Abstract (126 words)

Recently, there has been an increased interest in developing statistical methodologies for analyzing single case experimental design (SCED) data to supplement visual analysis. Some of these are simulation-driven such as Bayesian methods because Bayesian methods can compensate for small sample sizes, which is a main challenge of SCEDs. Two simulation-driven approaches: Bayesian unknown change-point model (BUCP) and simulation modeling analysis (SMA) were compared in the present study for four real datasets that exhibit "clear" immediacy, "unclear" immediacy, and delayed effects. Although SMA estimates can be used to answer some aspects of functional relationship between the independent and the outcome variables, they cannot address immediacy or provide an effect size estimate that considers autocorrelation as required by the What Works Clearinghouse (WWC) Standards. BUCP overcomes these drawbacks of SMA. In final analysis, it is recommended that both visual and statistical analyses be conducted for a thorough analysis of SCEDs.

Keywords: Bayesian; Markov chain Monte Carlo; Single case designs; Simulation Modeling Analysis; small samples

Comparing the Bayesian Unknown Change-Point Model and Simulation Modeling Analysis to Analyze Single Case Experimental Designs

Single case experimental designs (SCEDs) investigate change within an individual or a sampling unit rather than aggregate change for a group of individuals or units. Fields of applications of SCEDs include special education, psychology, and medicine, among others (e.g. Allen, Baker, Nuernberger, & Vargo, 2013; Guyatt, Sackett, Taylor, Chong, Roberts, & Pugsley, 1986). SCED studies are interrupted time series designs where an outcome variable is assessed repeatedly for an individual (or unit) over different phases. There is at least one baseline (phase A) and one intervention phase (phase B), with multiple observations both before and after treatment.

SCEDs have traditionally relied on visual analysis of graphs from multiple phases for determining the presence and magnitude of a treatment effect. Often visual analysis reports are supplemented with reporting phase means, medians, percentages, and effect sizes such as standardized mean differences or indices based on the amount of data overlap between phases (Parker, Hagan-Burke, & Vannest, 2007). Although visual analysis has definite advantages with analyzing SCED data, studies have shown that the presence of autocorrelation can confound the results of visual analysis. For instance, in data with autocorrelation, it is difficult to decompose patterns due to trends (slopes) versus patterns due to autocorrelated errors. Natesan Batley and Hedges (2020) conducted a simulation study to demonstrate the lack of accuracy of slope and autocorrelation estimates in SCEDs when both these parameters are estimated due to indeterminacy. Autocorrelation is almost impossible to detect by visual analysis alone (Kazdin, 2011; Thyer & Myers, 2011). The presence of autocorrelation increases Type-I errors (Matyas & Greenwood, 1990) and decreases interrater reliabilities (Brossart, Parker, Olson, & Mahadevan,

2006) in visual analysis. In fact, Jones, Weinrott, and Vaughn (1978) found that in data with moderate-high autocorrelations, visual analysis results were reduced to nearly chance levels. Therefore, there is increasing emphasis for more objective methodologies for analyzing SCED data and determining causal inferences. What Works Clearinghouse (WWC; Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, & Shadish, 2013), American Speech Language Hearing Association (2004), and the Council for Exceptional Children (CEC; Cook, Buysse, Klingner, Landrum, McWilliam, Tankersley, & Test, 2014) have all developed standards for SCEDs.

SCED data analyses pose challenges for many reasons. First, the sample sizes are often not adequate to carry out traditional analyses typically used with grouped data (Shadish & Sullivan, 2011). Second, the observations are not independent often because they are repeated measurements of the same individual. Therefore, the errors of SCED observations typically exhibit autocorrelation (Huitema, 1985). Most parametric and non-parametric analyses assume independence of observations. Third, although maximum likelihood-based approaches can be used to accommodate and model autocorrelations, these approaches require larger sample sizes. Finally, most SCED data are count or proportion data (Rindskopf, 2014; Shadish, Zuur, & Sullivan, 2014). This further exacerbates the issues with using traditional ANOVA and regression-type analyses with SCED data

Recently, there has been an increased interest in developing statistical methodologies that can address the problems posed by SCED data and supplement visual analysis. Examples of statistical developments for SCEDs *include* multilevel modeling (e.g., Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2013), semiparametric regression models (e.g., Shadish, Zuur, & Sullivan, 2014), fully Bayesian analysis (e.g., Natesan Batley, 2020a, 2020b; Natesan Batley,

Contractor, & Caldas, 2020; Natesan Batley, Shukla Mehta, & Hitchcock, 2020; Natesan, Minka, & Hedges, 2020; Natesan, 2019; Natesan & Hedges, 2017; Rindskopf, 2014), simulation-based analysis (e.g., Borckardt, Nash, Murphy, Moore, & O'Neal, 2008), and small sample corrections to standardized mean difference effect sizes comparable to the effect sizes estimated from conventional between-subjects designs (e.g., Hedges, Pustejovsky, & Shadish, 2012, 2013). Some of these approaches are simulation-driven because simulations can compensate for small sample sizes. The present study compares two such simulation-driven approaches: the Bayesian unknown change-point model (BUCP) and the simulation modeling analysis (SMA). Apart from both being simulation-driven Monte Carlo approaches, these two can be used to estimate intercepts, slopes, and autocorrelations of single-case design data. Readers may benefit from reading more basic material presented in Natesan (2019) and Natesan Batley, Contractor, and Caldas (2020). These two studies present the methodology and the models in detail along with software codes.

Bayesian methods use Markov chain Monte Carlo (MCMC) procedures to estimate the model parameters. The MCMC procedure is simulation-based. The estimates from many iterations for each parameter form the posterior distribution of the parameter. Because the posterior distribution is a probability distribution, statistical inferences made from these are more straightforward to interpret than those from traditional confidence intervals (e.g. Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2013; Kruschke, 2013). The Bayesian unknown change-point (BUCP) model estimates effect sizes while also taking into account the autocorrelation between the observations. Bayesian effect size estimate does not require small sample correction unlike the one proposed by Hedges, Pustejovsky, and Shadish (2012, 2013). Bayesian estimation is discussed in detail in the forthcoming sections.

