A Bayesian Rate Ratio Effect Size to Quantify Intervention Effects for Count Data in Single Case Experimental Research

Abstract

Single case experimental design (SCED) is an indispensable methodology when evaluating intervention efficacy. Despite longstanding success with using visual analyses to evaluate SCED data, this method has limited utility for conducting meta-analyses. This is critical because meta-analyses should drive practice and policy in behavioral disorders, more than evidence derived from individual SCEDs. Even when analyzing data from individual studies, there is merit to using multiple analytic methods since statistical analyses in SCED can be challenging given small sample sizes and autocorrelated data. These complexities are exacerbated when using count data, which are common in SCEDs. Bayesian methods can be used to develop new statistical procedures that may address these challenges. The purpose of the present study was to formulate a within-subject Bayesian rate ratio effect size (BRR) for autocorrelated count data which obviates the need for small sample corrections. This effect size is the first step towards building a between-subject rate ratio that can be used for meta-analyses. We illustrate this within-subject effect size using real data for an ABAB design and provide codes for practitioners who may want to compute BRR.

Keywords: single case experimental design; visual analysis; Bayesian; rate ratio; effect size; interrupted time-series

A Bayesian Rate Ratio Effect Size to Quantify Intervention Effects for Count Data in Single Case Experimental Research

The generation of evidence-based practices (EBPs) for students for whom typical instruction may not be effective must rely on research that meets strong methodological standards (e.g., Odom, Brantlinger, Gersten, Horner, Thompson, & Harris, 2005). One class of methods that can meet strong standards when evaluating intervention efficacy is the *single-case experimental design* (SCED). SCEDs can yield solid causal inference about treatment impacts and are of interest to federal agencies such as the Institute of Education Sciences (IES) (e.g., Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, & Shadish, 2010, 2013) and the National Institutes of Health (NIH) for n-of-1 designs which are a special case of SCEDs (Gabler, Duan, Vohra, & Kravitz, 2011). The What Works Clearinghouse (WWC), for example, now offers reports on special education interventions that are largely informed by SCED work (e.g., WWC, 2016) and IES funds SCED research to advance development of EBPs. The development of EBPs can be further advanced by systematically synthesizing SCED evidence, which represents a critical facet of behavioral disorders literature (cf. Briesch & Briesch, 2016; Chaffee, Briesch, Johnson, & Volpe, 2017; Dart, Collins, Klingbeil, McKinley, & VanDerHeyden, 2014; Kilgus, Riley-Tillman, & Kratochwill, 2016; Maggin, O'Keefe, & Johnson, 2011; Maggin, Chafouleas, Goddard, & Johnson, 2011; Soares, Harrison, Vannest, & McClelland , 2016). In principle, syntheses could be expanded by combining SCED effect sizes with impact estimates generated from other design types, such as randomized controlled trials and quasi-experiments (e.g., Hitchcock, Horner, Kratochwill, Levin, Odom, Rindskopf, & Shadish, 2014). Hence, EBP generation and evaluation is inextricably linked to SCED work and, as demonstrated later in this article, there is an ongoing need for methodological refinement to

related statistical analyses. We argue that expansion of analytic options could in turn, support behavioral analysis praxis, which entails combining research and practice (see for e.g., Nastasi & Hitchcock, 2016) and, more distally, practice via corresponding improvements in our understanding of evidence.

So what is the basis for arguing that there is need to refine statistical analyses of SCED data? In this article we focus on effect size estimation to address this question. Of course, several SCED effect size procedures exist, such as those based on percent of non-overlap data points between the baseline and intervention phases (Parker, Vannest & Davis, 2011) and standardized between-subject mean difference corrected for small sample sizes (Hedges, Pustejovsky, & Shadish, 2012, 2013). However, the former set of indices is problematic because they do not well account for outliers (Harrington & Velicer, 2015), cannot account for trend, and only measure non-overlap but not an actual effect size. The latter set of indices are a significant innovation in SCED analyses because they are between-subject effect sizes that can correct for small sample sizes assuming that data are intervally-scaled. However, it is more common in SCEDs to use count (e.g., the number of times some discrete behavior occurred) or proportion data (e.g., the percentage of time a student has appeared to be attentive in a classroom) (Rindskopf, 2014). Moreover, SCED data are often autocorrelated which means the error at a given time-point (say t) is systematically correlated with the error at a different time-point (say $t + l$). This is referred to as a l-lag autocorrelation (e.g. 1-lag, 2-lag, etc). This autocorrelation is the antithesis of the independence of observation assumption that is the basic tenet of all general linear models such as ANOVA and regression. Unfortunately, other commonly used effect sizes in behavioral research such as R-squared (Cohen, 1988) do not account for autocorrelations. Therefore, there is a need for new effect size procedures that do not necessarily replace existing approaches but can

at least be used in a supplementary fashion with existing procedures such as visual analyses so that SCED researchers can draw yet more information from their studies.

To be considered as a contribution to research and later practice, any new effect size estimation procedure should: (a) account for both autocorrelations and the scale of the data commonly used in SCEDs; (b) deal with small sample sizes, and (c) produce reliable interval estimates of uncertainty. To our knowledge the effect size we propose here, the Bayesian rate ratio (BRR) effect size, is the first to meet these needs. To demonstrate the BRR, in this article we use data from a published study of an ABAB design that was used to reduce disruptive behaviors of students in an urban fourth grade Math classroom (Lambert, Cartledge, Heward, & Lo, 2006). We apply the BRR to show how Bayesian statistical significance testing can be conducted using SCED count data, and we assess the degree to which visual analyses, the nonoverlap of all pairs (NAP) effect size, and the BRR produce both complementary and contradictory information about the intervention effect. Before these demonstrations are presented, we offer an overview of how Bayesian estimation can contribute to the analyses of SCED data. This is because understanding the potential contribution of the BRR to SCED work first requires a review of the challenges that come with analyzing SCED data.

Challenges in analyzing SCED data

One of the main reasons for use of SCEDs is the need to document a functional relation between specified independent and dependent variables. Essentially each person (or case) serves as the unit of analysis and his/her own control to generate strong causal evidence about intervention effects. Evidence of intervention efficacy is documented primarily through visual analyses that focus on changes from baseline to intervention in the level, trend, variability, immediacy of effect, data overlap, and consistency in the behavioral pattern for similar phases

(Gast & Ledford, 2014; Horner & Kratochwill, 2012). There is some, but not complete consensus among expert researchers on the decision-rules for making judgments regarding intervention effectiveness (e.g., Kratochwill et al., 2013), but the rules for visual analyses are not applied uniformly by behavioral and educational researchers (Horner, Swaminathan, Sugai, & Smolkowski, 2012). Furthermore, not all treatments exhibit immediacy effect and some treatment effects may not be visually striking, even though the overall data may show clinical and statistical effectiveness (Meadan, Snodgrass, Meyer, Fisher, Chung, & Halle, 2016). Therefore, although some researchers believe that visual analysis can be based on objective criteria (Horner, Carr, Halle, McGee, Odom, & Wolery, 2005; Roane, Rihgdahl, Kelley, & Glover, 2011), others see a need for quantitative methods to document intervention effects (e.g., Maggin, Chafouleas, Goddard, & Johnson, 2011; Parker et al., 2011).

We argue that, in principle, there is a need for both statistical and visual analysis to evaluate the causal validity of SCED findings via transparent, objective, and replicable procedures. Visual analysis primarily addresses the question of evidence of a functional relation between independent and dependent variables and statistical analysis quantifies the magnitude of the effect. We agree with authors who see visual analysis as an effective analytic approach but more information can be drawn from using multiple analytic methods and visual analyses do come with drawbacks.

Autocorrelation. A primary drawback from using visual analysis alone is based on the problem of autocorrelated errors (Harrington & Velicer, 2015), which is typical of SCED data given the need for repeated measures. Autocorrelation can contribute to decreased interrater reliability during visual analyses (Brossart, Parker, Olson, & Mahadevan, 2006) and increase in

Type I errors (Horner & Kratochwill, 2012; Lenovaz & Rapp, 2015; Maggin & Chafouleas, 2013).

If visual analyses were imperfect with respect to distinguishing autocorrelation from true performance change, one would hope to use statistical analyses to offer complimentary procedures so that researchers are better able to understand treatment effects. However, regularly used statistical methods of analyses such as ANOVA and regression are poorly suited for most SCED studies. To begin, ANOVA/regression-based (i.e., Ordinary Least Squares [OLS]) methods a) entail assuming that observations are independent (i.e., the antithesis of autocorrelation) and b) related analyses should be expected to contend with higher rates of Type II errors because SCEDs typically entail use of small sample sizes (Gresham, Sugai, & Horner, 2001).

