Bayesian Unknown Change-Point Models to Investigate Immediacy in Single Case Designs

Prathiba Natesan
University of North Texas

Larry V. Hedges
Northwestern University

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant **R305D170041** to Northwestern University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Accepted for publication in Psychological Methods 2016

Abstract

Although immediacy is one of the necessary criteria to show strong evidence of a causal relation in SCDs, no inferential statistical tool is currently used to demonstrate it. We propose a Bayesian unknown change-point model to investigate and quantify immediacy in SCD analysis. Unlike visual analysis that considers only 3-5 observations in consecutive phases to investigate immediacy, this model considers all data points. Immediacy is indicated when the posterior distribution of the unknown change-point is narrow around the true value of the change-point. This model can accommodate delayed effects. Monte Carlo simulation for a two-phase design shows that the posterior standard deviations of the change-points decrease with increase in standardized mean difference between phases and decrease in test length. This method is illustrated with real data.

Keywords: Bayesian estimation; single case designs; n-of-1 designs; Markov Chain Monte Carlo

Bayesian Unknown Change-Point Models to Investigate Immediacy in Single Case Designs

Single case designs (SCDs) are widely used to test the effects of interventions/treatments in education (e.g. Lambert, Cartledge, Hewrad, & Lo, 2006), psychology (e.g. Shih, Chang, Wang, & Tseng, 2014), and medicine (as *n-of-1* designs, Gabler, Duan, Vohra, & Kravitz, 2011). SCDs involve the repeated assessment of an outcome over time (i.e., a time series) within a case (which could be a child, a classroom, etc.), during one or more baseline phases and one or more treatment phases, where the experimenter controls the timing of the phases (Horner et al. 2005; Kratochwill & Levin, 2014). Thus, SCDs are a form of interrupted time series design. In many areas such as in research on treatments for low incidence disabilities (e.g. Autism spectrum disorders, moderate intellectual disability, schizoaffective disorder), it is difficult to assemble a substantial number of research subjects or implement a one-size-fits-all intervention. In these areas SCDs often provide a substantial part of the research evidence (e.g. Allen, Baker, Nuernberger, & Vargo, 2013; Lin & Chang, 2014; Neely, Rispoli, Camargo, Davis, & Boles, 2013; Shih, C.-H., Chang, Wang, & Tseng, 2014; Shih, C. -H., Chiang, & Shih, C. T., 2015; Shih, C.-H., Wang, Chang, & Kung, 2012). Decades of research experience has led to the development of considerable professional consensus on the methodological standards for SCDs. One example is the U.S. Department of Education's What Works Clearinghouse Pilot Standards for single-case designs (Kratochwill et al. 2013). Other examples include standards adopted by the American Speech-Language-Hearing Association (2004) and the Council for Exceptional Children (CEC) (Cook et al., 2014).

Because variation over time is a central feature of SCDs, analyses of SCDs have typically involved visual analysis of a plot of observations over time. Such plots are the standard format for reporting data and supporting data analysis in SCDs. Visual analysis has focused on

establishing that there is a stable pattern or functional relation among the observations in each phase and that there is a difference in the pattern of observations in baseline and treatment phases. The baseline phase is the phase where no intervention or treatment is administered. Observations are generally taken during this phase to establish a pattern of trend and stability before the treatment is administered. This phase is indicated as phase A. The treatment phase is the phase where the treatment is administered and is indicated as phase B. Observations in the treatment phase are expected to be a function of the treatment effect on the dependent variable. The difference between the patterns in treatment and baseline phases is evidence of a treatment effect. Often the differences in the patterns in baseline and treatment phases are striking, making treatment effects easy to identify. However this need not always be the case. When treatment effects are not striking, visual methods may lead to more ambiguous results.

There has recently been increased interest in developing statistical methods for analyzing single case designs to provide additional analytic tools to supplement visual analysis for SCDs (e.g. Hedges, Pustejovsky, & Shadish, 2012, 2013; Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2013; Shadish, et al, 2015; Shadish, Zuur, & Sullivan, 2014). These developments include applications of multilevel modeling (e.g., Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2013), semiparametric regression models (e.g., Shadish, Zuur, & Sullivan, 2014), and fully Bayesian analysis approaches (e.g., Rindskopf, 2014). There have also been attempts to develop methods for estimating effect sizes from SCDs that would be comparable to the effect sizes estimated from more conventional between-subjects designs (e.g., Hedges, Pustejovsky, & Shadish, 2012, 2013).

According to the WWC (Kratochwill et al., 2013), SCD analyses need to show three demonstrations of intervention effect along with no non-effects by (a) documenting the

consistency of level, trend, and variability within phases, (b) documenting immediacy, proportion of overlap of observations across phases, and comparing the observed and predicted patterns of the observations, and (c) examining anomalies and external factors. All the aforementioned statistical approaches focus on answering questions pertaining to modeling SCD data to measure differences in trends and levels, and measures to quantify these differences. These measures satisfy one criterion for demonstrating strong evidence of causality in SCDs as prescribed by Kratochwill et al. (2013). However, no inferential statistical analytic tool has focused on demonstrating immediacy which is another important criterion for demonstrating strong evidence of causality in SCDs.