Like the BUCP, SMA uses a simulation modeling procedure as the basis to counter the problem of small samples in SCEDs. The estimated parameters have associated p-values computed for the data. In addition, SMA also estimates the autocorrelations that occur due to repeated measurements on the same subject. However, SMA does not test SCED data according to the latest standards for establishing intervention effect in SCEDs (Cook et al., 2014; Kratochwill et al., 2013), particularly an effect size that considers autocorrelation in its computation and testing for immediacy. By the term establishing intervention effect, we mean the WWC standards on criteria for demonstrating evidence of relation between an independent variable and an outcome variable. These include documenting consistency of level, trend, and variability within each phase, immediacy of effect, and an effect size to demonstrate an intervention effect. Both SMA and BUCP are freely available and are easy to use.

In sum, there are two aims for this study. The first is to demonstrate the Bayesian methodology, which is the latest advance in simulation-driven approaches, for quantifying immediacy, and estimating the effect sizes for intercept differences that take into account autocorrelations. These are two aspects prescribed by the WWC standards required to establish a functional relationship between the independent and the outcome variables in SCEDs. The second aim is to compare Bayesian (BUCP) and non-Bayesian (SMA) methodologies in their effectiveness in assessing intervention effect in analyzing SCED data.

Specific advantages of Bayesian in SCEDs

Bayesian methods do not depend on asymptotic theory and work well with small samples, provided the prior distribution is appropriately specified (Gelman, et al., 2013). Therefore, the Bayesian effect sizes computed from BUCP models do not need small sample corrections. In the traditional frequentist framework, the statistical estimate is a fixed value with

an estimate of uncertainty known as standard error. Whereas, in Bayesian estimation, each parameter of interest has its own distribution. For example, when estimating mean and standard deviation, a posterior distribution is associated with each parameter. The posterior distribution can be summarized by its mean, median, mode, standard deviation, and credible or highest density intervals (HDI). For instance, a 95% credible interval is the interval spanning 95% of the posterior density. The values within the credible interval are more credible than the values outside the interval. Additionally, posterior distributions can be used to test regions of practical equivalence (ROPE, Kruschke, 2013) as opposed to conducting classical null hypothesis significance tests (NHST). Because posterior distributions are probability distributions, ROPE and credibile intervals (as opposed to confidence intervals) do not have replicability issues, unlike NHSTs and confidence intervals. Finally, Bayesian methodology provides modeling flexibility and can be extended to count and proportion data, which are more common than interval data in SCEDs.

Methods

The BUCP Methodology

In SCEDs change-point is the point when the intervention is introduced, which is the beginning of the intervention phase. This is referred to as the design change-point. In visual analysis change-point is known, and for an intervention to be significant, it is expected that the relationship between the independent variable and the outcome variable shifts in the desired direction from this point onwards for the rest of the phase. In BUCP analysis, however, the entire data from all phases are treated as one sequence of points and the BUCP algorithm searches for the change-point(s) in the sequence where there is a substantial change in the relationship between the independent variable and the outcome variable. That is, the change-point(s) is estimated. For example, when there is a clear immediacy (evidence of treatment effective at the

start of an intervention) the estimated change-point coincides with the design change-point; however, in case of delayed effect the estimated change-point will be some time after the start of the intervention. This approach is vastly different from the traditional visual analysis where the immediacy is determined by computing the difference in the medians of the first and last 3-5 observations of phases B and A, respectively in a two-phase study. This way of determining immediacy is highly subjective due to: (a) there being no guideline on how to interpret the magnitude of this difference to establish immediacy; in fact, this magnitude depends on the scale of the outcome variable; (b) not taking into account the patterns of all the data points in the phases; and (c) the median thus computed ignoring the autocorrelations between observations. The BUCP analysis on the other hand, statistically establishes immediacy and quantifies the effect size while accounting for autocorrelations.

The BUCP methodology is briefly described here for estimating intercepts[1]. For details see Natesan and Hedges (2017). For pedagogical purposes, the basic two-phase design is considered. However, the logic is the same for complex designs and data types. The observed value of the outcome variable *y* is assumed to be continuous and normally distributed.

The observed value at the first time point in phase 1, (y_{p1}), follows a normal distribution with mean $\hat{y}_{p1}$ and standard deviation σ_ε as shown in Equation 1:

$$y_{p1} \sim norm\left(\hat{y}_{p1}, \sigma_\varepsilon^2\right). \quad (1)$$

The predicted values in the following time points t are distributed as:

$$y_{pt}|H_{pt-1}, \Theta \sim norm\left(\hat{y}_{pt|(pt-1)}, \sigma_e^2\right). \quad (2)$$

[1] Only estimation of intercepts (instead of both intercepts and trends) is considered because Natesan Batley and Hedges (2020) recommend to not estimate intercept, slope and autocorrelations all in the same SCED analysis.

In Equation 2, H_{pt-1} is the past history, Θ is the vector of parameters, and σ_e is the white noise created by a combination of random error (σ_ε^2) and autocorrelation between adjacent time points (ρ). The relation between ρ (autocorrelation), σ_e (white noise), and σ_ε (random error) is

$$\sigma_e = \frac{\sigma_\varepsilon}{\sqrt{1-\rho^2}}. \quad (3)$$

The rest of the time-series follow a linear procedure with lag-1 autocorrelated errors (e.g. Harrop & Velicer, 1985; Velicer & Molenaar, 2013). The linear regression model without the slope parameter and the serial dependency of the residual (e_t) can be expressed respectively as,

$$\hat{y}_{pt} = \beta_{0p} \text{ and} \quad (4)$$

$$e_{pt} = \rho e_{pt-1} + \varepsilon. \quad (5)$$

In Equation 4, $\hat{y}_{pt}$ is the predicted value of the target behavior at time *t* in phase *p*; β_{0p} is the intercept of the linear regression model for phase *p*; e_{pt} is the error at time *t* for phase *p*; ρ is the autocorrelation coefficient; and ε is the independently distributed error. Let the time points in the baseline phase be denoted as $1, 2, \ldots, t_b$ and in the treatment phase as $t_{b+1}, \ldots, t_n$. Then the intercept β_{0p} can be modeled as:

$$\beta_{0p} = \begin{cases} \beta_{01}, \text{if } t \leq t_b \\ \beta_{02}, \text{otherwise} \end{cases}. \quad (6)$$

Immediacy is indicated when the mode of the posterior distribution of the change-point t_b is estimated to be the same as the design change-point coupled with small posterior standard deviation. This will be demonstrated in the forthcoming sections. The effect size is the standardized mean difference of the intercept estimates in the two phases under consideration.

