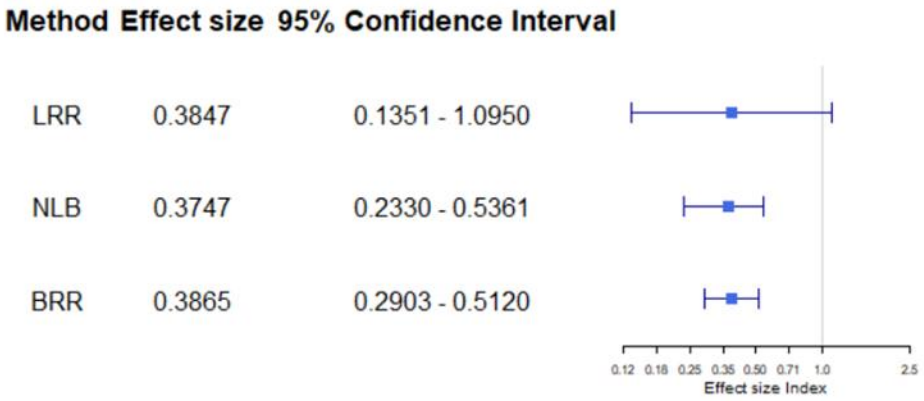

Figure 9: Baseline II to Intervention II (A2B2) Effect size estimates for second case

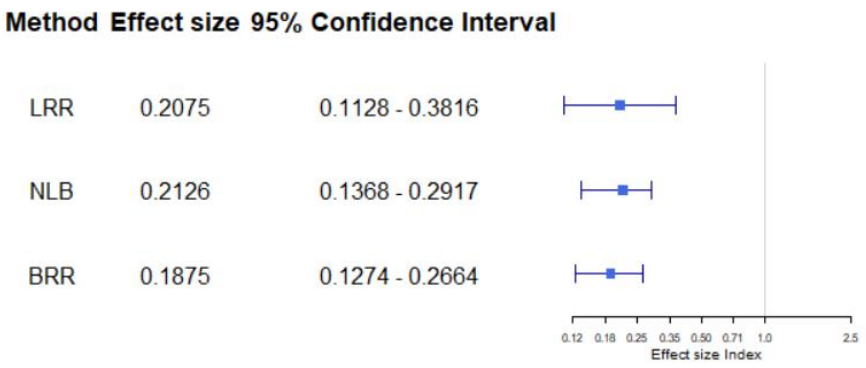

Figure 10: Baseline I to Intervention I (A1B1) Effect size estimates for third case

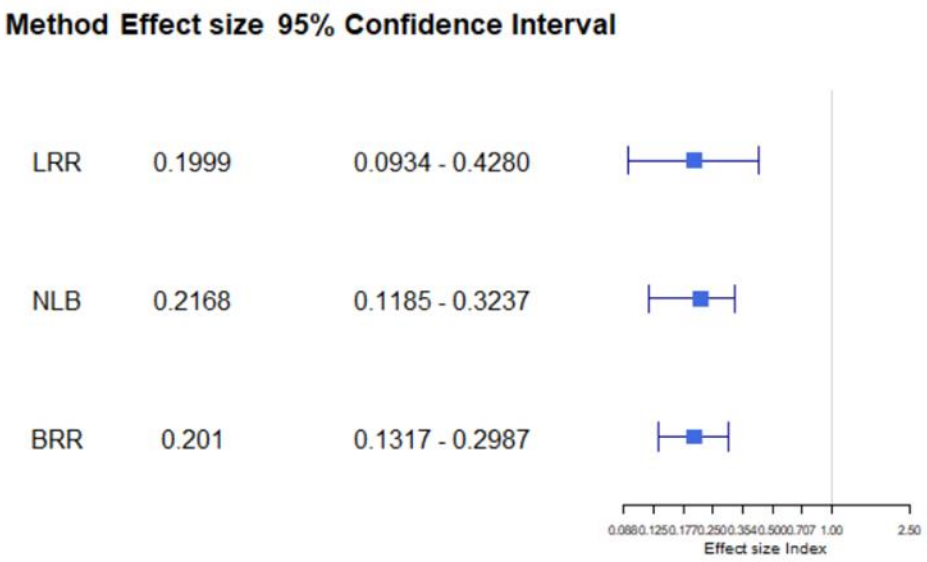

Figure 11: Baseline II to Intervention II (A2B2) Effect size estimates for third case

4.3 Simulation results

The aim of the 2nd research question is to examine the potential problems that one may encounter with the effect size estimates and standard error when using the three ES methods with autocorrelated and overdispersed simulated SCD data. This section presents the results for the 2nd research question.

In the generated datasets for all the 3 cases under all 12 simulation conditions for the A1B1A2B2 phases, there were no missing values and none of the generated data points equaled zero. The histogram of the generated 1000 datasets for a case under the twelve simulation conditions for A1B1A2B2 phases are presented in the Appendix D. Bias, RMSE, and coverage rates were calculated to determine how well the three different effect size estimators performed. The range of the 95% CI was also calculated. These quantities were calculated separately for effect size and 95% CI from phases A1B1 and phases A2B2. Each of these measures will be described next.

4.3.1 Bias and RMSE

As stated in the methods section, bias was calculated as the difference of the point estimate (estimated effect size) and the true parameter value. The average bias was calculated as the average across all three cases across all iterations. Figure 12, presents the average bias for the LRR, BRR, and NLB effect sizes of A1B1 phases for the 12 simulation conditions. For all the three methods, the average bias across all the three cases is very close to zero suggesting that the effect size estimates are almost equal to the true effect size value on average. Even in the worst cases, the bias for the point estimates were typically below .03 suggesting that all three methods produce reasonably unbiased estimates across all conditions tested. In general, bias seems to move upward with increase in overdispersion and is larger for higher autocorrelation

values. For all autocorrelation values, when overdispersion is less than 0.3, BRR has the lowest average bias and when overdispersion equals 0.3, LRR has the smallest bias as compared to BRR and NLB. The largest average bias was 0.0311 for NLB with overdispersion and autocorrelation equal to 0.3 and 0.4 respectively. Similarly, Figure 13 presents the average bias for effect size from the A2B2 phases. The trend is slightly different in that for all the 12 data conditions, LRR has the lowest average bias. As opposed to A1B1 phases, BRR has the largest average bias with overdispersion when autocorrelation is equal to 0.3 and 0.4 respectively. The RMSE of A1B1 effect size (Figure 14), suggests that the three methods have similar RMSE. Both NLB and LRR have slightly smaller RMSE values as compared to BRR indicating more accurate estimation. The RMSE of A2B2 effect size (Figure 15), clearly shows smaller RMSE values of NLB and LRR as compared to BRR indicating more accurate estimation. Compared to A1B1 phases, it is clearer in A2B2 phases. This may be because there is a smaller sample size in A2 and B2 phases compared to A1 and B1 phases (see Table 1) and BRR is most affected by smaller sample size. Clearly, the RMSE increases as both autocorrelation and overdispersion increases.

The average log bias of A1B1 effect size (Figure 16) is fairly close to zero for all the three methods. The average log bias of the A1B1 effect size for LRR and most of BRR (except for overdispersion value of 0.3) is negative. NLB has a positive average log bias of A1B1 effect size for all simulation conditions. The average log bias of A2B2 effect size (Figure 17) is close to zero but has both positive and negative values for all the three methods. In addition to the effect of autocorrelation and overdispersion, the difference in sample size for A1B1 and A2B2 phases might have also contributed to the log bias of the three methods. The log RMSE of A1B1 and

A2B2 phases (Figures 18 and 19 respectively) indicate NLB has the smaller values in general and thus indicates among the three methods, NLB provides more accurate estimation. The log RMSE for A2B2 phases is slightly higher than the log RMSE of the A1B1 phases. This might suggest, again, the influence of sample size on these performance measures.

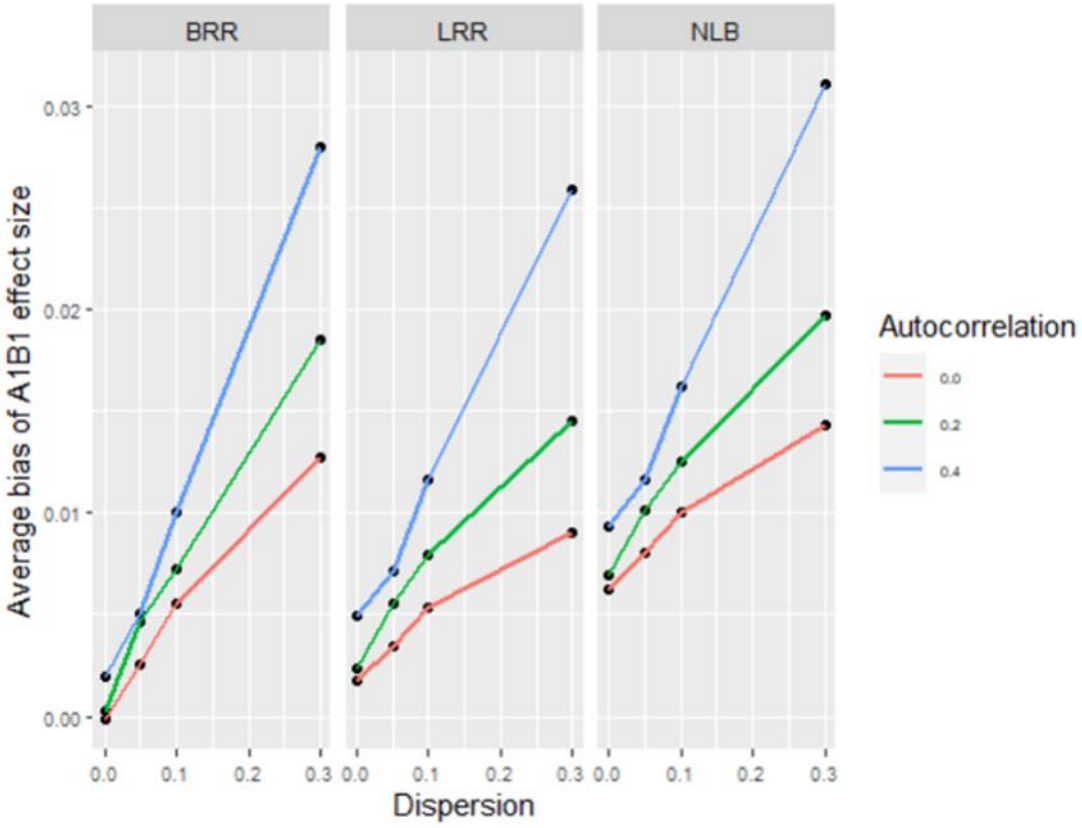

Figure 12: Average bias of LRR, BRR, and NLB effect sizes for A1B1 phases under simulation conditions.