On the other hand, OLS procedures can be used to detect the presence of autocorrelation. However, confidence intervals (CI) of autocorrelation estimates, which are needed to help us understand whether we can rule out autocorrelation, tend to be inaccurate because they tend to have undercoverage. Undercoverage means that CIs are narrower than they should be (Shadish, Rindskopf, Hedges, & Sullivan, 2013) and thereby makes it difficult to assess if autocorrelation is a concern. A subtler issue is that the challenges of inadequate autocorrelation diagnostics and small sample sizes interact. Autocorrelation estimates are often negatively biased and are accompanied by larger sampling errors because SCEDs typically have a small number of observations per participant in a study. Huitema and McKean (1994) and McKnight, McKean, and Huitema (2000) state that 50 observations per participant are about the minimum threshold needed to address these sampling error concerns. In contrast, a review of 809 SCEDs published in 113 studies in the year 2008 in 21 journals, studies typically had only 4-6 observations per

phase (Shadish & Sullivan, 2011). This is important because OLS confidence intervals have undercoverage, meaning fewer than expected autocorrelation confidence intervals contain the true value (Shadish, Rindskopf, Hedges, & Sullivan, 2013). This concern is exacerbated when there are a minimal number of data points per phase. In short, OLS procedures for assessing the presence of autocorrelation in SCED data may lead analysts to proceed with false confidence.

Effect sizes. Concerns with the use of standard statistical approaches move beyond autocorrelation, Type I, and Type II errors. Effect size estimates are also problematic because of the reasons discussed below. Standardized mean difference type effect sizes obtained from SCEDs require correction for small samples and require distributional assumptions that might not fit with typical analytic scenarios (Hedges, Pustejovsky, & Shadish, 2012, 2013). Of course, non-overlap indices represent a good option because they are free of distributional assumptions and can be applied to count data (Parker et al., 2011); there is reason after all for their longstanding use. However, NAP indices do not help researchers account for the distance between data points and consider only their non-overlap. This renders non-overlap between two closely spaced points the same as non-overlap between two widely spaced points. By logic, however, we expect the effect size of the former case should be greater than the effect size of the latter. Moreover, the standard errors proposed for NAP are not free of distributional assumptions and may be biased in the presence of autocorrelation. Due to space restrictions, we do not review all non-overlap SCED effect sizes options (see instead Parker et al., 2011). Furthermore, computing *p*-values and CIs for non-overlap metrics entail complex procedures (Parker et al., 2011). For all of these reasons, there is a need for different quantitative analytic solutions (e.g., Shadish et al., 2013). We argue that Bayesian methods can yield a viable solution that can overcome these challenges.

Bayesian Methods and SCEDs

A fundamental reason for why a Bayesian approach can be of use in the examination of SCED data is that it does not depend on large sample or asymptotic theory (Ansari & Jedidi, 2000; Ansari, Jedidi, & Jagpal, 2000). Bayesian methods also allow more direct probabilistic interpretation of parameters than do *frequentist* methods, which generally use OLS and the sort of null hypothesis testing procedures that are based on Fisher's work (Cohen, 1994). Bayesian estimation entails examining the posterior distributions of parameters such as intercepts, slopes, and effect sizes, and provides the probability (or credibility value) of each value an estimated parameter can take (Kruschke, 2013). Unfortunately, most applied researchers are not trained in Bayesian estimation (Natesan, Boedeker, & Onwuegbuzie, 2018). Perhaps as a result, Bayesian methods are not typically used in SCED work. Consider, for example, that of the 239 SCED articles published in the first half of 2018, only four mention the word Bayesian; of these four, none were empirical works (Natesan, 2019).

Fortunately, statistical methodologists have started to work out how Bayesian methods can be deployed to overcome various analytical challenges presented by SCED data. For instance, Moeyaert, Rindskopf, Onghena, & Van den Noortgate (2017) compared maximum likelihood and Bayesian estimation of multilevel modeling of SCED data. Natesan and Hedges (2017) proposed a Bayesian unknown change-point model that overcomes the small data and autocorrelation challenges of SCEDs by using Bayesian methodology. Natesan, Minka, & Hedges (In Press) extended this work further to include multiple phases such as the ABAB design. These works do not however address count data because such data require making different distributional assumptions, and as mentioned above, count data are more common in SCED work than interval data. Therefore, the present article describes how to use within-subject

Bayesian effect sizes and in particular addresses statistical complexities that arise from using count data (the Bayesian rate ratio or the BRR). It is of import to note that the BRR we present is a within-subject effect size and cannot be directly used in meta-analyses unlike the one proposed by Hedges et al. (2012). Nonetheless, we see this proposed BRR as the first step to build an equivalent between-subject effect size for count data that can be used for meta-analyses. Importantly, the programs used to compute the indices for ABAB designs are available to download for free from github (https://github.com/prathiba-stat/Bayesian-rate-ratio) along with annotations so that researchers can modify and input data for their own research. By demonstrating this method, discussing its advantages, and making the software codes accessible, this article can help researchers compute the BRR. In addition, since SCED researchers commonly use visual analyses, we show how to visually examine posterior density plots and regions of practical equivalence (ROPE), which we consider to be a part of the BRR process. With that background, there are three fundamental reasons for why a Bayesian approach should be considered when analyzing SCED data. We present these issues and then describe BRR.

Use of Bayesian Estimation: An Overview of Three Fundamental Issues

In Bayesian methods, each parameter estimate (an outcome that is calculated) represents a distribution of values; in contrast, when using frequentist methods one calculates a point estimate and applies a null hypothesis test. A Bayesian parameter can be of greater utility than a null hypothesis significance testing, and associated CI, derived from frequentist statistics. This is because a) a frequentist CI is often misunderstood in practice, and b) by itself does not support replication research, representing two fundamental issues that warrant use of Bayesian estimation. As for the first issue, unless the interpreter of frequentist results is well acclimated to how the process works, it can be easy to misconstrue a finding. To explain, when a frequentist

95% CI (or 68%, etc.) is constructed around a point estimate, such as a standardized mean difference effect size (SMD), this does not mean that there is a 95% chance that the observed difference is a true representation of a population difference (Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2013). Yet according to Cohen (1994) this is the incorrect interpretation many will make.

To explain, consider 10,000 samples. If one were to obtain CIs from these 10,000 samples then a 95% CI means that 95% of these CIs would contain the true value. In sum, constructing the frequentist CI entails using normal curve theory to provide a sense of how many of some number of (theoretical) sample draws contain the null value, and thereby gives a researcher a basis on which to consider whether to reject a null hypothesis. With that background, the frequentist CI is often misinterpreted as representing the probability that the point estimate (in this example, the observed SMD) is the population parameter, or close to the population parameter (again, see Cohen, 1994). But to be clear, one might be highly confident in rejecting a null hypothesis using frequentist methods but still have limited capacity in guessing the actual value of population parameter. In contrast, in Bayesian estimation the probability that a statistical estimate falls in the 95% credibility interval is much more straightforward. The chances that the observed SMD reflects the actual population value is, well, to be interpreted as 95% (Kruschke, 2015).

This connects to the second fundamental issue. This form of interpretation supports replication research, which has become an important topic in special education (e.g., Cook, 2014). This is because researchers should be using prior information to hypothesize (and empirically test) the size of a plausible treatment impacts on some outcome measure. As more information is gathered, the hypothesis becomes more refined. Hence, having a Bayesian

mindset entails continued thought about replication. If researchers are working with distributions of plausible treatment values, they will be in a stronger position to specify the strength of an intervention in advance of a study.

A third big picture issue is that Bayesian methods can more easily accommodate model complexities and several data types in a way that addresses the numerous concerns described earlier in this article. These methods can handle proportion and count data, which are common in SCED work (e.g., Rindskopf, 2014), and Shadish et al. (2013) found that Bayesian estimates of autocorrelation were more accurate than frequentist estimates. In all, we are not advocating that frequentist methods be discontinued in SCED research but we do argue that they are often poorly suited to SCED data analysis so should be used more sparingly. In contrast, Bayesian approaches do not entail the same drawbacks and they can complement visual analyses. With that background, we turn to BRR details.