An important aspect of Bayesian estimation is the use of priors to estimate the posteriors. In the BUCP program relatively uninformative priors are used so as to remain agnostic about our

beliefs about the posterior estimates (Natesan Batley & Hedges, 2020). For instance, β is sampled from a normal distribution with mean drawn from another normal distribution with mean 0 and precision .0001 which corresponds to a standard deviation of 100. The precision for the β value is sampled from a gamma distribution with shape and rate of 1 each. The autocorrelation is drawn from a uniform distribution ranging from -1 to 1, the plausible values for autocorrelation. The change-point can take on any discrete value from 3 to – 3 because at least three observations are needed per phase according to WWC standards and to discern a statistical pattern. In sum, all the priors considered are relatively uninformative, however, it is strongly recommended that more informative priors be used by researchers based on information about the study and from previous research (see Gill, 2015; Lambert, Sutton, Burton, Abrams, & Jones, 2005 for some general examples, and Natesan, 2019 for a specific SCED example). This is because, for small samples prior distribution plays a large influence on the resulting estimates, especially for the scale parameter. There is a great variation in the specification of relatively uninformative priors and their use could lead to different inferences (Lambert, et. al, 2005). Readers may download the BUCP program for implementing this analysis and producing the plots in the present study from github (https://github.com/prathiba-stat/BUCP). Runjags (Denwood, 2013) is a software package that runs using R (2016), both of which will be required for running the BUCP program.

In some SCEDs, delayed effects (i.e., latency) may be expected; that is, the effectiveness of the intervention is observed at a later time point after the introduction of the intervention. For instance, a drug or a chemotherapy treatment may take time to take effect or a child with autism may take time to learn how to use an iPad as a medium of communication. In such cases, it is necessary to acknowledge that the design change-point is different from the actual change-point,

that is, the point when the intervention begins to take effect. The BUCP model can systematically model a delayed effect and compute the correct effect size. This is an important distinction from traditional analysis where the delayed effect is ignored, and therefore the ensuing computation of the effect size estimate is inaccurate. Thus, by allowing the data to speak for themselves, the BUCP methodology can be used to test immediacy, latency, effect sizes, and testing for region of practical equivalence (ROPE; a Bayesian equivalent of statistical significance testing), all in a single analysis.

Bayesian statistical significance in BUCP

The 95% credible interval (CI) of the posterior distribution of standardized mean difference determines the limits for 95% of the credible values for the effect size under this distribution. The rule of thumb generally used in between-subject designs (i.e., 0.2, 0.5, 0.8 indicate small, medium, and large effects, respectively) cannot be used to interpret SCED effect sizes. This is because it is not uncommon to find standardized mean difference-type effect sizes such as 3 or higher in SCEDs (Shadish & Sullivan, 2011) and an effect size of 1 does not necessarily indicate a clinically significant effect. In this study, an effect size is tentatively considered to be clinically significant if it is three or higher. That is, the lower bound of the 95% CI of the posterior effect size distribution should be three or larger. However, this value was chosen for illustrative purposes only. Researchers are encouraged to choose values more appropriate for their research. Moreover, only in the Bayesian framework the null or the research hypothesis can be ‘accepted’ as opposed to being ‘not rejected’ as in the classical framework (Kruschke, 2013).

Simulation Modeling Analysis (SMA)

The SMA technique (Borckardt, Nash, Murphy, Moore, & O'Neal, 2008; Borckardt, 2006) was developed for analyzing particularly short streams ($n < 30$ per phase) of single-case time-series design data. SMA answers the question similar to that asked in a traditional NHST context: if there is no functional relationship between the independent and the outcome variable, what is the probability one would observe the relationship at least as large as is observed with observed data? Therefore, small *p*-values (e.g. $p < .05$) indicate phase effect. This program simulates several thousands of random data that have the same phase *n*-sizes and the same amount of autocorrelation as the observed data. Results from observed data are then compared to the distribution for the simulated random data to determine if the observed correlation is due to chance. The percentage of times the correlations of the simulated datasets are larger than the correlation from the observed data is an estimator for the *p*-value. Results of SMA include the estimates of autocorrelation, the mean and standard deviation values for the two phases, the *p*-value associated with the level change or phase effect, and an effect size (i.e., Pearson's *r*). The program also tests the data for different standard slope change models. Users can modify the program available on http://www.clinicalresearcher.org/software.htm for other types of SCEDs.

The SMA program is a freeware and easy-to-use tool for SCED researchers. However, it is not without disadvantages. Although SMA is a sound statistical procedure, it assumes that the estimated parameters used to generate data streams are a reasonable representation of the data characteristics. Secondly, SMA ignores autocorrelation when computing Pearson's correlation effect size. Therefore, it underestimates the effect size. Thirdly, most SCED data are count data or ratio data. It is unclear how SMA functions for such data (Borckardt & Nash, 2014). Fourthly, SMA does not function well for large data streams, although SCED data time-series are rarely longer than 10 data points per phase. Fifthly, it facilitates researchers to test several hypotheses

of slope differences. Testing multiple hypotheses leads to an increase in experiment-wise type-I error rate. Moreover, a researcher may be tempted to simply test each hypothesis at the traditionally used .05 threshold value and report only those they find statistically significant. Finally, the program cannot estimate delayed effects. The focus in SMA is to measure treatment effect assuming there is a clear immediacy and not test all aspects of intervention effect as prescribed by the WWC for SCEDs (e.g., immediacy, appropriate effect size). In sum, although SMA is very useful, it is not a one-stop-shop for complete SCED analysis.

Data

We selected three datasets from recent SCED literature that exhibited "clear" immediacy, "unclear" immediacy, or delayed effects (latency) to demonstrate the difference between the methods. Clear immediacy is exhibited if the data patterns in the two phases clearly supported the intervention change-point as the design change-point. Unclear immediacy is exhibited when the patterns within the phases show some inconsistency so that a researcher cannot clearly point out when the actual change began to take place. Delayed effects are exhibited when the actual change occurs at least one time point after the intervention was implemented. All of these grades of immediacy were determined by visual inspection. The team of four authors independently categorized the selected datasets as belonging to one of the immediacy types. There was 100% agreement in the categorization. The graphs were digitized using the digitizing software WebPlotDigitizer 3.11 (Rohatgi, 2017). Data that were coded manually were compared to the data coded by WebPlotDigitizer for each dataset to ensure the values were identical.