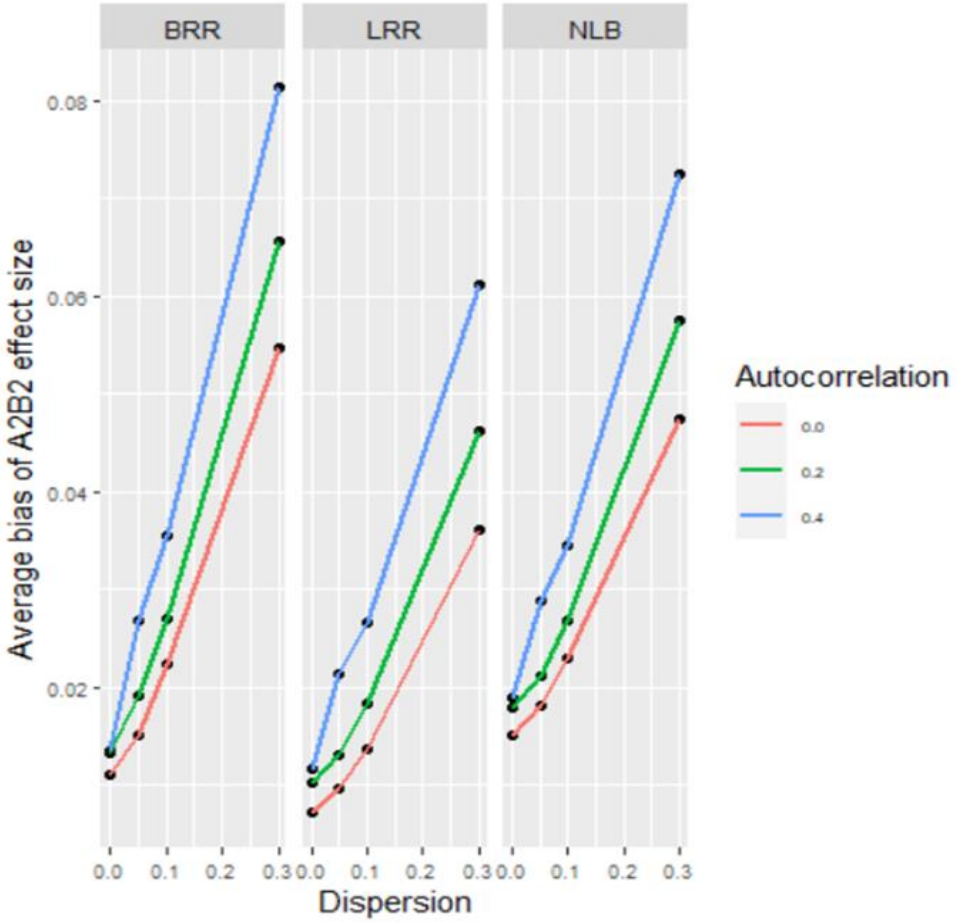

Figure 13: Average bias of LRR, BRR, and NLB effect sizes for A2B2 phases under simulation conditions

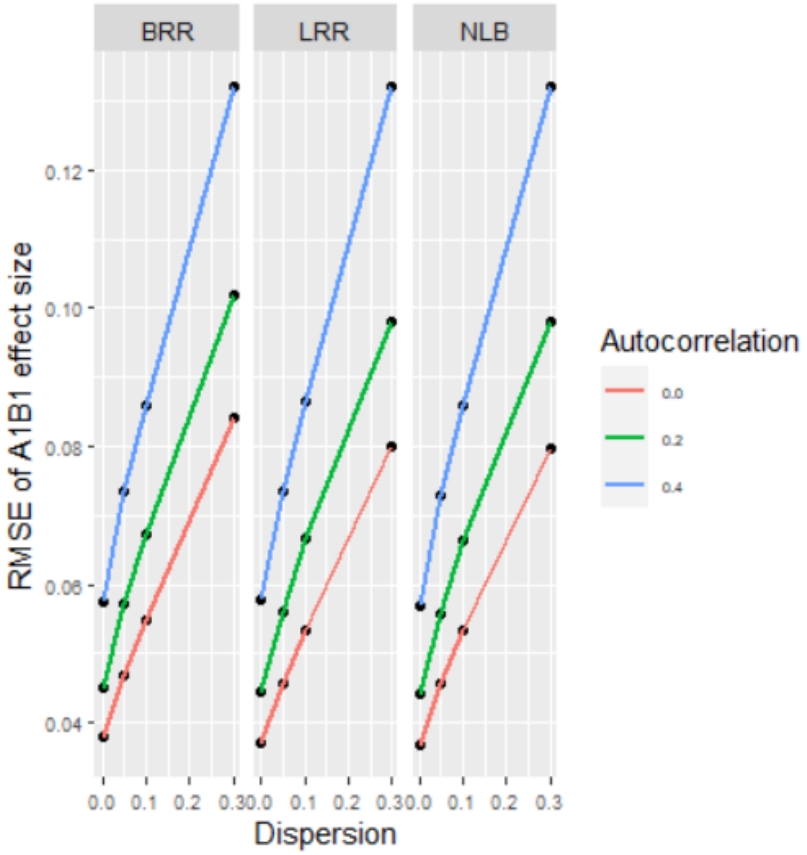

Figure 14: RMSE of LRR, BRR, and NLB effect sizes for A1B1 phases under simulation conditions.

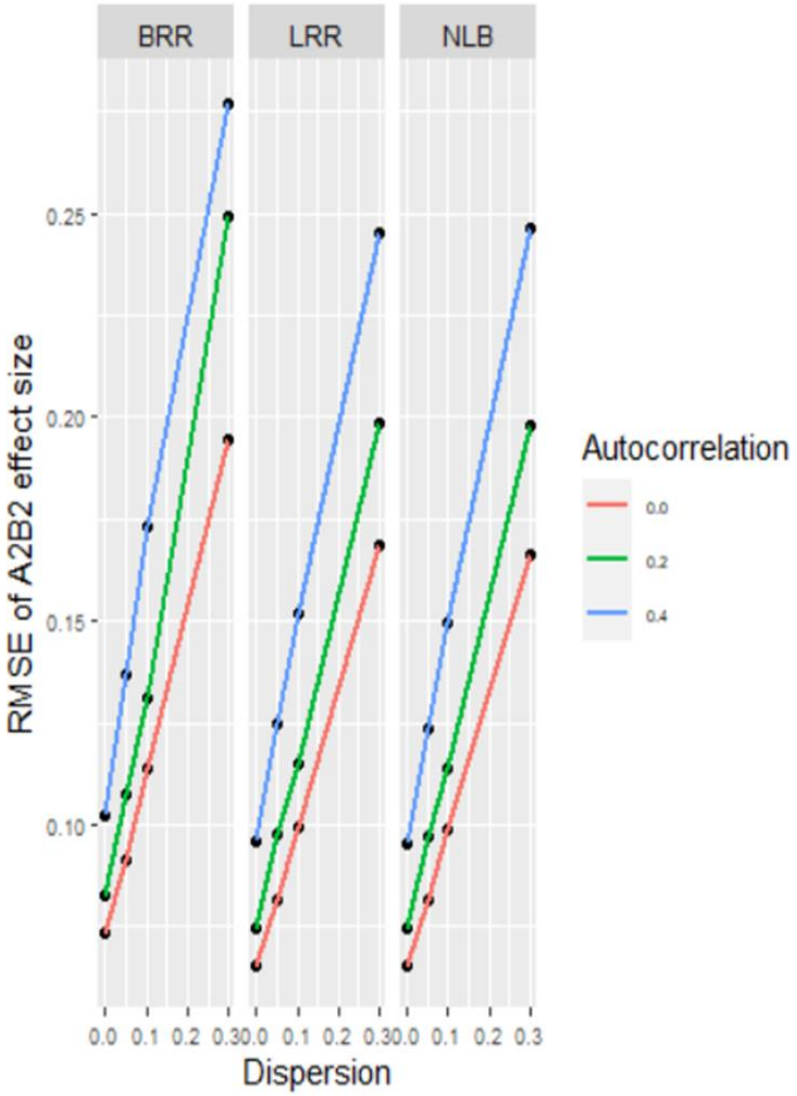

Figure 15: RMSE of LRR, BRR, and NLB effect sizes for A2B2 phases under simulation conditions.

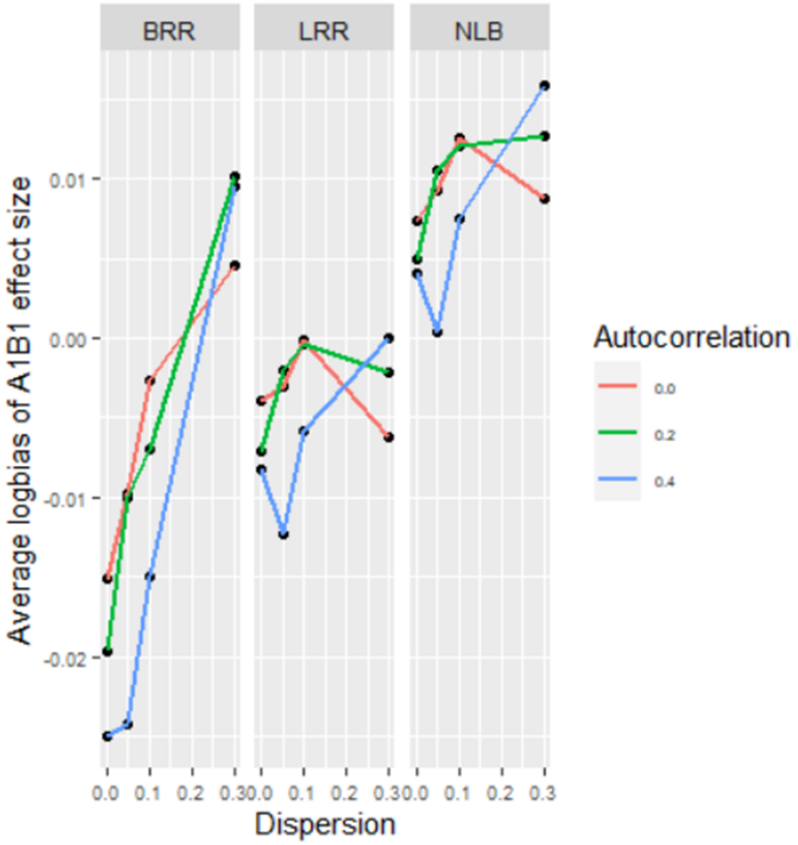

Figure 16: Average log bias of LRR, BRR, and NLB effect sizes for A1B1 phases under simulation conditions.