Applying the BRR to SCED Count Data

The Bayesian model used to analyze SCEDs in the present study is based on an interrupted time-series design that entails Bayesian estimation (Natesan, 2019; Natesan & Hedges, 2017, 2019; Natesan, Minka, & Hedges, 2019). As the name implies, an interrupted time-series design is a longitudinal design with time as the independent variable and has an outcome variable of interest tracked across time. A sudden introduction or withdrawal of a stimulus at a certain time-point causes an interruption in the pattern obtained until this time-point. Following this interruption, the outcome variable may follow a different pattern. This is the typical setup of an interrupted time-series design. Thus, SCEDs are variants of these designs. In fact, in the ABAB design that we will illustrate, there are three interruptions – baseline to intervention, removal of baseline from intervention, and reintroduction of intervention. We refer

to the Bayesian estimation of an interrupted time-series design as a Bayesian interrupted time-series (BITS) design. In our conceptualization of BITS, intercepts vary by phase (as in AB phases used in most SCEDs). We assume use of count data and in this approach the dependent variable is modelled using Poisson regression. We do not assume trend in the data that may appear due to anything other than autocorrelation because research has shown that BITS CIs of SCED data with autocorrelation and trend due to source other than autocorrelation severely underperform (Natesan & Hedges, 2019). In fact, the model confounds the patterns due to two sources of trend, that is, autocorrelation and trend from other sources such as a growth or decline in the outcome variable, that it is impossible to separate the variance that can be attributed to trend and the variance that can be attributed to autocorrelation. Natesan and Hedges (2019) recommend that for SCEDs, models that estimate only autocorrelation or trend due to other sources be estimated and not both in the same model. Therefore, only the simplest model, that is, the model with intercepts and autocorrelations alone is considered in the present study. The observed value at the first time point (y_{p1}) in Phase p follows a Poisson distribution with mean $\hat{y}_{p1}$ where $\hat{y}_{p1}$is the probability of obtaining a given response on the given model. The rest of the time series follows a Poisson procedure with 1-lag autocorrelated errors (e.g. Harrop & Velicer, 1985; Velicer & Molenaar, 2013). The predicted values in the rest of the time series are distributed as:

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

In Equation 1, H_{pt-1} is the past history, Θ is the vector of parameters, and Po refers to Poisson distribution. Essentially, what Equation 1 demonstrates is that the predicted value of the dependent variable at Time t in Phase p is Poisson-distributed as the probability of the predicted

value of the current data point given the past history, or the value of the previous data point. The generalized linear model and the serial dependency of the residual (e_t) can be expressed as,

$$\hat{y}_{pt} = \begin{cases} \exp(\beta_{01} + e_{pt-1}), if\ t \leq t_b \\ \exp(\beta_{02} + e_{pt-1}), otherwise \end{cases} \text{ and} \quad (2)$$

$$e_{pt-1} = \rho e_{pt-2} + \varepsilon. \quad (3)$$

In Equation 2, $\hat{y}_{pt}$ is the probability of the predicted value of the dependent variable at Time t in phase p; β_{01} and β_{02} are the means or intercepts of Phase 1 and Phase 2, respectively; e_{pt} is the error at Time t in Phase p; ρ is the autocorrelation coefficient; and ε is the independently distributed error. In Equation 3, e is white noise created by a combination of random error (ε) and autocorrelation between adjacent time-points (ρ). Their standard deviations are derived from Equation 4.

$$\sigma_e = \frac{\sigma_\varepsilon}{\sqrt{1-\rho^2}}. \quad (4)$$

Consider a design with only two phases: baseline and treatment. Let the time-points in the baseline phase be $1, 2, \ldots, t_b$ and in the treatment phase be $t_{b+1}, \ldots, t_n$. Then the intercept β_{0p} can be modeled as:

$$\beta_{0p} = \begin{cases} \beta_{01}, if\ t \leq t_b \\ \beta_{02}, otherwise \end{cases}. \quad (5)$$

The intercepts are drawn from normal distributions with hyperpriors (i.e., prior on a prior) in order to reduce the impact of prior specification on the estimates (Natesan, Nandakumar, Minka, & Rubright, 2016). The means of these normal distributions (μ_{0p}) are independently drawn from normal distributions with standard deviations for each phase independently drawn from gamma distributions.

$$\beta_{0p} \sim norm(\mu_{0p}, \sigma_p^2) \quad (6)$$

$$\mu_{0p} \sim norm(0, 100); p = 1, 2 \quad (7)$$

$$\sigma_p \sim gamma(1, 1). \quad (8)$$

Although the use of appropriate priors is very much a growing field and there is no generic guidance on whether there is a prior that works for all parameters (this is probably not possible), the general rule for use of priors is to use reasonable estimates with reasonable uncertainty specification. For instance, Natesan, Nandakumar, Minka, and Rubright (2016) conducted a study of prior comparisons that showed that using priors that matched the generating distribution produced comparably good estimates as hierarchical priors as used in the present study. However, using extremely less informative priors such as having a very large standard deviation led to improper posteriors. In general, when nothing is known about the estimates, a sensitivity analysis where different priors are tested to see how they affect the posterior distributions is recommended.

An effect size estimate of the treatment can be obtained from the posterior distribution of the rate ratio of the mean of the distribution from which the intercepts are drawn as given in Equation 9.

$$\mu_{ratio} = \frac{e^{\mu_2}}{e^{\mu_1}} \quad (9)$$

The rate ratio can be interpreted as the ratio of the rate between the treatment and the baseline phases. Larger rate ratio values are desirable for positive outcome variables because this would indicate the effectiveness of the intervention in increasing the occurrence of positive outcome variables in the treatment phase compared to that of the baseline phase. In the ABAB design, there will be 3 rate ratio effect sizes that will measure intervention effect, removal of intervention effect, and reintroduction of intervention effect.

The details of the Gibbs sampler are given in Appendix A. We now demonstrate how these concepts can be applied to a real dataset obtained from a SCED that implemented a function-based comprehensive behavioral intervention. This intervention was implemented to decrease problem behavior and increase socially appropriate behavior of four children in an elementary school.

ABAB Example

In the study by Lambert, Cartledge, Heward, and Lo (2006), the effect of response cards on disruptive behavior of urban fourth-grade students during Math lessons was measured. The baseline phase was with a single-student responding and the treatment phase was where each student would write a response to a question posed by the teacher. Students with frequent disruptive behaviors in the classroom were selected to participate in the study. We chose this study's data because it used ABAB design with count data, which was appropriate for demonstrating BRR for count data. The number of disruptive behaviors during single-student responding (SSR) and response card phase (RC) for the students was the outcome variable. The data are plotted in figure 1.

INSERT FIGURE 1 ABOUT HERE

Visual Analysis Results and Discussion

Visual analysis is commonly used to determine the existence of a functional relation between the independent and dependent variables and to specifically determine the stability of the behavioral pattern, change in the level of performance, immediacy of effect, direction of the trend line, and consistency in data across similar phases. In addition to visual analyses, the Nonoverlap of All Pairs (Parker et al., 2011) effect size was computed to determine the magnitude of effect. NAP is a non-parametric technique to measure overlap for two phases and

yields the percentage of improvement data across adjacent phases. It does not account for trend. Although Parker and Vannest (2009) claim that it is appropriate for nearly all data types and distributions, it cannot distinguish between various levels of non-overlap. For instance, a 100% non-overlap could be due to outliers or unusually big effects or very small effects.

As shown in Figure 1 and noted by the authors, the mean and median for Group A across blocks of sessions during the first baseline was 7 instances of disruptive behaviors. The mean decreased to 0.5 during the first intervention phase (median = 0). Similarly, the mean increased to 7.875 (median = 8) during the second baseline and then decreased to 2 (median = 2) when intervention was reinstituted. Data show an immediate effect where the average of the last three baseline I data points show an average of 6.33 which decreased to 0.67 for the first three data points in intervention I. A similar pattern was noted following a reversal and reinstitution of intervention, going from an average of 0.9.33 in baseline II to 2.66 for the first three data points for intervention II. The trend is not clearly discernible from the figure. Finally, data also show consistency in the pattern across similar phases when the independent variable was manipulated to document replications of effect. Results for the NAP effect size are given in Table 1. The results show that there is no overlap between the phases.