Datasets Characteristics

Dataset 1. This dataset was obtained from a multiple baseline design (MBD) by Coulter and Lambert (2015) and classified as showing clear immediacy. The dependent variable was the

percent of correct words read per minute in a preselected 150-200-word passage by participants with a learning disability and reading two grade levels below their same age peers. The researchers used visual analysis. Autocorrelation was not calculated; however the researchers reported an overall (inclusive of three participants) percentage of all non-overlapping data (PAND) effect size = 97.91% and 90% CI = [0.94, 0.99]. In addition to PAND, Coulter and Lambert (2015) reported a Pearson's *phi* value of .915, 90% CI = [.84, .98]. BUCP and SMA estimates for all 3 participants in the MBD were also computed as an extension of the first case to MBDs.

Dataset 2. Dataset 2 was obtained from one of the subjects of a MBD by Macpherson, Charlop, and Miltenberger (2015) who examined the effects of a portable video modeling intervention (using iPad®) on the verbal compliments and compliment gestures of children with autism. Dataset 2 contains the number of verbal compliments given by one participant in the baseline and intervention phases. The authors of the study used a one-tailed Wilcoxon signed rank test to conclude the presence of a statistically significant difference between the observations in the baseline and intervention phases. Effect sizes and autocorrelation were not reported. The data show a possible *delayed* effect.

Dataset 3. The third dataset was taken from Barber, Saffo, Gilpin, Craft, and Goldstein (2016) because it showed *unclear* immediacy. This study examined the efficacy of peer-mediated interventions (i.e., stay, play, and talk strategies) on the social communication skills of preschool children with autism. Barber et al. (2016) used SMA. A visual plot for the data suggests that it is unclear when the treatment effect started and whether there was a statistically significant treatment effect. SMA results showed no statistically significant level change at $\alpha =$

0.05 level (Pearson's $r = 0.729$; $p = 0.06$) but a statistically significant slope change ($r = 0.867$, $p = 0.02$). Autocorrelation was not reported.

Results

The results for SMA and BUCP analyses for all datasets are shown in Table 1. Figures 1-4 display the line charts and posterior plots for all data sets. Each figure has two parts: Part *a* displays the line chart of the data and Part *b* displays the posterior plots. For Datasets 2 and 3, for the sake of brevity, only posterior plots for phase means and the effect size are included.

TABLE 1 AND FIGURE 1 ABOUT HERE

Clear Immediacy

Dataset 1. Multiple Baseline Design. The results will be described in detail for George, followed by John and Mark. The vertical dotted line in Figure 1a separates the data for the child George from the two phases. There are 13 points in Phase A and 20 points in Phase B. Visual observation of Figure 1a shows clear immediacy at the end of Phase A at time point 13. It can be seen from Table 1 that the observed means and standard deviations (SD) for Phases A and B are 34.62 (8.18) and 61.0 (13.27), respectively. The SMA analysis showed that there is a statistically significant phase effect ($p < .01$). That is, the means differ significantly between Phase A and Phase B. Figure 1b shows Bayesian results for the same data: posterior distribution plots of means for Phase A and Phase B, effect size, change-point, and autocorrelation. The posterior plot for the change-point shows the mode at 13, which is in agreement with a visual inspection of Figure 1a. Therefore, one can conclude that BUCP correctly estimated the change-point. Posterior distributions of means of Phases A and B are non-overlapping and narrow distributions. As can be seen from Table 1, means and 95% CIs for Phase A and Phase B posterior distributions are: 33.93 [30.51, 37.28], and 61.21 [58.30, 64.13], respectively. These estimated

means are very close to the observed means, confirming our confidence in the estimation of the posterior distributions. The posterior distribution of the effect size has a mean of 5.5 with 95% CI [4.60, 6.38]. Suppose the researcher hypothesizes that an effect size greater than 3 shows a statistically significant treatment effect. Then the lower bound value of 4.6 of the 95% CI is much greater than the acceptable value of 3, indicating a large effect size and that 97.5% of credible values for the effect size are above a value of 4.6. Therefore, there is sufficient evidence to show that this effect size is statistically significant and the research hypothesis that the treatment effect is larger than 3 can be accepted.

INSERT FIGURE 2 ABOUT HERE

Figure 2a shows data plots for all three children, Mark, George, and John in Dataset 1, who are part of the MBD. Figure 2b shows their respective posterior distributions for the change-point. Visual observation data for Mark shows a clear immediacy at the end of Phase A at time point 10. It can be seen from Figure 2b that the BUCP modal estimate of the change point was also 10. This shows support for immediacy. The means for Mark in Table 1 show that there was a considerable increase in the percent correct score in phase 2 (46.1% to 71.5%). The 95% credible intervals for the effect size [4.41, 5.96] do not contain zero and are above 3, indicating statistically significant improvement. These results show that, for Mark, because of the intervention, the percent correct score in Phase 2 has increased, on an average, by more than five standard deviation units (mean of the effect size posterior distribution, 5.2).

BUCP estimate of change-point for the third child, John was 19, that is, the time-point immediately after the intervention started. Thus, again, there is support for immediacy. There was a considerable increase in the percent correct score in phase 2 (47.3% to 83.3%). Credible intervals for the effect size do not contain zero, indicating improvement. Furthermore, the lower

bound of the 95% credible intervals for the effect size is 6.42, much higher than the acceptable value of 3 for statistically significant difference. These results indicate that, for John, because of the intervention, the percent correct score in phase 2 has increased, on an average, by more than six SD units (mean of the effect size posterior distribution, 7.3).

Delayed Effect

Dataset 2. Visual observation of Figure 3a reveals a delayed effect, with the treatment effect appearing around time point 10 rather than at 8 when the intervention started. The observed means and standard deviations for Phase A and Phase B, in Table 1, are 4.92 (8.55) and 39.73 (43.13), respectively. The SMA results which presume 8 as the change-point indicate a statistically non-significant phase effect ($p > .19$). The BUCP analysis estimated the change-point at 11, correctly indicating a delayed effect. Note that this posterior could not be plotted in figure 3b because there is no variation in the change-point value in any of the iterations. Estimated means and 95% CIs for Phases A and B are 5.38 [2.37, 8.32] and 56.22 [52.27, 60.39], respectively. These means are very different from the observed means computed by considering 8 as the change-point. The posterior plot of the effect size is shown in Figure 3b with its mean at 10.19 and 95% CI [9.21, 11.22], indicating a large effect with 97.5% of credible values for the effect size above 9.21.