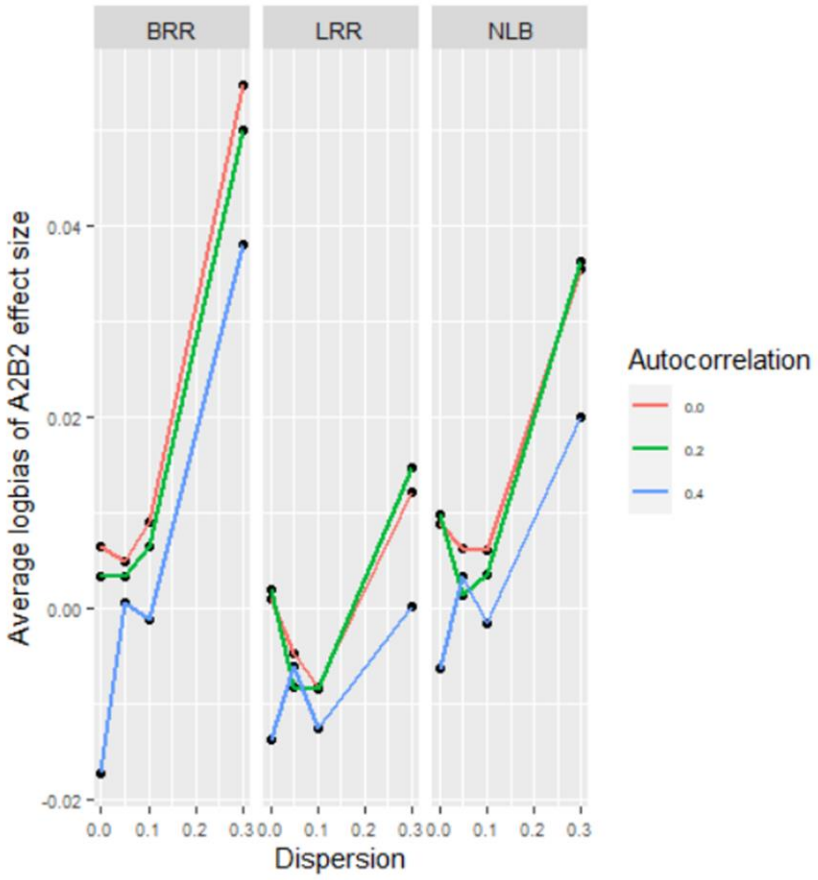

Figure 17: Average log bias of LRR, BRR, and NLB effect sizes for A2B phases under simulation conditions.

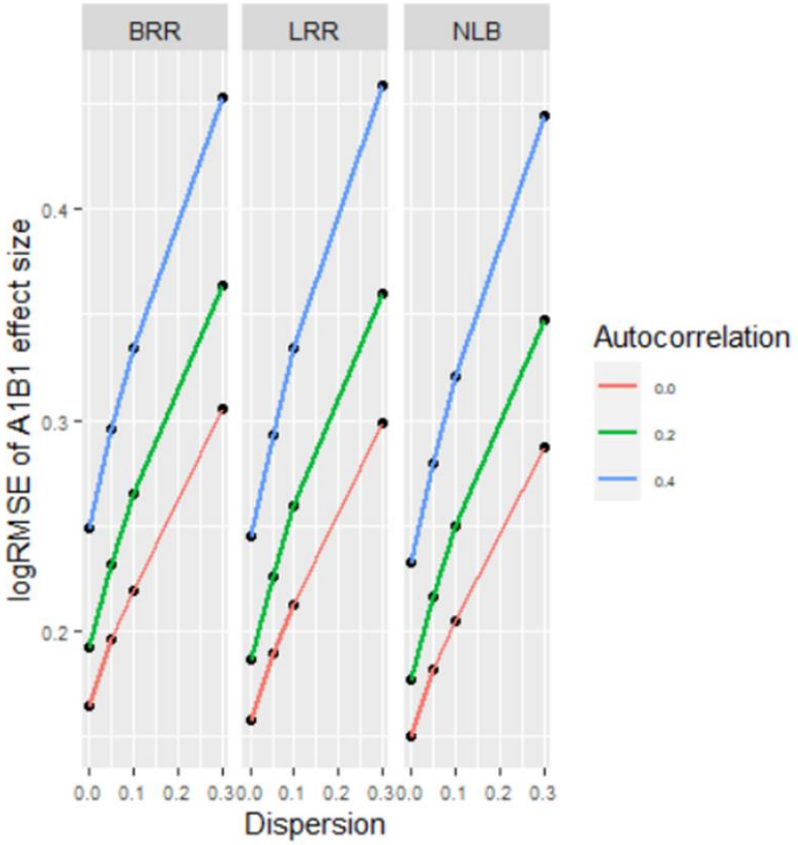

Figure 18: Log RMSE of LRR, BRR, and NLB effect sizes for A1B1 phases under simulation conditions.

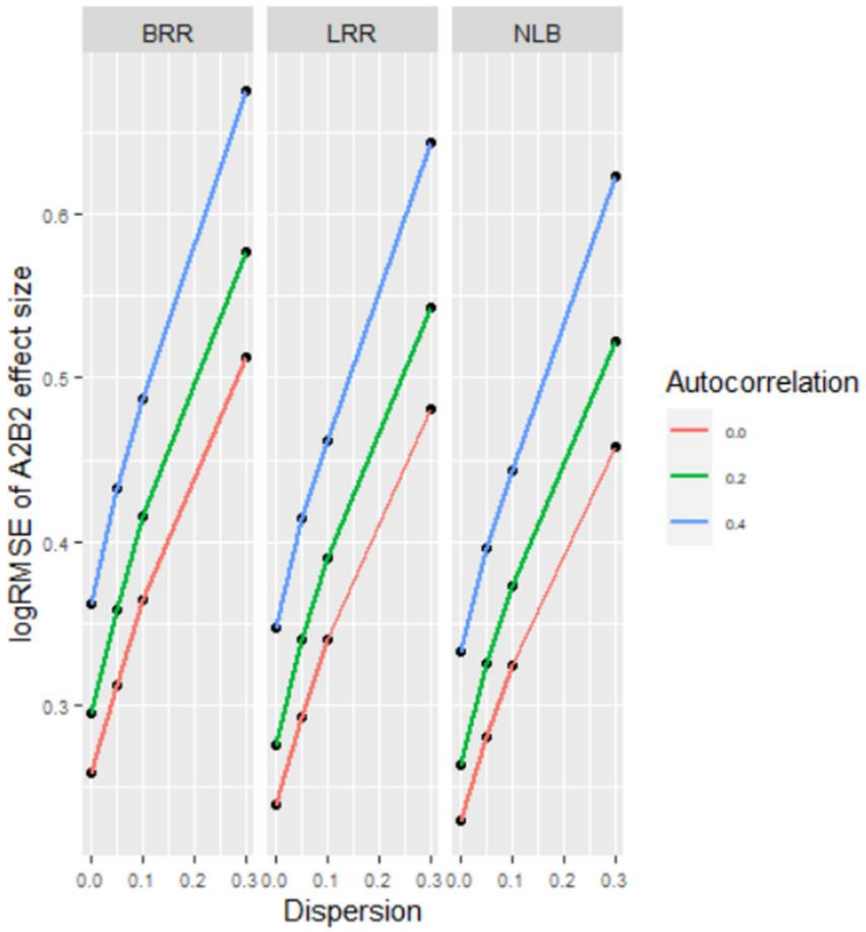

Figure 19: Log RMSE of LRR, BRR, and NLB effect sizes for A2B2 phases under simulation conditions.

4.3.2 Coverage

The Coverage indicator shows the proportion of 95% confidence intervals from LRR and 95% credible intervals from BRR and NLB that contain the true effect size value. The coverage for phases A1B1 and A2B2 are presented in figures 20 and 21 respectively, and results are similar. Among all the three effect size methods, only NLB has a coverage rate of 95% for one condition (no autocorrelation and negligible overdispersion). In general, the coverage of BRR credible intervals is lower than that of LRR and NLB. The coverage of BRR and NLB steadily declines as overdispersion and autocorrelation increases, with BRR having steeper decline than NLB for the same data conditions. The coverage rate of LRR decreases as the autocorrelation increases. It is important to note the change in overdispersion does not seem to affect the coverage rate of LRR as shown by a somewhat steady coverage across the different overdispersion values for the different autocorrelation levels.