INSERT TABLE 1 ABOUT HERE

Statistical Analyses

JAGS 4.0.0 (Just another Gibbs sampler, Plummer, 2003) was used to fit the data. The R package runjags (Denwood, 2016) runs parallel chains and iterates the model estimates until convergence. Runjags checks convergence using two convergence diagnostics: the multivariate potential scale reduction factor (MPSRF, Brooks & Gelman, 1998) and Heidelberger and Welch's convergence diagnostic (Heidelberger & Welch, 1983). Four chains were run. The

corresponding JAGS code is given in github. The estimates are shown in Table 2.

INSERT TABLE 2 ABOUT HERE

Level changes. The estimated levels of the outcome variable in both baseline and intervention phases are approximately equal to the means reported in the visual analysis section (6.45 and 0.5, respectively). Readers can compare the exponent of the β_{01} (from equation 6) means and quantiles with the reported means. For instance, the mean in the first baseline phase was 7 and in the intervention phase was 0.5 in the visual analysis. The posterior mean of β_{01} was 6.45 $[\exp(1.86) = 6.45]$. The difference in the mean of the outcome variable of the Bayesian and visual analyses estimates was less than or equal to 1. The posterior means of autocorrelations of the data ranged from .04 to .28. The corresponding posterior standard deviations ranged from .18 to .27. The posterior standard deviations are rather small for all phases. The posterior means of autocorrelations are not negligible for two phases indicating that the results would be suspect if autocorrelations were not modeled. However, the autocorrelation was negligible for the third phase.

INSERT FIGURE 2 ABOUT HERE

Rate ratio effect size. Posterior density plots of the rate ratio effect sizes are given in Figure 2. To recap, the rate ratio is interpreted as a reduction or increase in treatment compared to the baseline. Therefore, the decrease in the outcome variable was .06 times of what it was in baseline phase I. When considering 95% of the highest density interval (HDI) of the posterior density, the outcome variable in intervention phase I is .02 to .17 times of what it was in baseline phase I. This shows that disruptive behaviors in the first intervention phase were only 0.2 to 17% of what they were in the baseline phase. Similarly, the outcome variable in baseline phase II is 2.9 to 23 times higher than it was in intervention phase I and the outcome variable in intervention

phase II is 0.144 to 0.38 times lower than in baseline phase II. The researcher is now free to choose a region of practical equivalence (ROPE, Kruschke, 2013) where the null hypothesis that there was no effect can be accepted based on what values the researcher deems to be negligibly different from the null. The posterior distribution of the rate ratio in Figure 2 gives both the probable values of the rate ratio and their corresponding probabilities (i.e., probability density). For instance, the effect size between baseline phase I and intervention phase I is peaked at .06 (mode) and its probable values run from approximately .02 to .2 with 95% of the values lying between .02 and 0.17. This is the 95% highest density interval.

Suppose we decide that a treatment is effective only if the outcome variable is decreased to not more than 40% of the original frequency of disruptive behaviors. We see that none of the posterior distributions for the phase changes between baseline I and intervention I, and baseline II and intervention II contain the value of .4. Similarly, the posterior between phase change from intervention I to baseline II does not contain the reciprocal of .4 which is 2.5. The null can be accepted for all phase changes because the probabilities of the rate ratio of outcome variable being less than .4 times for phase change between baseline and intervention and greater than 2.5 times for phase change between intervention and baseline are 100% for all 3 phase changes as seen in Figure 2. The vertical lines at .4 and 2.5 in the figure show the hypothesized value chosen by the researcher and the percentage value in the figure represents the probability mass that falls on the right side of the hypothesized value for the first and the last phase changes. Obviously, this direction is reversed for the change in phases between intervention I and baseline II because here the researcher would be looking for an increase in the outcome variable to at least 2.5 times the value in the intervention I phase.

In sum, we can see that the results of the statistical analysis support the results of visual analysis with respect to the median and mean estimates of the outcome variable in each phase and the presence of immediacy effect. However, what BRR adds over the visual analysis is that it produces a statistically sound effect size with posterior distribution which can be used to make decisions about the statistical significance of the effect, and produces estimates of autocorrelation, and other statistics along with their respective posterior distributions. This is clearly an addition to the existing protocol that is generally used for analyzing count data in SCEDs.

Limitations

One key limitation of the BRR that we present, relative to synthesis work, is it is based on within-subject and not between subject contrasts. Hence, it is not (yet) appropriate for use in synthesizing effects size estimates derived from group-design studies like randomized controlled trials. Fundamentally, the variance properties in SCEDs and group-design studies tend to be very different, rendering different playing fields (see for e.g., Lipsey and Wilson, 2001). We do anticipate that with extensive simulation work a procedure that allows for syntheses across effect sizes is possible, but for now BRR should be seen as a way to gain statistical insights into single studies and facilitate syntheses across multiple SCEDs. There is a learning curve associated with implementing the codes that are attached to this article. But the rewards for computing BRR are well worth the effort as we have demonstrated in the previous sections. The appropriate use of priors needs to addressed at this juncture. Improper priors can lead to improper posteriors. Therefore, it is recommended that researchers try various prior specifications for their analyses and test if the posteriors are sensitive to prior specification. This type of sensitivity analysis can produce more confidence in the posterior estimates. Finally, the fact that this is a ratio presents

the researcher with a possible set of two problems: (1) when the denominator value is zero or very close to zero, (2) when the denominator value has a very large posterior standard deviation. Although one could logically truncate the posterior by removing the lowest and the highest, say 1% of the posterior estimates to compute the rate ratio, this is a crude fix. Thus, this are remains an avenue for further research.

Conclusion

The main purpose of the present study was to present the BRR effect size for supplementary use in the analysis of SCED data. The model we presented estimates autocorrelation along with BRR. The estimates of the intercepts from both visual and statistical analysis were similar. The autocorrelation estimates could not be computed using visual analysis, but their rather high values from the statistical analysis shows that this statistic cannot be ignored. The BRR also produced plots that showed regions of practical equivalence which are a nice addition to the visual plots in SCED analysis. Researchers can visually detect the magnitude of the effect based on the posterior distributions of the rate ratio. However, we cannot altogether abandon visual analysis because of their simplicity and ease of use. As mentioned before, Bayesian estimation has a steep learning curve associated with it and many articles, workshops, and courses need to be made available to help applied researchers use this method. Shiny apps and easy to use software tools would also help improve the user-friendliness of the methodology.

BRR can (a) deal with count data, (b) handle small samples, (c) produce interval estimates of uncertainty, and (d) complement visual analyses via ROPE. In principle, BRR could also be used in SCED research syntheses (a point to be demonstrated in future work). Hence, BRR is a new and useful analytic tool for SCED analyses. We have shared the programs used to produce rate ratio effect size estimates and the ROPE comparisons.

The fact that the rate ratio estimation is made possible even for such short time-series is compelling evidence of the flexibility of Bayesian modeling. To date, the model we presented is the only inferential statistical procedure that estimates intercepts and effect sizes, accounts for autocorrelations and small sample sizes, and works with count data. This form of estimation can thus yield better understanding of SCED data and could, in principle, support SCED research syntheses. However, recall the limitation that the effect size used here is still based on a within-subjects design. Further work is needed to understand if and how Bayesian procedures might quantitatively synthesize within-subject and between-subject effects.

We think that researchers who use SCEDs when combining research and practice (praxis) or researchers who engage in synthetic SCED work will be interested in use of the BRR as either an alternate or complementary analytic approach. Furthermore, given the amount of SCED synthesis work conduced in the field, exploring the use of new effect sizes that account for several difficulties with OLS methods and can be used to complement visual analyses should provide a basis for the long-term viability of the BRR. The BRR will be of further utility within psychology research (and science-practitioners who conduct SCEDs) if user-friendly internet freeware can be developed and tested, which is a step that will be pursued in future work.

References

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics, 17*(3), 251-269. doi: 10.3102/10769986017003251

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

Ansari, A., Jedidi, K., & Jagpal, S. (2000). A hierarchical Bayesian approach for modeling heterogeneity in structural equation models. *Marketing Science, 19*, 328-347. doi: 10.1287/mksc.19.4.328.11789

Briesch, A.M., & Briesch, J.M. (2016). Meta-Analysis of behavioral self-management interventions in single-case research. *School Psychology Review, 45*, 3-18. doi:. 10.17105/SPR45-1.3-18

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*, 531-563. doi: 10.1177/0145445503261167

Chaffee, R.K, Briesch, A.M., Johnson, A.H., & Volpe, R.J. (2017). A meta-Analysis of class-wide interventions for supporting student behavior. *School Psychology Review, 46*(2), 149-164. doi: 10.17105/SPR-2017-0015.V46-2

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (second ed). Hillsdale, NJ: Lawrence Erlbaum.