INSERT FIGURE 3 ABOUT HERE

Suppose this delayed effect were expected as a function of the type of intervention. In such a case, because of the delayed effect, SMA has erroneously concluded that there is no treatment effect. On the other hand, the BUCP analysis, by correctly detecting the delayed effect and correctly estimating the change-point, showed a statistically significant treatment effect with a

large effect size. These results show how misleading the results can be when immediacy is assumed to happen at the start of the treatment.

Unclear Immediacy

Dataset 3. Visual observation of Figure 4a indicates that there is immediacy; however, it is not clear when. In other words, there is a delayed effect, but unlike Figures 1a or 3a, it is not clearly distinguishable visually when the change-point occurred. The observed means and SDs for Phase A and Phase B as shown in Table 1 are 3.14 (3.14) and 14.44 (5.48), respectively. The SMA results, which assume the time point 7 as the change-point showed a statistically significant result ($p < .05$). However, the BUCP analysis estimated the change-point to occur much later than when the intervention was implemented, that is, at 16 resulting in the estimated means and 95% credibility intervals as 4.31 [-1.35, 8.89] and 14.93 [1.92, 24.48], for Phases A and B, respectively. Of course, this throws some questions about the reliability and validity of the data. Specifically, are the changes in the data due to an intervention or just random fluctuations? If the researcher had sufficient information to support that the change in the data pattern is due to the intervention effect and there is a reason for the change-point to be occurring at the estimated change-point, he/she could continue with computing the effect size. The effect size posterior mean is 2.82 with 95% CI [0.14, 5.37]. The lower bound value of 0.14 for the CI is certainly below the cut-off value of 3, and more than 50% of the credible values are below the cut score of 3, clearly indicating a statistically non-significant result. Therefore, the researcher would 'accept' the null that there is no statistically significant treatment effect according to the criterion set in this study.

INSERT FIGURE 4 ABOUT HERE

This data set also illustrates the importance of accurately estimating the change-point. For Datasets 2 and 3, SMA results presuming the change-point to be at the time of intervention reached an erroneous conclusion while the BUCP analysis detected the delayed effect and appropriately estimated the change-point and subsequently, the posterior means. Of course, what the BUCP cannot determine is whether this delay is expected due to the nature of the intervention. This can only be determined substantively.

Discussion and Conclusions

This study explained and discussed BUCP, a recently developed statistical methodology, for measuring effect sizes that account for autocorrelations and do not require small sample corrections for SCEDs. BUCP also investigates and quantifies immediacy in SCED studies. This study compared and contrasted the performance of BUCP with SMA, two simulation-driven approaches for analyzing SCED data. A limitation of the BUCP effect size is that it is a within subject effect size, that is, the variance used to compute it comes from the measurements of a single subject. This means that the BUCP effect size in the current study cannot be aggregated across studies in a meta-analytic context.

Determining immediacy is an important aspect of establishing intervention effect in SCEDs. However, until now, there have been no criteria as to what a meaningful difference between means/medians is to establish immediacy. Given that this difference is computed only for 3-5 data-points in each phase, testing statistical significance in the classical framework is unreasonable. This difference depends on the range of the outcome variable as well. For instance, problem behaviors may range from zero to 20 in a given time-period while performing computer mouse operation may range from zero to a few hundred. Thus, an immediacy value of 12 may indicate strong immediacy while an immediacy value of 85 may indicate weak

immediacy, depending on the scale of the outcome variable. The BUCP model, on the other hand, is sensitive to patterns that show weak immediacy compared to those that show strong immediacy. This sensitivity is indicated in the shape of the posterior distribution of the change-point. If there is a clear single mode in the posterior and this mode aligns with the time point when the intervention was implemented, there is evidence to support immediacy (Natesan & Hedges, 2017).

Although visual analysts study latency, it is unclear how statisticians would deal with such delayed effects. There is also little guidance on this in existing standards. In fact, there may be no one-size-fits-all decision when it comes to delayed effect being a threat to intervention effect. When delayed effect is not considered, the effect sizes are underestimated. Therefore, examining immediacy in an objective manner is important in SCEDs. The BUCP methodology is a useful technique in this regard. It considers the entire data pattern in the two phases and estimates the change-point.

Only in the Bayesian framework the null or the research hypothesis can be ‘accepted’ as opposed to being ‘not rejected’ as in the classical framework (Kruschke, 2013). Since the publication of SMA in 2008, several standards for establishing intervention effect in SCEDs have been published (e.g. Cook et al., 2014; Kratochwill et al., 2013). Unlike BUCP, SMA does not model two aspects of these standards, effect sizes that account for autocorrelation and immediacy of intervention effect. For example, when there was clear immediacy, the means of the two methods were comparable. However, when there was delayed effect, only BUCP was able to identify and incorporate this delayed effect in its effect size computation. Hence, SMA is a good technique to analyze SCED data where there is clear immediacy, but falls short and provides inaccurate information about the effectiveness of an intervention when there is delayed

immediacy. The BUCP analysis, on the other hand, is an effective tool in estimating the effectiveness of an intervention even in cases of delayed immediacy. In this sense, the BUCP analysis can also serve as a diagnostic tool.

Because the BUCP methodology is a Bayesian technique, it reveals a wealth of information about the possibilities of statistics of the parameters in the form of posterior distributions. In traditional analysis, mean or median are considered acceptable values to evaluate the outcome variable. However, posterior distributions are obtained based on repeated Monte Carlo simulations of a combination of the prior and the likelihood (data). Depending upon the shape of this distribution and the contextual information such as the sample size, one can examine the mean or the median of this distribution, in combination with credibility intervals of the desired length. The present study also demonstrated how regions of practical equivalence could be built around a hypothesized value to test statistical significance in the Bayesian framework. In addition, BUCP offers more modeling flexibility over SMA by being able to incorporate the scale of the data.

This study has illustrated and highlighted the strengths of the performance of BUCP compared to SMA on a limited set of data. However, a more extensive simulation study that compares the performance of the two models in a more systematic manner would provide a more complete comparison of the two approaches. The BUCP model is not without its drawbacks. Apart from the long computing time (about 47 seconds on average per analysis) and the learning curve, even though BUCP is highly sensitive to data patterns, a unimodal clear change-point estimate is not conclusive evidence of immediacy. Therefore, visual inspection of the data must always accompany interpreting statistical estimates in SCED data analysis. The two aspects of

data analysis, visual and statistical, together can evaluate the causal validity of SCED findings via transparent, objective and replicable procedures.