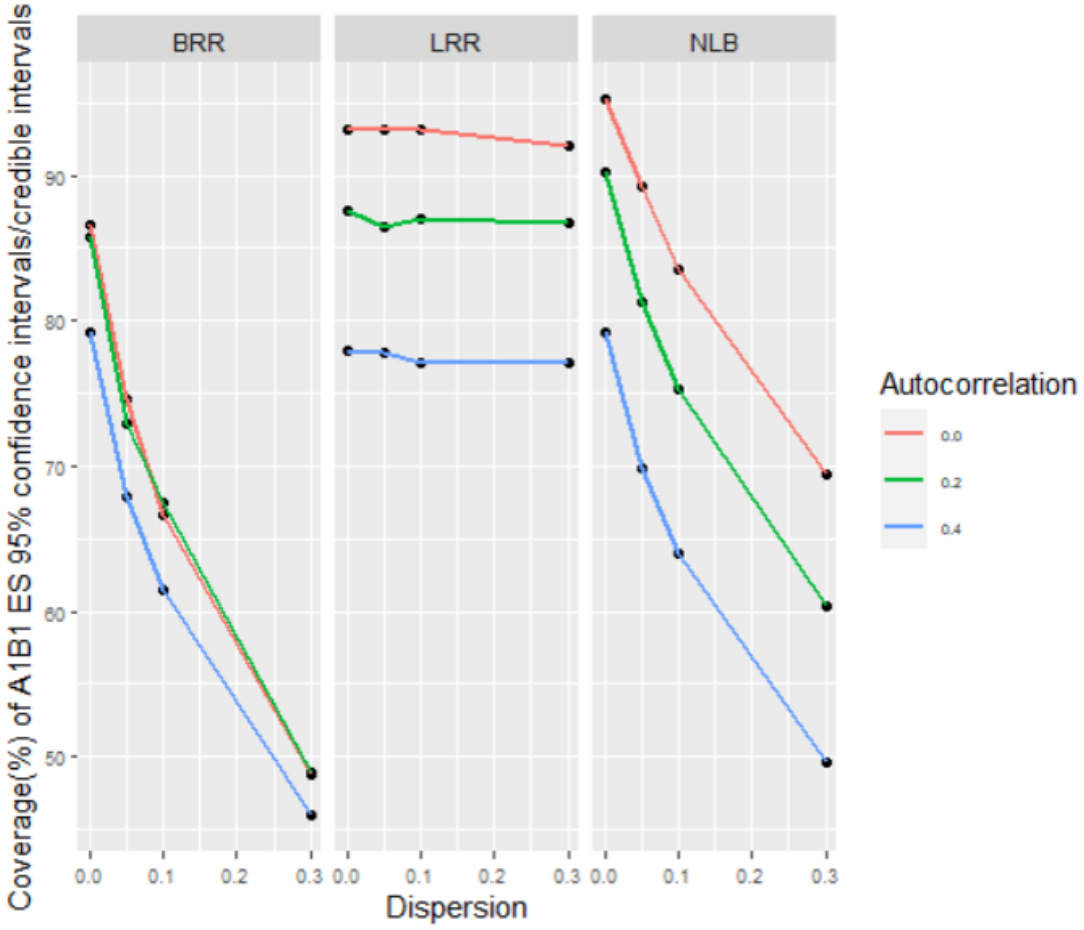

Figure 20: Coverage of LRR, BRR, and NLB effect sizes 95% confidence intervals/credible intervals for A1B1 phases under simulation conditions

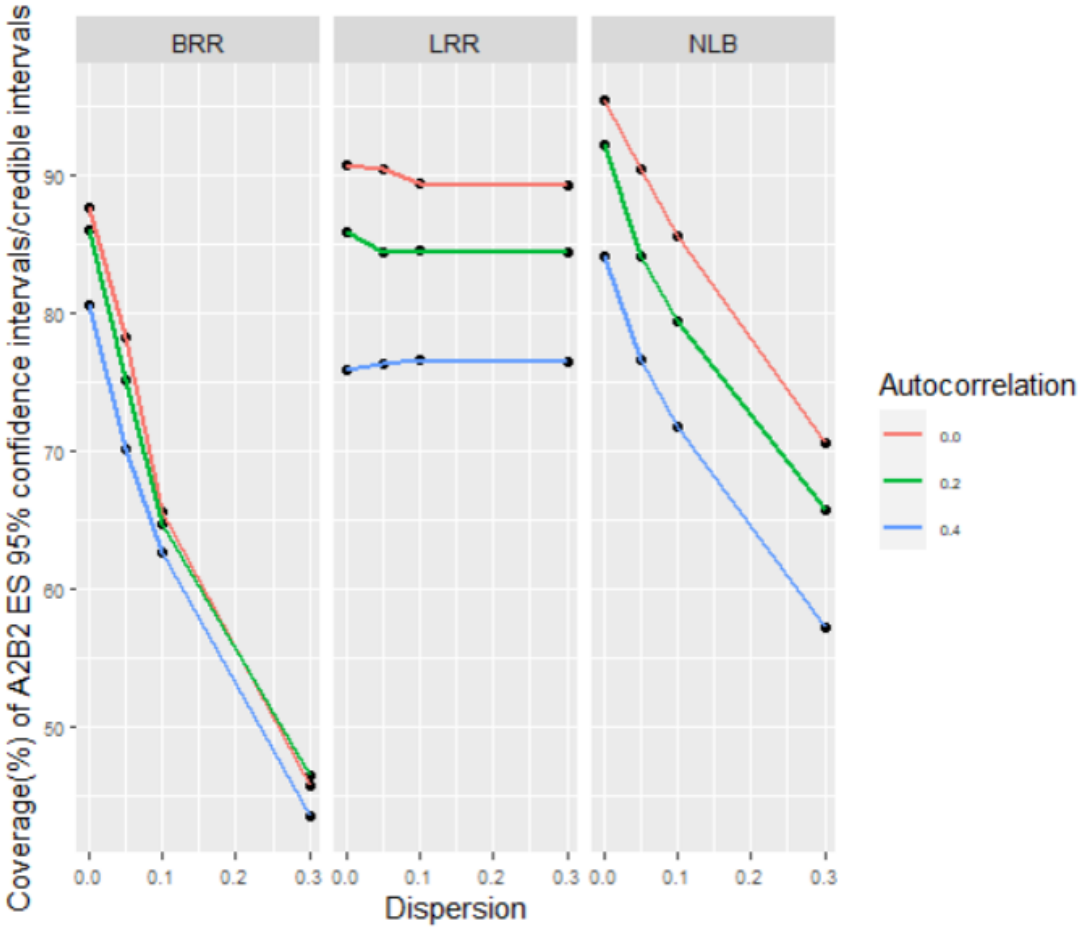

Figure 21: Coverage of LRR, BRR, and NLB effect sizes 95% confidence intervals/ credible intervals for A2B2 phases under simulation conditions

4.3.3 Average range of 95% confidence/credible intervals

The average range of the 95% confidence intervals was calculated as the difference between the 97.5th and 2.5th percentile. The exponentiation of the LRR 95% confidence interval endpoints transforms them to the same scale as the BRR and NLB. Thus, taking the average range of the ends of the 95% confidence interval from the three methods allows everything to be in the same metric of the confidence range of the counts. For phases A1B1, the average range was similar for NLB and LRR for the three autocorrelation values when overdispersion was negligible as shown in Figure 22. As the overdispersion increased, the average 95% confidence interval range of LRR increased sharply. Similarly, for phases A2B2, as shown in Figure 23, the average 95% confidence interval range of LRR was higher than BRR and NLB as the overdispersion increased.

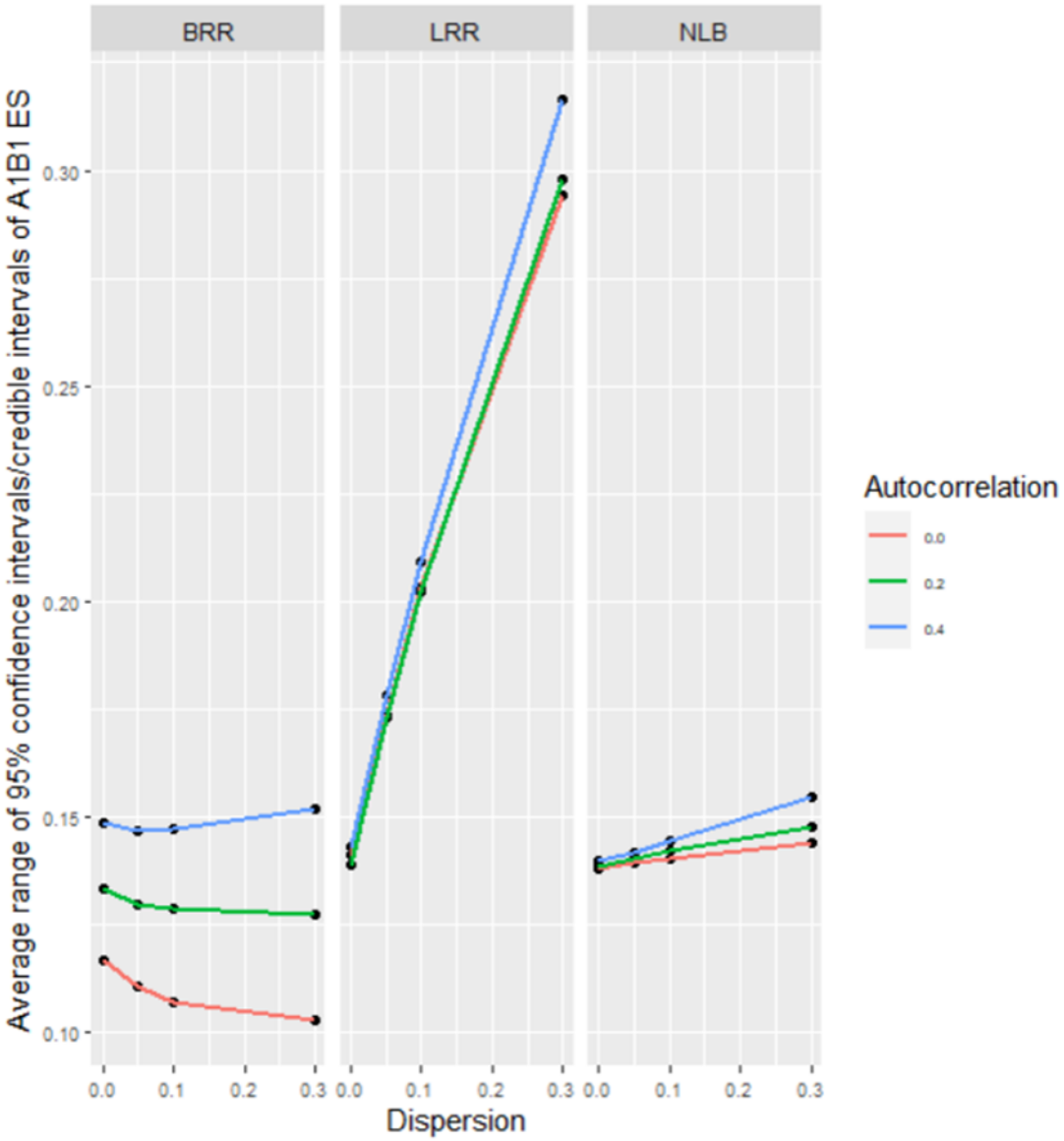

Figure 22: Average range of 95% confidence intervals/credible intervals for A1B1 phases for LRR, BRR, and NLB effect sizes under simulation conditions