Cook, B.G. (2014). A call for examining replication and bias in special education research. *Remedial and Special Education, 35*(4), 233-246. doi: 10.1177/0741932514528995

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997-1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Dart, E.H., Collins, T.A., Klingbeil, D.A., & McKinley, L.E. (2014) Peer management interventions: A meta-analytic review of single-case research. *School Psychology Review, 43*(4), 367-384. doi: 10.17105/SPR-14-0009.1

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

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.

Gast, D. L. & Ledford, J. R. (2014). *Single subject research methodology in behavioral sciences* (2nd ed.). New York, NY: Routledge.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). London, England: Chapman & Hall.

Gelfand, A.E., & Smith, A.M.E (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85,* 398-409. doi: 10.1080/01621459.1990.10476213

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.

Gresham, F. M., Sugai, G., & Horner, R. H. (2001). Interpreting outcomes of social skills training for students with high-incidence disabilities. *Exceptional Children, 67*, 331-344. doi: 10.1177/001440290106700303

Harrop, J. W. & Velicer, W. F. (1985). A comparison of alternative approaches to the analysis of interrupted time-series. *Multivariate Behavioral Research, 20*, 27-44. doi: 10.1207/s15327906mbr2001_2

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

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. doi: 10.1002/jrsm.1052

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across studies. *Research Synthesis Methods, 4*, 324-341. doi: abs/10.1002/jrsm.1086

Heidelberger, P. & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research, 31*, 1109-44. doi: 10.1287/opre.31.6.1109

Hitchcock, J. H., Horner, R. H., Kratochwill, T. R., Levin, J. R., Odom, S. L., Rindskopf, D. M & Shadish, W. R. (2014). The What Works Clearinghouse Single-Case Design Pilot Standards: Who will guard the guards? *Remedial and Special Education 35*(3), 154-152. doi: 10.1177/0741932513518979

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. doi: 10.1177/001440290507100203

Horner, R. H., & Kratochwill, T. R. (2012). Synthesizing single-case research to identify evidence-based practices: Some brief reflections. *Journal of Behavioral Education 21*, 266-272. DOI:10.1007/s10864-012-9152-2

Horner, R., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). *Expanding analysis of single case research*. Washington, DC: Institute of Education Science, U.S. Department of Education.

Huitema, B. E., & McKean, J. W. (1994). Two biased-reduced autocorrelation estimators: r_{F1} and r_{F2}. *Perceptual and Motor Skills, 78*, 323–330. doi:10.2466/pms.1994.78.1.323

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

Kilgus, S.P., Riley-Tillman, C., & Kratochwill, T.R. (2016) Establishing interventions via a theory-driven single case design research cycle. *School Psychology Review, 45*(4), 477-498. doi: 10.17105/SPR45-4.477-498

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M & Shadish, W. R. (2010). *What Works Clearinghouse: Single-case design technical documentation. Version 1.0 (Pilot)*. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_scd.pdf.

Kratochwill, T.R., Hitchcock, J., 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. doi: 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.

Lambert, M.C, Cartledge, G., Heward, W.L., & Lo, 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,* 88–99.

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis.* Thousand Oaks, CA: Sage.

Lenovaz, M. J., & Rapp, J. T. (2016). Using single-case experiments to support evidence-based decisions. *Behavior Modification, 40*, 377-395. doi: 10.1177/0145445515613584

Maggin, D. M., & Chafouleas, S. M. (2013). Introduction to the special series: Issues and advances of synthesizing single-case research. *Remedial and Special Education, 34*, 3-8. doi:10.1177/0741932512466269

Maggin, D. M., O'Keefe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985–2009. *Exceptionality, 19,* 109-135. doi: 10.1080/09362835.2011.565725

Maggin, D, M., Chafouleas, S.M., Goddard, K.M., Johnson, A.H. (2011). A systematic evaluation of token economies as a classroom management tool for students with challenging behavior. *Journal of School Psychology, 49*(5), 529-554. soi: 10.1016/j.jsp.2011.05.001

Matyas, T. A. & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341–351. doi: 10.1901/jaba.1990.23-341/full

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to

analyze linear models with autoregressive error terms. *Psychological Methods, 5*, 87–101. doi:10.1037/1082-989X.5.1.87

Meadan, H., Snodgrass, M. R., Meyer, L. E., Fisher, K. W., Chung, M. Y., & Halle, J. W. (2016). Internet-based parent-implemented intervention for young children with autism. *Journal of Early Intervention, 38*, 3-23. doi: http://dx.doi.org/10.1177/1053815116630327

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. doi:10.1037/met0000136

Nastasi, B. K. & Hitchcock, J. H. (2016). *Mixed methods research and culture-specific interventions: Program design and evaluation*. (The New Mixed Methods Research Series). Thousand Oaks, CA: Sage.

Natesan, P. (2019). Fitting Bayesian Models for Single-Case Experimental Designs: A Tutorial. *Methodology, 15,* 147-156. https://doi.org/10.1027/1614-2241/a000180.

Natesan, P., Boedeker, P., & Onwuegbuzie, A. J. (2018). Adopting a meta-generative way of thinking in the field of education via the use of Bayesian methods. Paper presented at the Annual meeting of the American Educational Research Association, Toronto, Canada.

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

Natesan, P. & Hedges, L. V. (2019). Accurate model vs. accurate estimates: A study of Bayesian single-case experimental designs. Presented at the annual meeting of the American Educational Research Association, Toronto, Canada.

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. (2016). Bayesian Prior Choice in IRT estimation using MCMC and Variational Bayes. *Frontiers in Psychology: Quantitative Psychology and Measurement, 7*, 1-11.

Natesan, P., Minka, T., & Hedges, L. V. (In Press). Investigating immediacy in multiple phase-change single case experimental designs using a Bayesian unknown change-points model. *Behavioral Research Methods.*

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children, 71*, 137–148. doi: 10.1177/001440290507100201

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect Size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*, 303-322. doi:10.1177/0145445511399147

Parker, R. I. & Vannest, K. J. (2009). An improved effect size for single-case research: nonoverlap of all pairs. *Behavior Therapy, 40*, 357-367. doi: 10.1016/j.beth.2008.10.006

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. In Proceedings of the 3rd international workshop on distributed statistical computing.

Roane, H. S., Rihgdahl, J. E., Kelley, M. E., & Glover, A. C. (2011). Single-case experimental designs. In W. W. Fisher, C. C. Piazza, & H. S. Roane (Eds.), *Handbook of applied behavior analysis* (pp. 132–147). New York: Guilford Press.

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

Shadish, W. R., Rindskopf, D. M., Hegdes, 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 treatment effects in 2008. *Behavior Research Methods, 43*, 971-980. doi: 10.3758/s13428-011-0111-y

Soares, D.A., Harrison,J.R., Vannest, K.J., McClelland, S.S. (2016) Effect Size for token economy use in contemporary classroom settings: A meta-analysis of single-case research. *School Psychology Review, 45*(4), 379-399. doi: 10.17105/SPR45-4.379-399

Strickland-Cohen, M. K., & Horner, R. H. (2015). Typical School Personnel Developing and Implementing Basic Behavior Support Plans. *Journal of Positive Behavior Interventions, 17*, 83 –94. DOI: 10.1177/1098300714554714

Velicer, W. F., & Molenaar, P. (2013). Time series analysis: Research methods in psychology. In J. Schinka and W. F. Velicer (eds). *Volume 2 of Handbook of Psychology,* 2nd Ed. New York: John Wiley & Sons, pp. 628–660.