References

Allen, M.B., Baker, J.C., Nuernberger, J.E., & Vargo, K.K. (2013). Precursor manic behavior in the assessment and treatment of episodic problem behavior for a woman with a dual diagnosis. *Journal of Applied Behavior Analysis, 46*, 685-688. https://doi.org/10.1002/jaba.57

American Speech-Language-Hearing Association. (2004). *Evidence-Based Practice in Communication Disorders: An Introduction* [Technical Report]. Available from http://shar.es/11yOzJ or http://www.asha.org/policy/TR2004-00001/.

Barber, A. B., Saffo, R. W., Gilpin, A. T., Craft, L. D., & Goldstein, H. (2016). Peers as clinicians: Examining the impact of Stay Play Talk on social communication in young preschoolers with autism. *Journal of communication disorders*, *59*, 1-15.

Borckardt, J. J. (2006). *Simulation modeling analysis: Time series analysis program for short time series data streams* (Version 8.3.3). Charleston, SC: Medical University of South Carolina.

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008).

Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. *American Psychologist*, *63*(2), 77-95. doi: 10.1037/0003-066X.63.2.77

Borckardt, J. J., & Nash, M. R. (2014). Simulation modelling analysis for small sets of single subject data collected over time. *Neuropsychological Rehabilitation*, *24*(3-4), 492-506. doi: 10.1080/09602011.2014.895390

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.

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

Coulter, G. A., & Lambert, M. C. (2015). Access to general education curriculum: The effect of preteaching key words upon fluency and accuracy in expository text. *Learning Disability Quarterly*, *38*(4), 248-256.

Council for Exceptional Children. (2014). *Council for Exceptional Children Standards for Evidence-Based Practices in Special Education*. Arlington, VA: Council for Exceptional Children. http://www.cec.sped.org/~/media/Files/Standards/Evidence%20based%20Practices%20and%20Practice/EBP%20FINAL.pdf

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

Gill, J. (2015). *Bayesian Methods: A social and behavioral sciences approach* (3rd ed.). Boca Raton, FL: CRC Press.

Guyatt, G., Sackett, D., Taylor, D.W., Chong, J., Roberts, R., & Pugsley, S. (1986). Determining optimal therapy – randomized trials in individual patients. *New England Journal of Medicine, 314*, 889–892.

Harrop, J. W., & Velicer, W. F. (1985). A comparison of three alternative methods of time series model identification. *Multivariate Behavioral Research, 20*, 27-44.

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.

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.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7*, 107–118.

Jones, R. R., Weinrott, M., & Vaughn, R. S. (1978) Time-series analysis in operant research. *Journal of Applied Behavior Analysis, 10,* 151-166.

Kazdin, A. E. (2011). *Single-Case Research Designs* (2nd ed.). NY: Oxford University Press.

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.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General, 142*, 573-603.

Lambert, Sutton, Burton, Abrams, & Jones (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, *24*: 2401–2428.

Macpherson, K., Charlop, M. H., & Miltenberger, C. A. (2015). Using portable video modeling technology to increase the compliment behaviors of children with autism during athletic group play. *Journal of autism and developmental disorders*, *45*(12), 3836-3845.

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.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2013). Three-level analysis of single-case experimental data: Empirical validation. *The Journal of Experimental Education*, *82*(1), 1-21.doi: 10.1080/00220973.2012.745470

Natesan Batley, P. (2020a). Use of Bayesian Estimation in the Context of Fully Integrated Mixed Methods Models. In J. Hitchcock and A. J. Onwuegbuzie (Eds.) *Routledge Handbook for Advancing Integration in Mixed Methods Research*. Routledge.

Natesan Batley, P. (2020b). Bayesian Analyses with Qualitative Data. In A. J. Onwuegbuzie and B. Johnson (Eds.) *The reviewer's guide to mixed methods research analysis.* Routledge.

Natesan Batley, P., Contractor, A. A., & Caldas, S. (2020). Bayesian Time-Series Models in Single Case Experimental Designs: A Tutorial for Trauma Researchers. *Journal of Traumatic Stress, 33*, 1144-1153. DOI: https://doi.org/10.1002/jts.22614.

Natesan Batley, P., Shukla Mehta, S. & Hitchcock, J. (2020). Integrating Visual and Statistical Analyses in Single Case Experimental Research using Bayesian Unknown Change-Point Models. *Behavioral Disorders.* DOI: 10.1177/0198742920930704.

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. & 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., Minka, T., & Hedges, L. V. (2020). Investigating immediacy in multiple phase-change single case experimental designs using a Bayesian unknown change-points model. *Behavior Research Methods, 52,* 1714-1728. DOI: https://doi.org/10.3758/s13428-020-01345-z

Natesan Batley, P. & Hedges, L. V. (2020). Accurate model vs. accurate estimates: A study of Bayesian single-case experimental designs. *Behavior Research Methods*.

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

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.

R development core team. (2016). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, http://www.R-project.org.

Rindskopf, D. (2014). Nonlinear Bayesian analysis for single case designs. *Journal of School Psychology, 52*, 179-189.

Rohatgi, A. (2017). WebPlotDigitizer. Austin, TX. http://arohatgi.info/WebPlotDigitizer.

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.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology, 52*, 149–178.

Thyer, B. A. & Myers, L. L. (2011). The quest for evidence-based practice: A view from the United States. *Journal of Social Work, 11*, 8-25.

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

Table 1

Description of Datasets and Results from SMA and BUCP

Data Description						SMA Results		BUCP Results			
Dataset		n_A	n_B	$Mean_A$ (SD)	$Mean_B$ (SD)	*r**	*p*-value	Change Point	$Mean_A$ [95% HDI]	$Mean_B$ [95% HDI]	ES Mean [95% HDI]
Dataset 1 (MBD)											
	George	13	20	34.62 (8.18)	61 (13.27)	0.75	0.004	13	33.93 [30.51, 37.28]	61.21 [58.30, 64.13]	5.5 [4.60, 6.38]
	Mark	10	20	46.1 (4.65)	71.1 (8.52)	0.85	0	10	46.09 [43.05, 49.15]	71.48 [69.4, 73.48]	5.2 [4.41,5.96]
	John	19	13	47.8 (8.08)	83.7 (10.59)	0.89	0.0012	19	47.27 [44.49,49.93]	83.28 [79.98, 86.63]	7.3 [6.42, 8.18]
Dataset 2		8	9	4.92 (8.55)	39.73 (43.13)	0.48	0.192	11	5.38 [2.37, 8.32]	56.22 [52.27, 60.39]	10.19 [9.21, 11.22]
Dataset 3		7	16	3.14 (3.14)	14.44 (5.48)	0.73	0.049	16	4.31 [-1.35, 8.89]	14.93 [1.92, 24.48]	2.82 [0.14, 5.37]