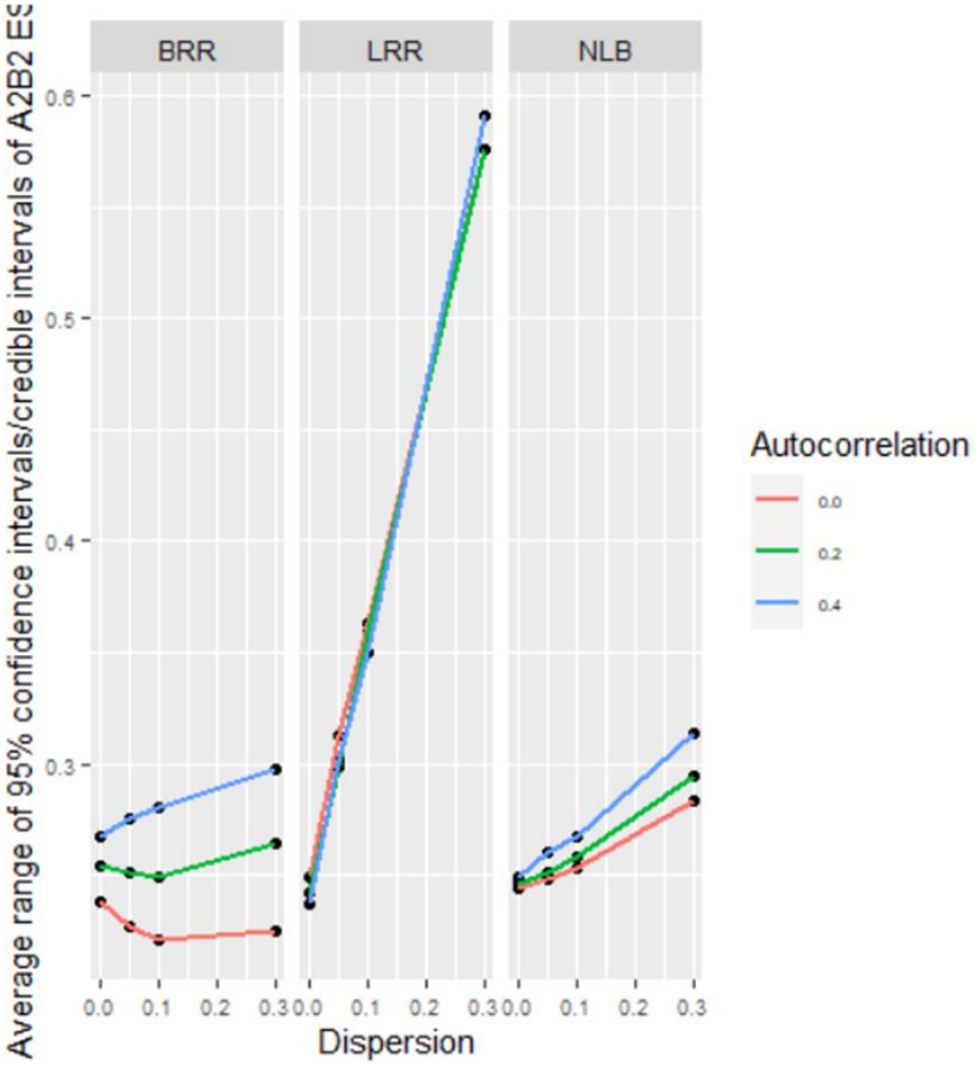

Figure 23: Average range of 95% confidence intervals / credible intervals for A2B2 phases for LRR, BRR, and NLB effect sizes under simulation conditions.

4.3.4 Coverage of autocorrelation estimate from BRR

The coverage of BRR autocorrelation estimate 95% credible intervals are equal to or close to 95% for all the three autocorrelation values with negligible overdispersion for both A1B1 and A2B2 phases. As the overdispersion increases there is a steep decline in the coverage for all the autocorrelation values.

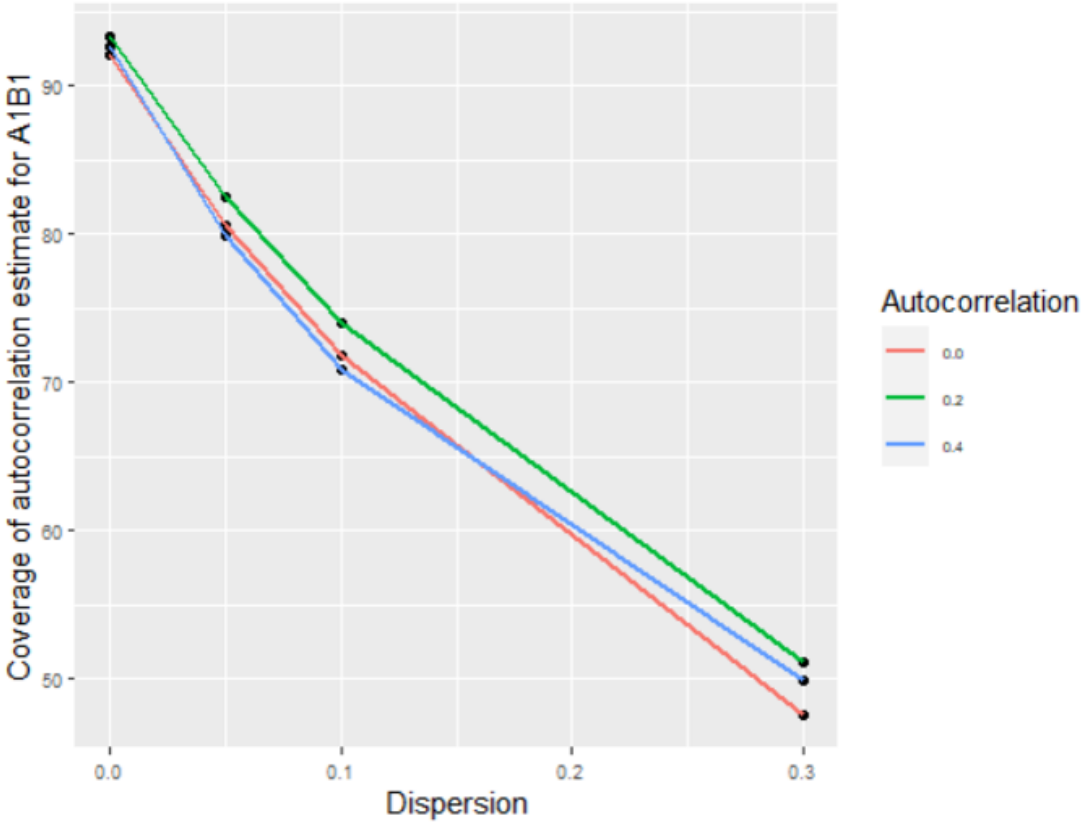

Figure 24. Coverage of autocorrelation estimate of A1B1 phases from BRR model.

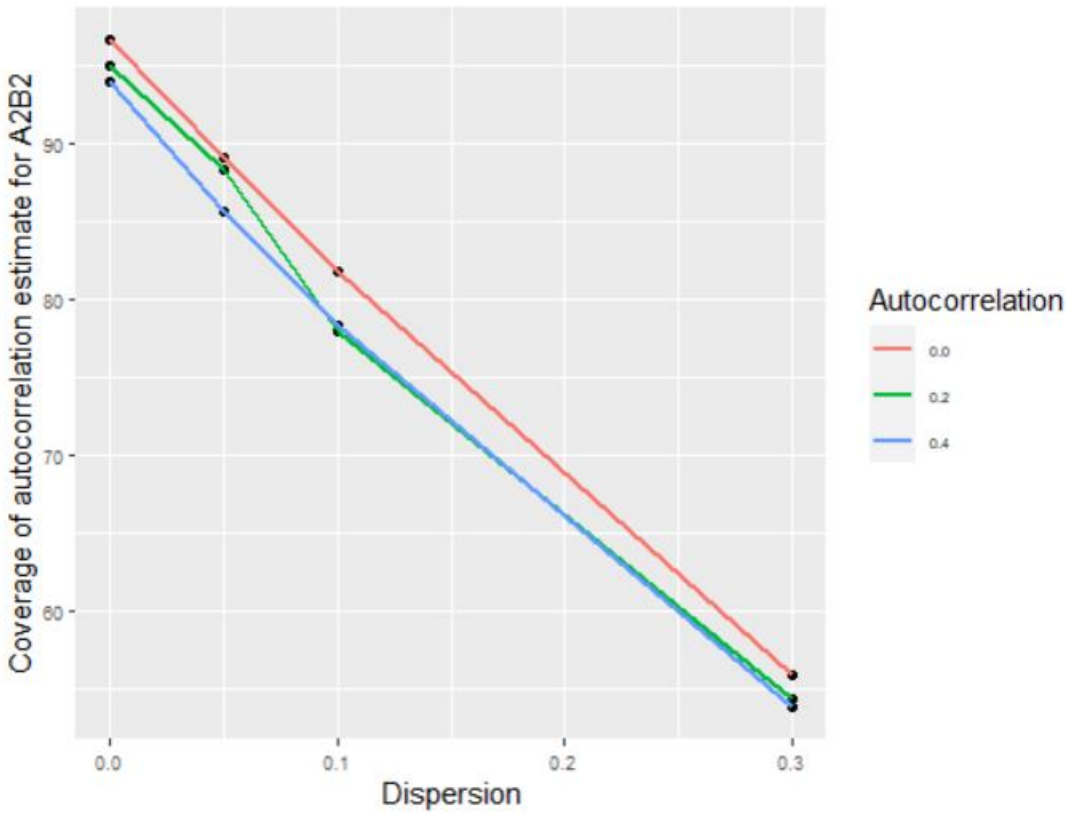

Figure 25. Coverage of autocorrelation estimate of A2B2 phases from BRR model.

4.4 Benefits and Challenges

All the three ES methods use statistical analysis to estimate the SCD effect size and its standard error. This information can be used to determine the effectiveness of a treatment under study as stated in the WWC standards and procedures handbook (WWC, 2022a). While it is not feasible to exhaustively list here all the benefits and challenges with each method, some of the points (ease of using a method, types of estimates produced, etc.) for each of the three methods are discussed.