What Works Clearinghouse (2016). *Functional behavioral Assessment-based interventions*. Retrieved from https://ies.ed.gov/ncee/wwc/EvidenceSnapshot/667

Appendix A

Gibbs Sampler for the Bayesian Rate Ratio (BRR)

Sampling Algorithm

A posterior distribution is typically obtained from a sampling algorithm. The Gibbs sampler is one of the most frequently used Markov chain Monte Carlo (MCMC) methods in Bayesian estimation (Albert, 1992; Gelfand & Smith, 1990; Geman & Geman, 1984). Let us consider a time-series SCED data $Y = (y_1, y_2, \dots, y_n)$ such that the functional relationships between the dependent and the independent variables differ based on the phase, which is unknown. In Equation 9, $\theta_1 = g(\beta_{01},\ \sigma_\varepsilon, \rho)$ and $\theta_2 = g(\beta_{02}, \sigma_\varepsilon, \rho)$ where g is a function of the parameters within parentheses.

$$y_t = \begin{cases} \theta_1 \; if \; t \; \leq t_b, \\ \theta_2 \; otherwise \end{cases}. \tag{A1}$$

Equation A1 is a different way of presenting the model in equations 1-8. In the parameter vector $\Theta = (\beta_{01}, \beta_{02}, \sigma_\varepsilon, \rho, t_b)$, all parameters are independent *a priori*. The posterior distribution $\pi(\Theta|Y)$ can be obtained using the Gibbs sampler.

A generic Gibbs sampler follows an iterative process. Consider a simple bivariate normal distribution with parameter vector (x_1, x_2). A set of starting values are assigned to the vector at step 0 of the iteration. The value of each parameter is updated iteratively holding all other values constant. The process is Markov chain because the value of parameter x_1 at the jth iteration depends only on the value of parameter x_2 at the $(j-1)th$ iteration. Figure A1a shows this process. Note that in Figure A1a when each parameter is updated, it moves along the value of the other parameter and only along its own axis. That is, when x_1 is updated, the sampler moves along the x_1 direction. Similarly, when x_2 is updated, the sampler moves along the x_2 direction. The first few sampled values are allowed to *burn-in* in order to avoid the effect of the starting

values on the estimates. The rest of the sampled values form the posterior distribution of the parameter as shown in Figure A1b.

INSERT FIGURES A1a AND A1b ABOUT HERE

When we extend the Gibbs sampler from the bivariate case discussed above to a multivariate case for the model in equation 2, the parameter vector is $(\beta_{01},\ \beta_{02},\ \sigma_{\varepsilon}, \rho, t_b)$. We assign a set of starting values, S to the vector at step 0 of the iteration. Let the iteration be indexed using the variable j and then the following eight-step process can be followed.

Step 1: Set $j = j + 1$

Step 2: Sample $\left(\beta_{01}^{j} \middle| \beta_{02}^{j-1}, \sigma_{\varepsilon}^{j-1}, t_b^{j-1}, \rho^{j-1}, Y\right)$

Step 3: Sample $\left(\beta_{02}^{j} \middle| \beta_{01}^{j}, \sigma_{\varepsilon}^{j-1}, t_b^{j-1}, \rho^{j-1}, Y\right)$

Step 4: Sample $(\sigma_{\varepsilon}^{j} | \beta_{01}^{j}, \beta_{02}^{j},\ t_b^{j-1}, \rho^{j-1}, Y)$

Step 5: Sample $(t_b^{j} | \beta_{01}^{j},\ \beta_{02}^{j}, \sigma_{\varepsilon}^{j}, \rho^{j-1}, Y)$

Step 6: Sample $(\rho^{j} | \beta_{01}^{j},\ \beta_{02}^{j}, \sigma_{\varepsilon}^{j}, t_b^{j}, Y)$

Step 7: Sample $(\hat{Y}^{j} | \beta_{01}^{j},\ \beta_{02}^{j}, \sigma_{\varepsilon}^{j}, t_b^{j}, \rho^{j}, Y)$, where $\hat{Y}^{j}$ is the vector of predicted values of Y at the jth iteration

Step 8: Return to Step 1 until desirable number of iterations is complete.

For the current algorithm, Figure A1b would be expanded to multiple dimensions and is based on values from the posterior distributions, which are more informative than a frequentist point estimate and standard error, which do not treat the estimate as a parameter. This means the estimate has no shape in the frequentist framework. Because the posterior density is made up of the possible values the parameter can take and their associated probabilities, the posterior density is a probability density function and can be interpreted as such.

To summarize, the most pertinent issue to social scientists is the fact that the probability of the true parameter value lies within a 95% interval; representing a more straightforward interpretation compared to a CI from the frequentist framework (Gelman et al., 2013).

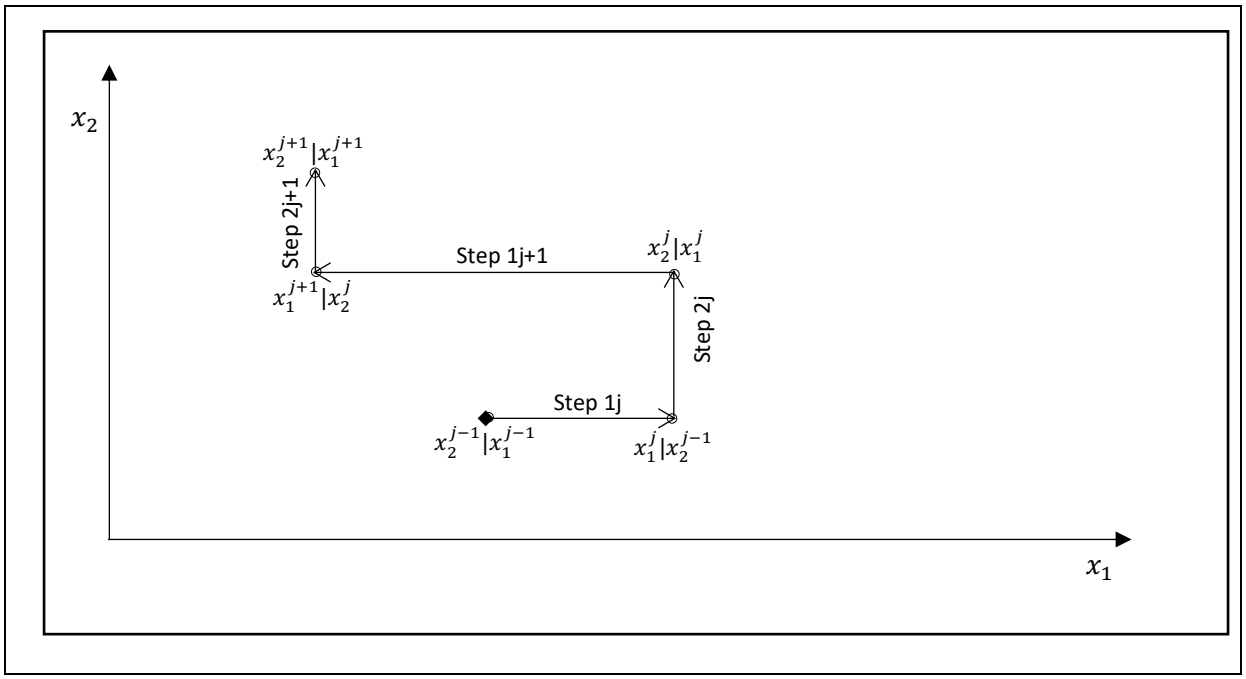

Figure A1a: Gibbs sampler where x_1 and x_2 are updated at the jth and the (j+1)th iteration