Note. n = number of data points in phases A and B; HDI = high density interval or credibility interval; ES = effect size that accounts for autocorrelation; *p* = mean difference *p*-value. **r* = correlation between the DV and the phase (does not account for autocorrelation)

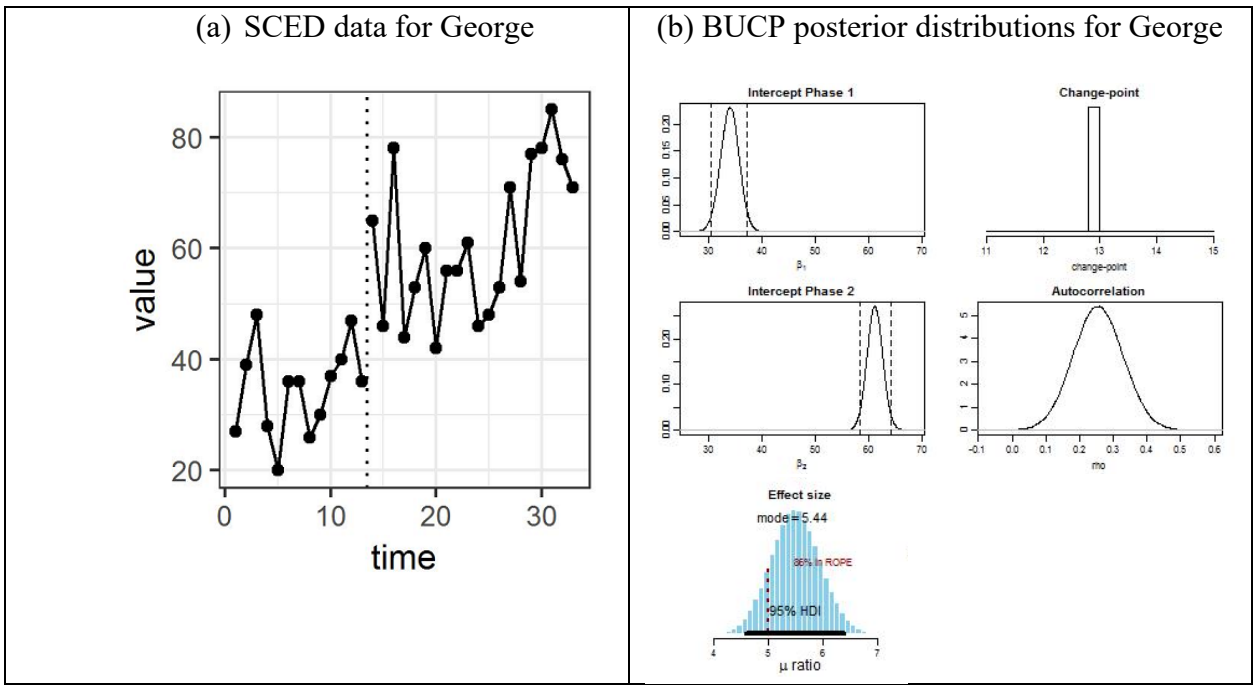
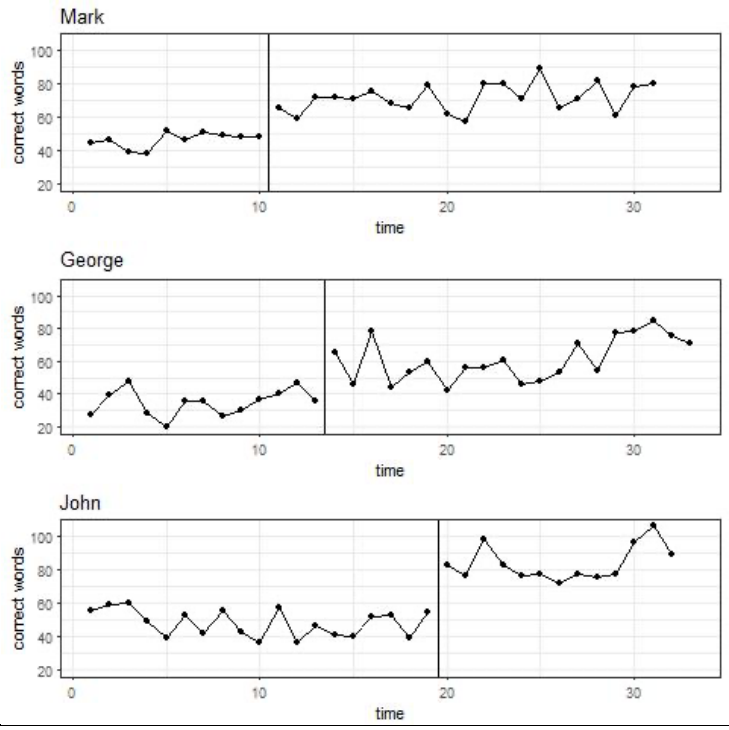


Figure 1a-b. SCED raw data for George in Dataset 1 and posterior distributions obtained with BUCP.

(a) SCED data for dataset 1 MBD



(b) BUCP posterior distributions for dataset 1 MBD

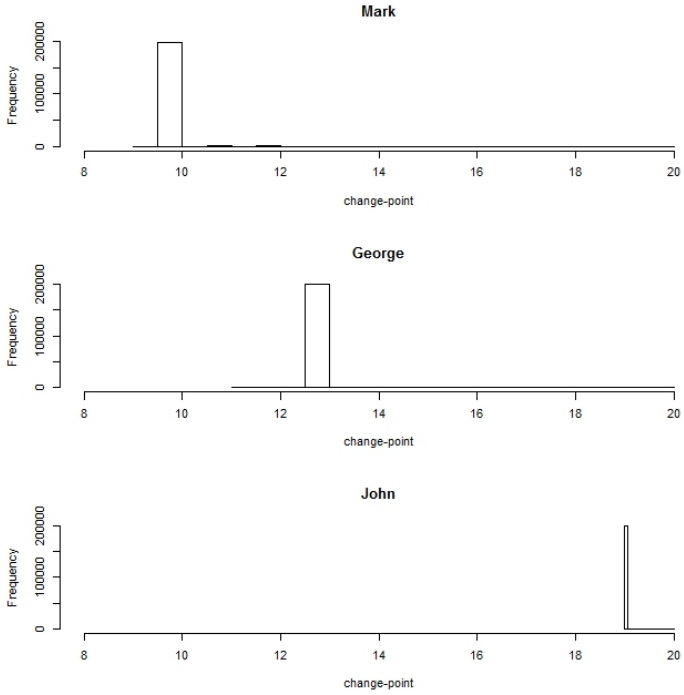


Figure 2a-b. SCED raw data for MBD dataset 1 paired with the posterior distribution for change-points obtained with BUCP.