4.4.1 LRR method

LRR is the simplest among the three ES methods under comparison, both in terms of estimation and in terms of interpretation. It can be used for a single SCD study or for meta-analysis. LRR can be used with both MBD and ABAB designs and with outcomes that are on a ratio scale. Thus, LRR ES can be calculated with frequency counts, rates, proportions, or percentage metrics. It has an existing R package “SingleCaseES”, which is user-friendly and one reasonably familiar with R can quickly and easily analyze their data. The options within the R package allow a user to produce ES estimates that are based on truncated means and corrected for small sample bias. In terms of the outputs, the LRR ES are provided in log metric and in percentage of change metric, which is an intuitive metric that is easier to understand and explain. However, presenting results in percentage change might pose a challenge in interpretation when there is a large value for the percentage change estimate. LRR does not provide estimates of the phase means, but users can calculate those fairly easily themselves. Alternatively, as there is likely interest in producing phase mean estimates, this could be easily added to the R package.

4.4.2 NLB method

NLB uses multilevel modeling and thus can answer questions such as “what is the average treatment effect across the cases within a study”, “How much variability is there in the effect size across cases” in-addition to providing effect sizes for each case within a study. As it employs a Bayesian approach, it can be used to produce informative conclusions such as “the probability that an effect of the treatment is positive”, or “the probability that an effect of the treatment is larger/smaller than ‘X’”. It can be used with count outcome that follows a Poisson or binomial distribution or a more complex distribution. It can also be used for continuous outcomes with a nonlinear relationship by specifying the appropriate functional form of the model (for details, refer Rindskopf, 2014). It can be used with both MBD and ABAB designs. Using NLB one can answer interesting questions, and one can even ‘accept’ the null hypothesis, but the Bayesian approach has a steep learning curve. Using multi-level modeling coupled with Bayesian estimation might be a challenge in itself for many researchers. In addition, users need to adapt existing WINBUGS code, write their own programs to run this model, or use some existing package/programs. As an example, researchers can use brms (Burkner P, 2021) R package that can be used to fit Bayesian multilevel models with a range of distributions and link functions. This uses Stan to fit the model, and the stan codes are generated automatically. Additionally, a major challenge for researchers can be which priors to use while using the Bayesian methods as using vague versus informative priors might produce different estimates. A study including a sensitivity analysis with different priors is yet to be done.

4.4.3 BRR method

This is the only method that accounts for autocorrelation. Since this method uses Bayesian estimation like the NLB method, it can also provide answers to interesting questions and one can also accept the null hypothesis. Moreover, R codes available in the GitHub site produce ES estimates when the treatment is introduced and when the treatment is removed. The program also provides a nice graph of the posterior density plots of the effect size estimate (not shown in this dissertation). If a researcher chooses, then they can specify the region of practical equivalence (ROPE), and conduct hypothesis testing. This is a within-subject effect size that can be used with an ABAB study design. The main challenge with this is there is some discrepancy between the model described in the paper and that estimated via the R codes on GitHub. If this is corrected, then the end user can easily use the R codes available to obtain the phase means, ES point estimates, standard errors, and even autocorrelation estimates.

Chapter 5

DISCUSSION

The main purpose of this dissertation is to help other SCD researchers and applied practitioners understand and interpret the ES estimates from the three different SCD effect size indices for count outcomes: LRR, NLB, and BRR, so that they can make better informed decisions about which one to use in their own research study. This is important because, though these three methods are for count outcome data, they use either different statistical modeling or a different estimation framework (Bayesian or frequentist), and they may assume the presence or absence of autocorrelation, which is frequently present in SCD data. Thus, a comparison of these methods in terms of assumptions and interpretations, and on the performance of these ES indices when autocorrelation and overdispersion (common characteristics of SCD count data) are present is warranted.

WWC (2022a) have emphasized on comparability of different SCD effect sizes by transforming them to a common metric. The findings from this dissertation empirically shows it is possible to compare the ES estimates from these three count effect size indices, i.e., they can be transformed to a common metric. If a researcher is planning to use either one of these methods in their study, then this dissertation can be a resource (in addition to the methods paper of these three ES indices) that highlights not only how to use, interpret, and compare the ES estimates but to what degree one should be aware of the potential problems if autocorrelation and overdispersion is

present in the data and which method might be better to use in a given situation. Thus, the results from this dissertation can largely benefit both new and existing SCD researchers and practitioners who use/plan to use SCD in their research and practice.

Based on results from this dissertation, all the three methods provide ES estimates in familiar metric that is interpretable. LRR provides percentage change metric while BRR and NLB provides ratio of counts metric. In terms of understandability, LRR is the easiest method and does not require knowledge of advanced statistics to run the model and get the parameter estimates. While the ES estimates that are obtained from the NLB and BRR are interpretable, a challenge in itself is how comfortable and confident is a researcher in using the complex statistical model and Bayesian estimation techniques used by BRR and NLB. Thus, these two methods are not easily understandable, and the understandability depends on the SCD researcher who is using it.

The effect size estimates from the three methods LRR, NLB, and BRR are rate ratios of average level in the treatment and baseline phases and thus can be converted to a common metric. The analysis results of Schmidt (2007) disruptive behavior converted to a common metric allowed for comparisons among the three ES estimates. Thus, for SCD with count outcomes, the ES estimate obtained from these three methods can be compared as they can be converted to a common metric; specifically, a ratio of phase means. This conversion cannot be done for the standard error estimates as they are in different metrics. However, the coverage rates of the 95% confidence/credible interval from the three methods can be compared because the lower and upper 95% limits are the 2.5^{th} and 97.5^{th} percentiles respectively, and these can be transformed to the ratio of phase means scale.

The performance of the three methods for a typical single-case study with ABAB design and three cases with/out autocorrelation and with/out overdispersion count outcome was examined with simulation. The simulation results from this dissertation clearly indicates that all three ES indices produce reasonably unbiased estimates of the treatment effect under various autocorrelation and/or overdispersion values. The highest raw bias calculated was 0.08 which is less than 0.1 and is quite low. The average log bias is also fairly close to zero. The RMSE results suggests BRR RMSE values are larger than that of LRR and NLB suggesting these two methods provide more accurate estimates compared to BRR. One can be quite confident about producing an ES estimate that is reasonably close to the true effect size of the population under study using any of the three indices. If the main priority of the study is getting an unbiased estimate of the treatment effect, then the preference of using one method over the other will be driven by other considerations since all the three ES indices estimates are nearly equivalent in terms of bias.

In terms of efficiency, NLB is the only ES that has 95% coverage when there is no autocorrelation and negligible overdispersion. However, the presence of autocorrelation and/or overdispersion seems to greatly affect the standard error, thus resulting in decrease of coverage, especially for BRR and NLB. Both BRR and NLB methods use Bayesian estimation and weakly informative priors (standard normal distribution i.e., Normal (0,1) for phase means and vague prior ((gamma (1,1)) for precision and an uninformative prior (uniform (-1,1)) for autocorrelation was used. These priors were chosen by using the priors used in the BRR R codes and NLB paper as reference. Both the methods use same prior for the phase means and precisions, however NLB method has more data. It uses data from three cases and thus uses

shrinkage estimates, and thus “borrows strength” from other cases while estimating the parameter for each case (Raudenbush & Bryk, 2002). BRR, even though it accounts for autocorrelation, it uses information from only one case to produce estimates, and the influence of a weakly informative prior for phase means may cause lower coverage rates than NLB and LRR. However, only with sensitivity analysis this can be further explored. Moreover, the difference in the sample size in A1B1 and A2B2 phases also indicates that both bias and RMSE might be affected by the difference in sample sizes, but again this is yet to be examined.

In case of LRR, for zero autocorrelation, the coverage across the different overdispersion values is close to 90%. Pustejovsky (2018), has indicated that the SE for a single study might not be valid when the outcome measures are autocorrelated and SE estimates generally tend to under-estimate the true sampling variability, though this is not an issue in meta-analysis because robust variance estimation is used. Thus, interpreting in terms of SE might not be appropriate. Interpreting in terms of coverage, the coverage seems to decrease as the autocorrelation increases, though the effect of increasing overdispersion seems pretty non-existent. This might be because LRR does not assume any distribution for the count outcomes while other two models assume the data follow a Poisson distribution (which assumes the mean and variance are equal).

Looking at the average range of the 95% CI, the average range of the 95% CI for BRR and NLB is not very wide though it is generally seen that using the fully Bayesian methods like in NLB and BRR, it takes into account uncertainty about all other parameters, thus the standard error of the estimates may be larger, resulting in wider confidence intervals (Rindskopf, 2014). Compared to these Bayesian estimation

models, LRR has a relatively wider confidence interval range. With zero autocorrelation and negligible dispersion, the average range of LRR is similar to BRR and NLB, but as indicated in the LRR paper, with increase in autocorrelation, the estimate of the standard error of ES might not be valid, and thus even though the coverage rates and the range are comparable, but the standard error of the LRR is driving much wider CI ranges than the BRR and NLB credible intervals.

Thus, based on the findings, if a researcher has multiple cases, and is confident that there is no autocorrelation and negligible overdispersion in the data, then NLB might be the best ES measure (both unbiased and efficient) to use as it allows one to produce both within and across cases effect sizes, and also obtain posterior distributions of the parameters because it uses Bayesian estimation. LRR uses the simplest and most understandable model, and it is easy to produce estimates using the "SingleCaseES" R package. BRR is the only method that accounts for autocorrelation, and if a researcher wants to use a model that accounts for autocorrelation, then this might be the method to use, as it provides estimates for both ES and autocorrelation. However, the BRR model appears to struggle when the data includes both autocorrelation and overdispersion. If the researcher is worried about overdispersion in the data, then LRR appears to be the most robust method. If there is presence of both autocorrelation and overdispersion in the data, then none of these methods seems to be working well though.

5.1 Limitations of this study

There are several limitations of this dissertation. Firstly, it is comparing count ES indices that assume no trend in the data and only counts out of a fixed interval are considered. The performance results of the three ES indices presented in this

dissertation are based on only 1,000 Monte Carlo replications, though 5,000 MC replications were generated. The simulation study used only one set of priors for BRR and NLB methods, sensitivity analysis with other priors may result in different bias and coverages for these methods. For BRR and NLB, even with thinning of MCMC chain, there may still be autocorrelation in the plausible values, which could lead to underestimation of the variance of the posterior distribution of the effect size. There is no clear explanation as to why coverage rates for LRR and BRR were less than 95% even when there is absence of autocorrelation and negligible overdispersion. Difference in phase sample size might be affecting the performance measures results. However, the effect of different sample sizes on these measures are not studied in this dissertation.