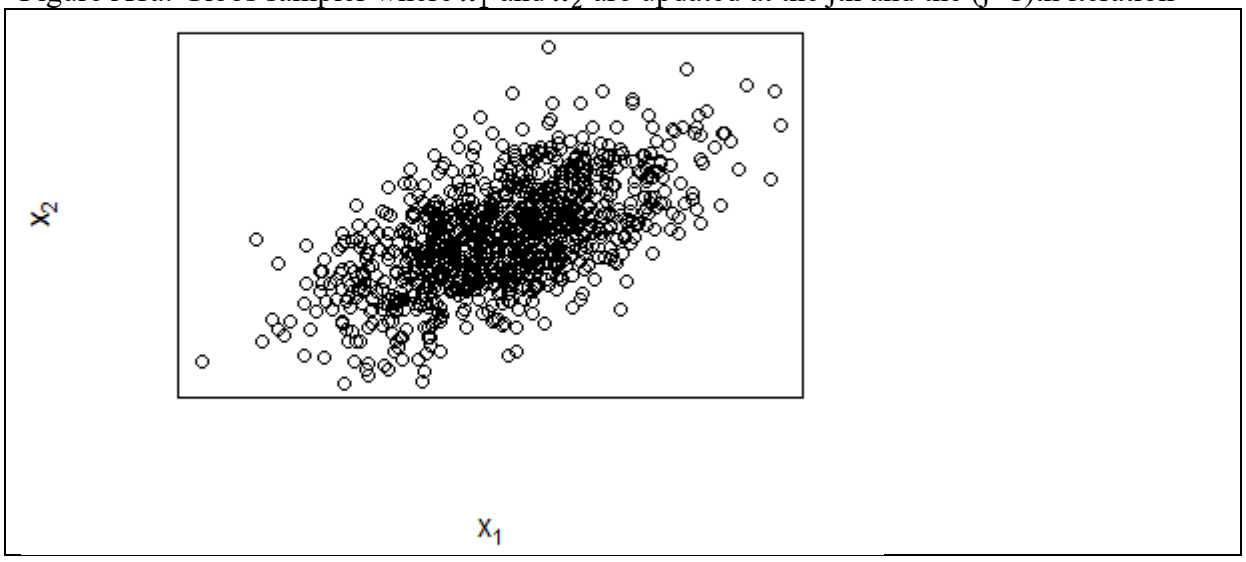

Figure A1b: Sampled values (joint posterior) of x_1 and x_2

Table 1

NAP Results

Label	S	PAIRS	NAP	VARs	SD	Z	P Value	CI 85%
A1 vs B1	-48	48	0	240	15.49	-3.1	0.002<>1	-1<>-.535
B1 vs A2	48	48	1	240	15.49	3.1	0.002<>1	.535<>1
A2 vs B2	-66	72	.0417	432	20.78	-3.2	0.001<>1	-1<>-.501
Combined	-66	336	0.3237	-	0.1576	0	0.071<>1	0.066<>0.582

A1, B1, A2, and B2 refer to baseline I, intervention I, baseline II, and intervention II, respectively.

Table 2
Parameter Estimates from BRR

Parameter	Lower95	Median	Upper95	Mean	SD
Baseline I to Intervention I					
ρ	-0.05	0.30	0.58	0.28	0.18
μ_1	1.53	1.87	2.19	1.86	0.17
μ_2	-1.77	-0.66	0.33	-0.69	0.54
σ_1	0.36	1.20	4.45	1.84	16.44
σ_2	0.36	1.19	4.46	1.75	2.42
μ_{ratio}	0.03	0.08	0.19	0.09	0.04
Intervention I to Baseline II					
ρ	-0.30	0.25	0.74	0.23	0.27
μ_1	-1.31	-0.33	0.69	-0.34	0.51
μ_2	1.59	1.99	2.30	1.97	0.18
σ_1	0.36	1.20	4.43	1.77	3.85
σ_2	0.36	1.19	4.43	1.77	2.87
μ_{ratio}	4.18	10.20	26.22	11.43	5.81
Baseline II to Intervention II					
ρ	-0.36	0.05	0.41	0.04	0.20
μ_1	1.75	2.02	2.28	2.02	0.14
μ_2	0.16	0.66	1.12	0.65	0.24
σ_1	0.37	1.20	4.43	1.79	3.65
σ_2	0.38	1.20	4.44	1.76	3.17
μ_{ratio}	0.17	0.26	0.40	0.26	0.06

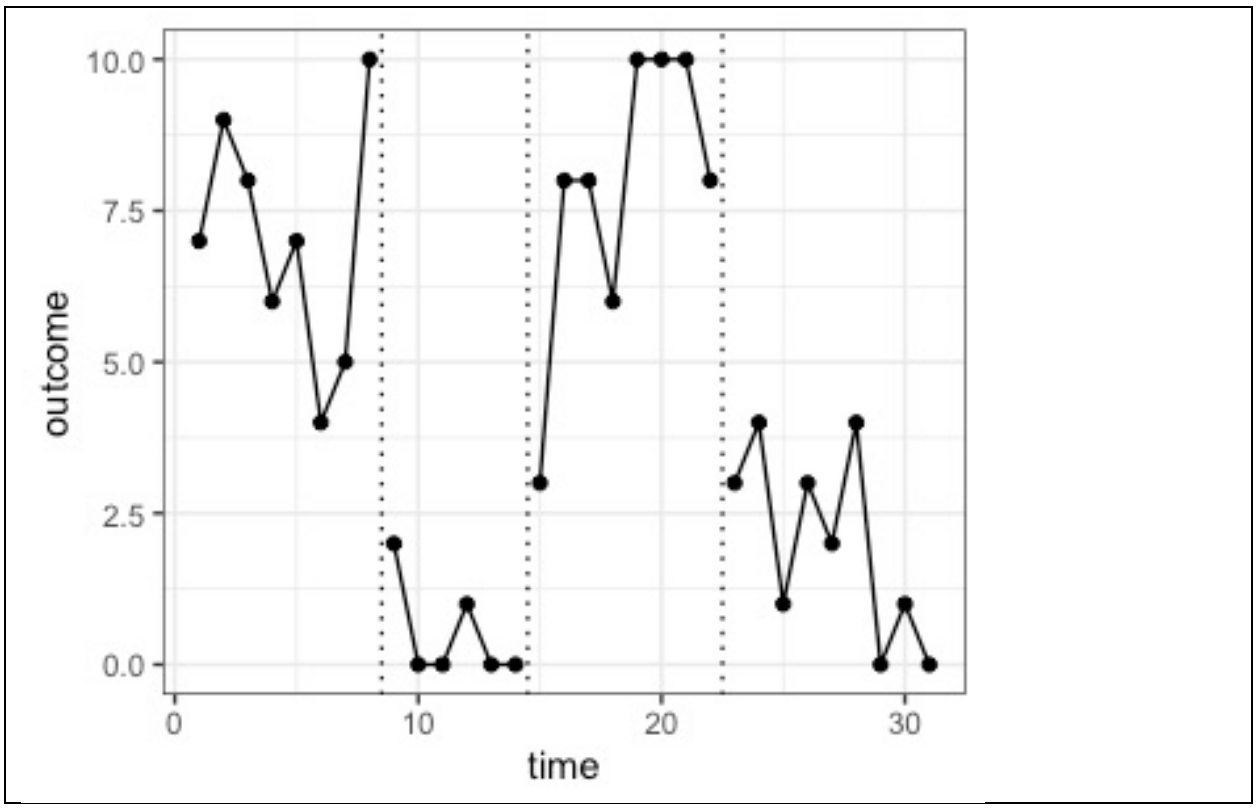

Figure 1. Data Plot of Shih et al. (2015)