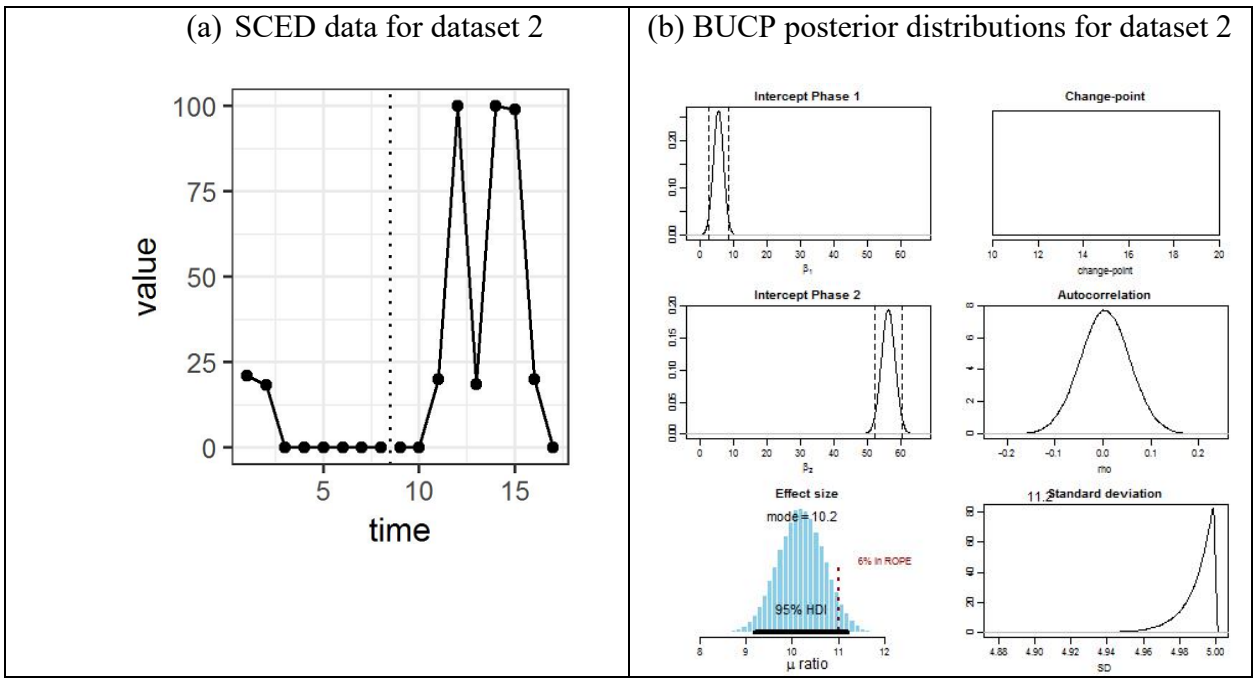


Figure 3a-b. SCED raw data for dataset 2 paired with the posterior distributions obtained with BUCP.

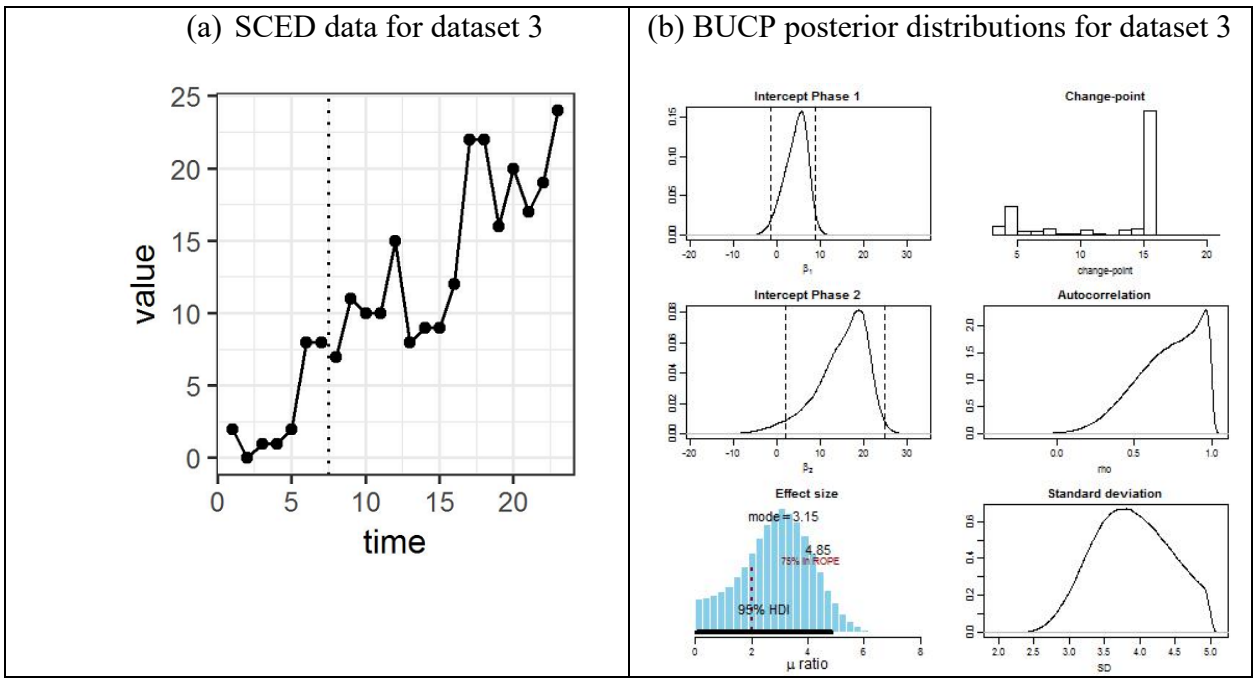


Figure 4a-b. SCED raw data for dataset 3 paired with the posterior distributions obtained with BUCP.