5.2 Future Research and Recommendation

The results from this dissertation suggest that characteristics of SCD count data such as autocorrelation and overdispersion affect the standard error estimates for the three count models compared. Thus, the next step is to explore the potential sources of underestimation of parameter variance in these models and if possible, incorporate additional parameters to address autocorrelation and/or overdispersion. For example, negative binomial models can be used to address both count outcomes and overdispersion. Also, there is a need to develop methods for sensitivity analyses to assess robustness of results from one SCD study when autocorrelation and/or overdispersion are present. The effect of different sample sizes on the estimates from the three methods is yet to be examined.

Some improvements can be made to make the ES method and estimates more accessible to the wider researcher community who are new or existing SCD users. For

example, simple phase means could be included in the output from the LRR method. Likewise, the R codes for BRR automatically provide the phase means in raw counts in addition to log counts. However, there exists no GitHub page or R package for estimating the NLB model.

Among numerous reasons, flexibility of SCD and accessibility of SCD datasets have contributed towards increase in the use of SCD in diverse fields. With advancement in the statistical methods available for SCDs, the contribution of SCDs in identifying evidence-based practices can be huge in coming years. To make these statistical methods used in SCD more accessible, the impetus is not only upon the methodologists to develop methods that cater to the nuances of SCDs and present them in understandable and simple manner but also upon the applied practitioners and researchers to make an attempt to learn these new methods and try to implement in their studies and eventually contribute towards evidence-based practices in their respective fields.

REFERENCES

Agresti, A. (2015). Foundations of linear and generalized linear models, John Wiley & Sons, Incorporated. ProQuest Ebook Central.https://ebookcentral.proquest.com/lib/udel-ebooks/detail.action?docID=1895564.

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika, 65,* 475–497. http://dx.doi.org/10.1007/BF02296339.

Barger-Anderson, R., Domaracki, J. W., Kearney-Vakulick, N., & Kubina, R. M. (2004). Multiple baseline designs: The use of single-case experimental design in literacy research. *Reading Improvement, 41*, 217-225.

Biglan, A., Ary, D., & Wagenaar, A. C. (2000). The value of interrupted time-series experiments for community intervention research. *Prevention Science, 1*(1), 31–49. doi: 10.1023/a:1010024016308

Bolker, B. M. (2008). *Ecological Models and Data in R.* Princeton University Press. https://doi.org/10.2307/j.ctvcm4g37

Brooks, S., & Gelman, A. (1998). Some issues in monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7,* 434-455. Doi: 10.1080/10618600.1998.10474787

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior modification, 30(5),* 531–563. https://doi.org/10.1177/0145445503261167

Bürkner P (2021). "Bayesian Item Response Modeling in R with brms and Stan." *Journal of Statistical Software*, **100**(5), 1–54. doi:10.18637/jss.v100.i05.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), Single case research design and analysis (pp. 187–212). Mahwah, NJ: Erlbaum.

Byiers B. J., Reichle J., & Symons F. J. (2012). Single-subject experimental design for evidence-based practice. *American Journal of Speech-Language Pathology, 21*, 397–414. doi: 10.1044/1058-0360(2012/11-0036)

Cannon, J. E., Guardino, C., Antia, S. D., & Luckner, J. L. (2016). Single-case design research: Building the evidence base in the field of education of deaf and hard of hearing students. *American Annals of the Deaf*, 160(5), 440–452. doi: 10.1353/aad.2016.0007

Carsey, T. M., & Harden, J. J. (2014). Monte Carlo simulation and resampling methods for social science. Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781483319605

Center, B. A., Skiba, R. J., & Casey, A. (1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education,19 (4),* 387–400.

Chen, M., & Pustejovsky, J. E. (2022). Multilevel meta-analysis of single-case experimental designs using robust variance estimation. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000510

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. W. (2014). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education, 39,* 305–318.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of personality assessment*, 91(2), 121-136.

Davis, R. A., Dunsmuir, W. T., & Wang, Y. (1999). Modeling time series of count data. *Statistics textbooks and monographs*, *158*, 63-114.

Declercq, L., Jamshidi, L., Fernandez-Castilla, B., Berevtas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (2019). Analysis of single-case experimental count data using the linear mixed effects model: A simulation study. *Behavior Research Methods, 51*, 2477-2497. https://doi.org/10.3758/s13428-018-1091-y

Denwood, M. J. (2016). runjags: An R package providing interface utilities, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software, 71*. Doi:10.18637/jss.v071.i09

Dollaghan, C. A. (2007). Handbook for evidence-based practice in communication disorders. Baltimore, MD: Brookes.

Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study of multilevel-modeling approaches. *Behavior Research Methods, 42,* 930–943. Doi:10.3758/BRM.42.4.930.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., … Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6,* 151–175. doi: 10.1007/s11121-005-5553-y

Gabler, N. B., Duan, N., Vohra, S., & Kravitz, R. L. (2011). N-of-1 trials in the medical literature: A systematic review. *Medical Care, 49,* 761– 768. doi:10.1097/MLR.0b013e318215d90d

Gage, N.A., & Lewis, T. J. (2013). Analysis of effect for single-case design research. *Journal of Applied Sport Psychology, 25,* 46–60. DOI:10.1080/10413200.2012.660673

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6* ,721–741. doi:10.1109/TPAMI.1984.4767596.

Gierut, J. A., Morrisette, M. L., & Dickinson, S. L. (2015). Effect size for single-subject design in phonological treatment. *Journal of speech, language, and hearing research, 58(5),* 1464–1481. https://doi.org/10.1044/2015_JSLHR-S-14-0299

Glass, G.V. (1978). Integrating findings: The meta-analysis of research. *Review of Educational Research, 5,* 351–371.

Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research, 50,* 162–183. https://doi.org/10.1080/00273171.2014.973989

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives, 2(3),* 167–171. https://doi.org/10.1111/j.1750-8606.2008.00060.x.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3,* 224–239. http://dx.doi.org/10.1002/jrsm.1052

Hersen, M., & Barlow, D. H. (1976). Single-case experimental designs strategies for studying behavior change. Elmsford, NY: Pergamon Press Ltd.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165-179. https://doi.org/10.1177/001440290507100203

Kazdin, A. E. (1982). Single-case experimental designs in clinical research and practice. *New Directions for Methodology of Social & Behavioral Science, 13,* 33–47

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

Kame'enui, E. (2006). The national special education research agenda: Inside the matrix. Paper presented at the Council for Exceptional Children convention and expo, Salt Lake City, UT

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue, *Psychological Methods, 15*, 122-144. doi: 10.1037/a0017736

Kratochwill, T. R., & Levin, J. R. (2014). Meta- and statistical analysis of single-case intervention research data: Quantitative gifts and a wish list. *Journal of School Psychology, 52,* 231–235. Doi:10.1016/j.jsp.2014.01.003

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

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34,* 26–38. Doi:10.1177/0741932512452794

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*, 573–603. http://dx.doi.org/10.1037/a0029146

Kruschke, J. K. (2015). Doing Bayesian data analysis: A Tutorial with R, JAGS, and Stan (2nd ed.). Burlington, MA: Academic Press/Elsevier.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychon Bull Rev*, *25*, 155–177. https://doi.org/10.3758/s13423-017-1272-1

Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y. Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions*, *8*(2), 88-99.

Law, J., Garrett, Z., & Nye, C. (2004). The efficacy of treatment for children with developmental speech and language delay/disorder: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *47*, 924–943.

Ledford, J.R., Lane, J.D., & Severini, K.E. (2017). Systematic Use of Visual Analysis for Assessing Outcomes in Single Case Design Studies. *Brain Impairment*, *19*, 4 - 17.

Lobo, M. A., Moeyaert, M., Cunha, A. B., & Babik, I. (2017). Single-case design, analysis, and quality assessment for intervention research. Journal of Neurologic Physical Therapy, 41, 187–197. Doi:10.1097/NPT.0000000000000187

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, *10*(4), 325-337.

May, H. (2004). Making statistics more meaningful for policy research and program evaluation. *American Journal of evaluation*, *25*(4), 525-540.

Moeyaert, M., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017). Multilevel modeling of single-case data: A comparison of maximum likelihood and Bayesian estimation. *Psychological Methods*, *22*, 760–778. https://doi.org/10.1037/met0000136

Montgomery, D.C., Peck, E.A., and Vining, G.G. (2006) Introduction to Linear Regression Analysis. 4th Edition, John Wiley & Sons, Inc., Hoboken.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074-2102

Muller, K. E., Edwards, L. J., Simpson, S. L., & Taylor, D. J. (2007). Statistical tests with accurate size and power for balanced linear mixed models. *Statistics in Medicine*, *26*, 3639–3660. http://dx.doi.org/10.1002/sim.2827

Natesan, P. (2019). Fitting Bayesian models for single-case experimental designs: A tutorial. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *15*(4), 147-156. doi: 10.1027/1614-2241/a000180

Natesan, P. & Hedges, L. V. (2017). Bayesian unknown change-point models to investigate immediacy in single case designs. *Psychological Methods*, *22*, 743-759. doi: 10.1037/met0000134

Natesan Batley, P., Nandakumar, R., Palka, J. M., & Shrestha, P. (2020). Comparing the Bayesian unknown change-point model and simulation modeling analysis to analyze single case experimental designs. *Frontiers in Psychology*, 11, 3960. doi: 10.3389/fpsyg.2020.617047

Natesan Batley, P., Shukla Mehta, S., & Hitchcock, J. H. (2021). A Bayesian rate ratio effect size to quantify intervention effect for count data in single case experimental research. *Behavioral Disorders*. doi: 10.1177/0198742920930704

National Reading Panel. (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 (00-4769). Washington, DC: National Institute of Child Health & Human Development.

(NCER 2015-002). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology : An International Journal of Experimental Educational Psychology.*, *25* (2–3), 313–324. doi: 10.1080/0144341042000301238.