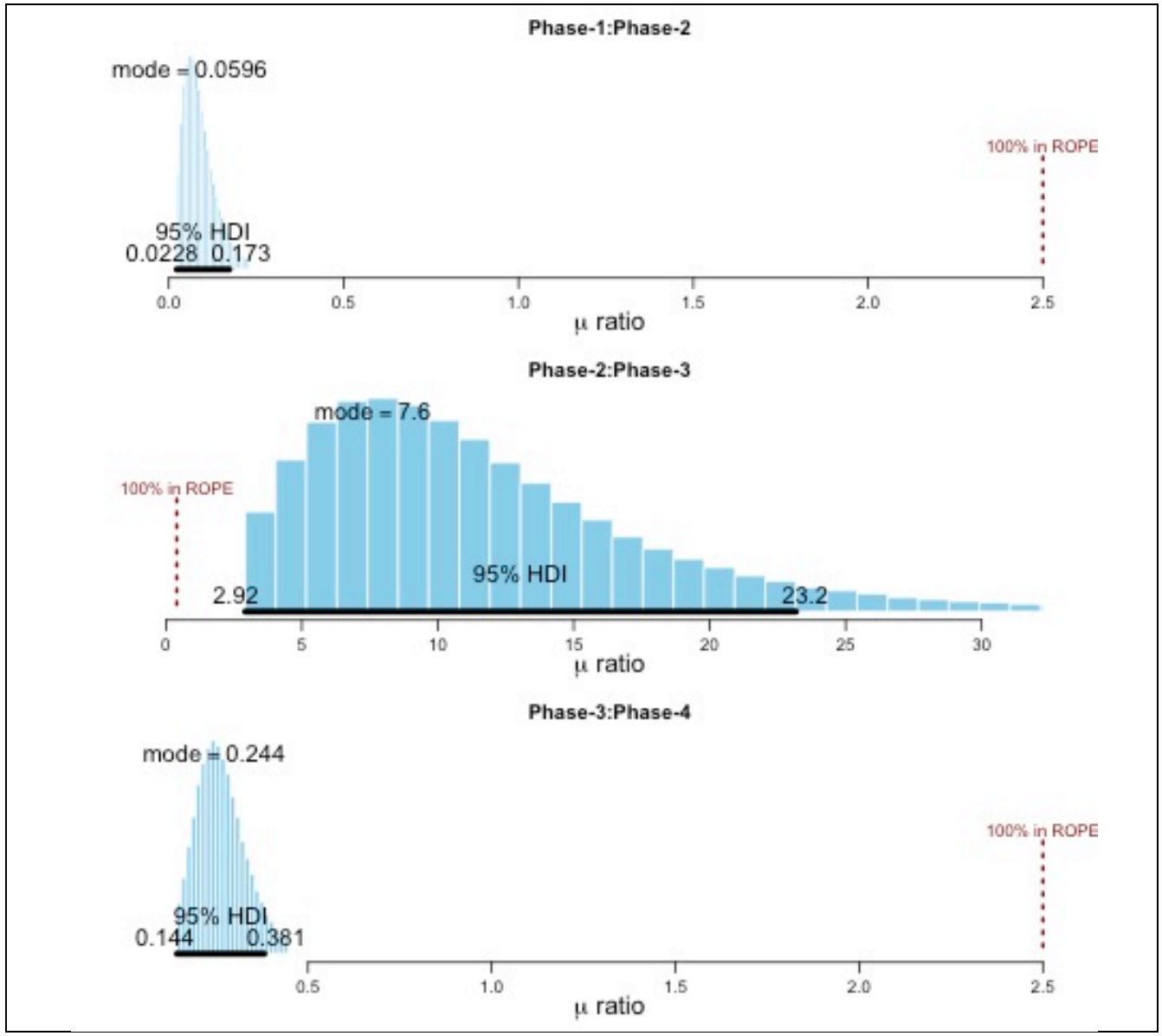

Figure 2: Posterior density plots of the rate ratio of the intercepts

Appendix A

Gibbs Sampler for the Bayesian Rate Ratio (BRR)

Sampling Algorithm

A posterior distribution is typically obtained from a sampling algorithm. The Gibbs sampler is one of the most frequently used Markov chain Monte Carlo (MCMC) methods in Bayesian estimation (Albert, 1992; Gelfand & Smith, 1990; Geman & Geman, 1984). Let us consider a time-series SCED data $Y = (y_1, y_2, \dots, y_n)$ such that the functional relationships between the dependent and the independent variables differ based on the phase, which is unknown. In Equation 9, $\theta_1 = g(\beta_{01},\ \sigma_\varepsilon, \rho)$ and $\theta_2 = g(\beta_{02}, \sigma_\varepsilon, \rho)$ where g is a function of the parameters within parentheses.

$$y_t = \begin{cases} \theta_1 \; if \; t \; \leq t_b, \\ \theta_2 \; otherwise \end{cases}. \qquad \text{(A1)}$$

Equation A1 is a different way of presenting the model in equations 1-8. In the parameter vector $\Theta = (\beta_{01}, \beta_{02}, \sigma_\varepsilon, \rho, t_b)$, all parameters are independent *a priori*. The posterior distribution $\pi(\Theta|Y)$ can be obtained using the Gibbs sampler.

A generic Gibbs sampler follows an iterative process. Consider a simple bivariate normal distribution with parameter vector (x_1, x_2). A set of starting values are assigned to the vector at step 0 of the iteration. The value of each parameter is updated iteratively holding all other values constant. The process is Markov chain because the value of parameter x_1 at the jth iteration depends only on the value of parameter x_2 at the $(j-1)th$ iteration. Figure A1a shows this process. Note that in Figure A1a when each parameter is updated, it moves along the value of the other parameter and only along its own axis. That is, when x_1 is updated, the sampler moves

along the x_1 direction. Similarly, when x_2 is updated, the sampler moves along the x_2 direction. The first few sampled values are allowed to *burn-in* in order to avoid the effect of the starting values on the estimates. The rest of the sampled values form the posterior distribution of the parameter as shown in Figure A1b.

INSERT FIGURES A1a AND A1b ABOUT HERE

When we extend the Gibbs sampler from the bivariate case discussed above to a multivariate case for the model in equation 2, the parameter vector is $(\beta_{01},\ \beta_{02},\ \sigma_\varepsilon, \rho, t_b)$. We assign a set of starting values, S to the vector at step 0 of the iteration. Let the iteration be indexed using the variable j and then the following eight-step process can be followed.

Step 1: Set $j = j + 1$

Step 2: Sample $(\beta_{01}^j | \beta_{02}^{j-1}, \sigma_\varepsilon^{j-1}, t_b^{j-1}, \rho^{j-1}, Y)$

Step 3: Sample $(\beta_{02}^j | \beta_{01}^j, \sigma_\varepsilon^{j-1}, t_b^{j-1}, \rho^{j-1}, Y)$

Step 4: Sample $(\sigma_\varepsilon^j | \beta_{01}^j, \beta_{02}^j,\ t_b^{j-1}, \rho^{j-1}, Y)$

Step 5: Sample $(t_b^j | \beta_{01}^j,\ \beta_{02}^j, \sigma_\varepsilon^j, \rho^{j-1}, Y)$

Step 6: Sample $(\rho^j | \beta_{01}^j,\ \beta_{02}^j, \sigma_\varepsilon^j, t_b^j, Y)$

Step 7: Sample $(\hat{Y}^j | \beta_{01}^j,\ \beta_{02}^j, \sigma_\varepsilon^j, t_b^j, \rho^j, Y)$, where $\hat{Y}^j$ is the vector of predicted values of Y at the jth iteration

Step 8: Return to Step 1 until desirable number of iterations is complete.

For the current algorithm, Figure A1b would be expanded to multiple dimensions and is based on values from the posterior distributions, which are more informative than a frequentist point estimate and standard error, which do not treat the estimate as a parameter. This means the estimate has no shape in the frequentist framework. Because the posterior density is made up of the possible values the parameter can take and their associated probabilities, the posterior density is a probability density function and can be interpreted as such.

To summarize, the most pertinent issue to social scientists is the fact that the probability of the true parameter value lies within a 95% interval; representing a more straightforward interpretation compared to a CI from the frequentist framework (Gelman et al., 2013).

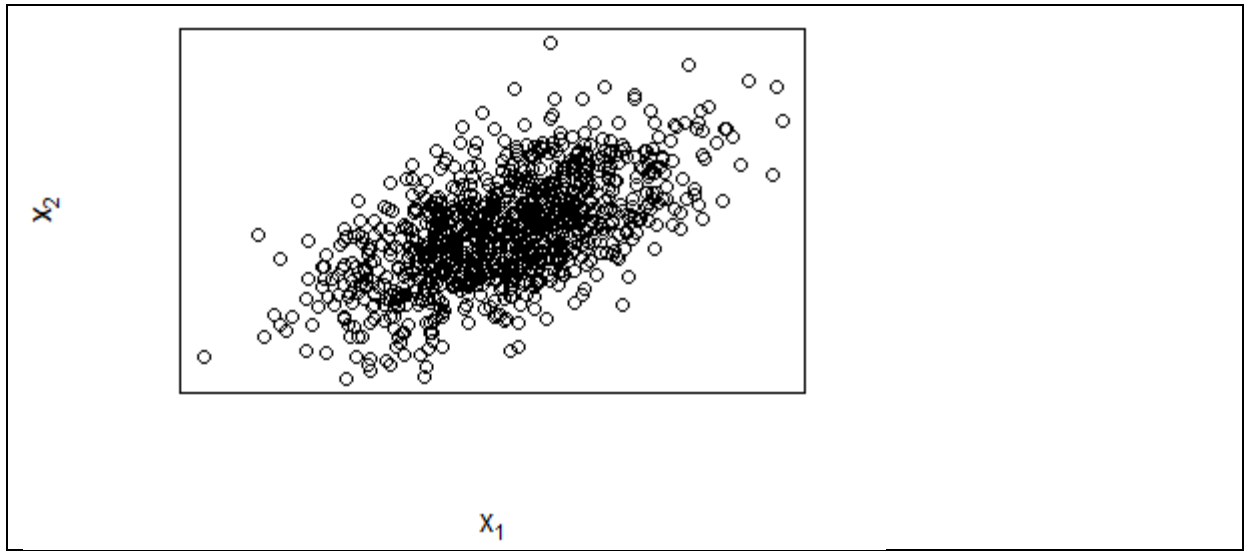

Figure A1b: Sampled values (joint posterior) of x_1 and x_2