Parker, R.I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine overlap techniques. *Behavior Modification*, *35*(4), 303-22. doi: 10.1177/0145445511399147

Payne, E.H., Gebregziabher ,M., Hardin, J.W., Ramakrishnan, V., & Egede, L,E (2018). An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data. *Commun Stat Simul Comput*. Jul 5;47(6):1722-1738. doi: 10.1080/03610918.2017.1323223. PMID: 30555205; PMCID: PMC6290908.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Retrieved from https://www.r-project.org/conferences/DSC-2003/Drafts/Plummer.pdf

Plummer, M. (2012). JAGS Version 3.3. 0 user manual.

Prathiba-stat (2020). Bayesian-rate-ratio. https://github.com/prathiba-stat/Bayesian-rate-ratio

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-Comparable Effect Sizes in Multiple Baseline Designs: A General Modeling Framework. *Journal of Educational and Behavioral Statistics*, *39*(5), 368–393. https://doi.org/10.3102/1076998614547577

Pustejovsky, J. E. (2015). Measurement-comparable effect sizes for single-case studies of free-operant behavior. *Psychological Methods*, *20*(3), 342.

Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, *68*, 99-112. https://doi.org/10.1016/j.jsp.2018.02.003

Pustejovsky, J. E., Chen, M., & Swan, D. M., (2021). SingleCaseES : A calculator for Single-Case Effect sizes. R package version 0.5.0. https://cran.r-project.org/web/packages/SingleCaseES/index.html

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.

Rindskopf, D. (2014). Nonlinear Bayesian analysis for single case designs. *Journal of School Psychology, 52*, 179-189. http://dx.doi.org/10.1016/j.jsp.2013.12.003

Schmidt, A. C. (2007). The effects of a group contingency on group and individual behavior in an urban first-grade classroom. University of Kansas. Retrieved from http://gradworks.umi.com/14/43/1443719.html.

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*, 109-122. https://doi.org/10.1016/j.jsp.2013.11.009

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: new applications and some agenda items for future research. *Psychological methods*, *18*(3), 385.

Shadish, W., & Rindskopf, D. (2007). Methods for evidence-based practice: Quantitative Synthesis of Single-Subject Designs. *New Directions for Evaluation, 2007,* 95–109.

Shadish, W. R., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2013). Bayesian estimates of autocorrelations in single-case designs. *Behavioral Research Methods, 45*, 813-821. doi: 10.3758/s13428-012-0282-1

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971-980. doi: 10.3758/s13428-011-0111-y

Sigurdsson, S. O., & Austin, J. (2006). Should we be measuring effect size in applied behavior analysis? Retrieved March 29, 2006 from the Organizational Behavior Management Network, http://www.obmnetwork.com/resources/articles/main/Sigurdsson_EffectSize.htm

Slavin, R. E. (1984). Meta-analysis in education: How it has been used? *Educational Researcher, 13,* 6–15

Smith, J.D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods. 17*(4), 510-550. doi: 10.1037/a0029312

Swan, D., Pustejovsky, J. E., & Beretvas, N. (2020). The impact of response-guided designs on count outcomes in single-case experimental design baselines. *Evidence-Based Communication Assessment and Intervention, 14,* 82–107. https://doi.org/10.1080/17489539.2020.1739048

.

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology, 52*, 213-230. http://dx.doi.org/10.1016/j.jsp.2013.12.002

Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. 2008. Rating the methodological quality of single-subject designs and N-of-1 trials: Introducing the Single-Case Experimental Design (SCED) Scale. *Neuropsychological Rehabilitation, 18(4),* 385–401. Doi:10.1080/09602010802009201

Tate, R. L., Rosenkoetter, U., Vohra, S., Horner, R., Kratochwill, T., Sampson, M., & Wilson, B. (2016). The single-case reporting guideline in behavioral interventions (SCRIBE) 2016 statement. *Journal of Clinical Epidemiology, 73,* 142–152. Doi:10.1016/j.jclinepi.2016.04.006

Therrien, W. J., Zaman, M., & Banda, D. R. (2011). How can meta-analyses guide practice? A review of the learning disability research base. *Remedial and Special Education, 32(3),* 206–218. https://doi.org/10.1177/0741932510361266

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental studies using hierarchical linear models. *School Psychology Quarterly, 18,* 325-346. http://dx.doi.org/10.1521/scpq.18.3.325.22577

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, and Computers, 35,* 1-10. http://dx.doi.org/10.3758/BF03195492

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today, 8,* 196–209. http://dx.doi.org/10.1037/h0100613

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2,* 142–151. https://doi.org/10.1080/17489530802505362

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic bulletin & review, 25*(1), 143-154.

What Works Clearinghouse. (2022a). *What Works Clearinghouse procedures and standards handbook, version 5.0.* U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). This report is available on the What Works Clearinghouse website at https://ies.ed.gov/ncee/wwc/Handbooks.

What Works Clearinghouse. (2022b). Reviews of Individual Studies. https://ies.ed.gov/ncee/wwc/reviewedstudies#/RatingId:99%7CStudyDesignId:4%7COnlyStudiesWithPositiveEffects:false%7CSetNumber:1%7CEssaRatingId:0,1,2,3

Wilson, B. A. (2011). 'Cutting Edge' developments in neuropsychological

rehabilitation and possible future directions. *Brain Impairment, 12(1),* 33–42. Doi:10.1375/brim.12.1.33

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education, 44(1),* 18–28. https://doi.org/10.1177/0022466908328009

Appendix A

GIBBS SAMPLER EXAMPLE

Gibbs sampler follows an iterative procedure to repeatedly sample from the conditional distribution of one variable given all of the other variables.

Suppose, a parameter vector $\Theta = (\mathrm{A}, \mathrm{B}, \mathrm{C})$ is to be estimated. Using the Gibbs sampler the posterior distribution of the parameter vector given the observed data, $\pi(\Theta|Y)$ can be obtained. A set of starting values is assigned to the parameter vector at step 0. Let the iterations be indexed using the variable i.

Step 1: Set i =i+1.

Step 2: Draw $(A^i \mid B^{i-1}, C^{i-1}, Y)$

Step 3: Draw $(B^i \mid A^i, C^{i-1}, Y)$

Step 4: Draw $(C^i \mid A^i, B^i, Y)$

Step 5: Draw $(\hat{Y}^i \big| A^i, B^i, C^i, Y)$, where $\hat{Y}^i$ is the predicted Y values at the ith iteration.

Step 6: Return to step 1

The above shown steps are for one iteration of the sampler. The iterations are continued until convergence.

Appendix B

NONLINEAR BAYESIAN WINBUGS CODES FOR PROPORTION DATA

```
model {
for (i in 1: z) {                                                    # z total observations
        logit (theta[i]) <- b0[subj[i]] + b1[subj[i]] * phase[i]      #level-1 model
        y[i] ~ dbin (theta[i], n)                                     #count out of a n number of  trials
                                                                        modeled binomially
}
for (j in 1: m) {                                  # m cases
        b0[j] ~ dnorm (mu0 , prec0)                #intercept treated as a random effect
        b1[j] ~ dnorm (mu1, prec1)                 #treatment treated as a random effect
        case.0[j] <- b0[j]                         # log-odds of the outcome expected during
                                                     baseline for each case

        case.1[j] <- case.0[j] + b1[j]             # log-odds of the outcome expected during
                                                     treatment for each case

        case.pA.0[j] <- 1/(1+exp(-1*case.0[j]))
        # log-odds of the outcome expected during baseline transformed to proportion
        # for each case

        case.pB.1[j] <- 1/(1+exp(-1*case.1[j]))
          # log-odds of the outcome expected during treatment transformed
          # to proportion for each case
        }
mu0 ~ dnorm (0, 0.01)                  # prior on baseline mean
mu1 ~ dnorm (0, 0.01)                  # prior on the treatment effect
prec0 ~ dgamma (0.01, 0.01)            # prior on baseline precision
prec1 ~ dgamma (0.01, 0.01)            # prior on treatment precision
sig0 <- 1/sqrt (prec0)                 # standard deviation of intercept
sig1 <- 1/sqrt (prec1)                 # standard deviation of phase effect
var0 <- 1/prec0                        # variance in baseline
var1 <- 1/prec1                        # variance in treatment

phaseB <- mu0 + mu1                    # treatment mean
p.A <- exp (mu0)/(1+exp ( mu0))        # probability of outcome in the baseline
p.B <- exp (phaseB)/(1+exp(phaseB))    # probability of outcome in the treatment
p.AB <- p.A – p.B                      # The average treatment effect
}
```

Appendix C

NONLINEAR BAYESIAN WINBUGS CODES FOR COUNT DATA

```
model {
for (i in 1: z) {                                          # z total observations
        log(lambda[i]) <- b0[subj[i]]+b1[subj[i]]*phase[i]   #level 1 model
        y[i]~ dpois(lambda[i])                             # count outcome modeled with
                                                           Poisson distribution parameter
                                                           lambda
        }
for (j in 1: m) {                                  # m cases
        b0[j] ~ dnorm (mu0 , prec0)                #intercept treated as a random effect
        b1[j] ~ dnorm (mu1, prec1)                 #treatment treated as a random effect
        case.0[j] <- b0[j]
        # log count of the outcome expected during baseline for each case

        case.1[j] <- case.0[j] + b1[j]
        # log count of the outcome expected during treatment for each case

        case.pA.0[j] <- exp (1 * case.0[j])
         # log count of the outcome expected during baseline transformed to count for each
case

        case.pB.1[j] <- exp (1 * case.1[j])
        # log count of the outcome expected during treatment transformed to count for each
case
}
mu0 ~ dnorm (0, 0.01)                  # prior on baseline mean
mu1 ~ dnorm (0, 0.01)                  # prior on the treatment effect
prec0 ~ dgamma (0.01, 0.01)            # prior on baseline precision
prec1 ~ dgamma (0.01, 0.01)            # prior on treatment precision
sig0 <- 1/sqrt (prec0)                 # standard deviation of intercept
sig1 <- 1/sqrt (prec1)                 # standard deviation of phase effect
var0 <- 1/prec0                        # variance in baseline
var1 <- 1/prec1                        # variance in treatment effect

phaseB <- mu0 + mu1                    # treatment mean
p.A <- exp (mu0)                       # average count of the outcome in the baseline
p.B <- exp (phaseB)                    # average count of the outcome in the treatment
p.AB <- p.A – p.B                      # The average treatment effect
}
```

Appendix D

HISTOGRAMS OF GENERATED DATA FOR CASE1 PHASE1