We propose the Bayesian unknown change-point model to estimate and quantify immediacy in SCDs. To that end, the present study investigates the performance of Bayesian unknown change-point models in estimating and quantifying immediacy in SCDs with two phases. The simulation section of the study investigates the feasibility of the model in commonly occurring data conditions in SCDs including phase lengths, autocorrelations, and standardized mean difference between phases. The performance of the model informs us about its applicability to multiple baseline design data. The illustration section of the study applies the model to two real datasets to demonstrate the utility of the model. Bayesian inference of various parameters including constructing the region of practical equivalence (ROPE) is demonstrated. A sample code is provided in the appendix to facilitate the reader to use this model.

We only consider a biphasic (two-phase) AB design in this study which is quasi-experimental and is not adequate to show evidence of causality by itself. This is because the AB design shows only one demonstration of intervention effect. The multiple baseline design (MBD) where the timing of the intervention is staggered across multiple participants or the ABAB

design which contains 4 phases are the commonly used designs that can provide strong evidence of causality. Nonetheless, this is the first step in investigating the feasibility of unknown change-point models to investigate and quantify immediacy. This bi-phasic investigation avoids the effects of external factors on the accuracy of the estimates such as between person differences and the accuracy of previous change-points affecting the accuracy of subsequent change-points which may be expected in MBD and ABAB designs, respectively. Moreover, all SCDs are based on phase changes making the present investigation a necessary first step before extending to other complex designs.

Immediacy

Until now immediacy has been established through visual analysis and computing the change in the mean or median levels between the last three-to-five data points from phase 1 and the first three-to-five data points in phase 2 (Horner, Swaminathan, Sugai, & Smolkowski, 2012). The importance of immediacy depends on whether the researcher expects a gradual or a rapid shift in the dependent variable following the treatment. A delayed effect is generally considered a threat to internal validity because there is less convincing evidence that the change in the dependent variable was due to the manipulation of the independent variable. An inferential statistical tool that can demonstrate immediacy or delayed effect can be a valuable addition to the SCD researchers' toolkit because it can add evidence to support internal validity. While current methods use only 3-5 data points per phase to establish immediacy, the Bayesian unknown change-point model uses all data points to establish immediacy.

Currently used statistical approaches assume that the change in the dependent variable takes place immediately where it was intended. But this may not necessarily be realistic in cases where there is a delayed effect. A delayed effect happens when there is no change in the

observations immediately following a phase change but rather a change after a time period. A washout effect is a special case of delayed effect when there is no change in the observations immediately following withdrawal of an intervention but rather a change after a time period. For instance, consider a treatment that takes an unknown time to create a change in the dependent variable (e.g. a drug that takes time to be absorbed into the body) and a design with a baseline phase followed by the treatment phase, such as an AB design. This is an example of a delayed effect. Until the treatment begins to have an effect, the first few observations in the treatment phase will not reflect the trajectory that occurs under treatment. Instead, these first few observations still reflect the baseline trajectory. Evaluating the treatment effect or computing effect sizes by considering these observations as being part of the trajectory produced by the treatment will be erroneous. A similar argument could be made about the first few baseline observations after a treatment phase (for example in an ABA design) if the treatment takes some time to "wear off" (such as a drug that needs time to be excreted from the body). This is an example of a washout effect. Here the first few baseline observations may still reflect the treatment trajectory, not the baseline trajectory. The exact delay times may not be known in advance and may vary across cases. As Duan, Kravitz, and Schmid (2013) noted, successful n-of-1 trials must ascertain gradual and/or delayed effects or account for these with appropriate analytic strategies to untangle these from long-term treatment effect. To date, no inferential statistical method has been used to investigate or quantify immediacy which is a necessary criterion for strong evidence of causality in SCDs. Reporting inferential statistical evidence of immediacy can add to evidence of internal validity in SCDs.

Let us consider the middle two phases of an ABAB data from Neely, Rispoli, Camargo, Davis, and Boles (2013) that is shown in Figure 1. The Bayesian unknown change-point model

estimated the change-point to be at 8 in each of the 200,000 iterations after burning in the first 10,000 iterations. The visual plot shows a possible delayed, effect from phase B1 to A2 and the mean difference between the three last and first points in phases 1 and 2, respectively is 27.67. If there were a delayed effect, the standardized effect size, d would be 6.1. If a researcher ignores this possibility, the effect size would be 5.83 – an underestimation of the effect.

INSERT FIGURE 1 ABOUT HERE

The problem with ignoring the possibility of delayed effects is multi-fold. First, the validity of the data is threatened because the data no longer represent what they are defined to represent. Second, the effect size is misestimated. Third, this inaccurate estimation will in-turn affect meta-analyses. This is a significant drawback given that the purpose of some single-case designs such as n-of-1 trials is to ultimately meta-analyze and find patterns across studies. Finally, the reasons for delayed effects remain unexplored.

Significance

An unknown change-point model can help demonstrate immediacy which is a criterion for strong evidence of causality. A simple change-point model which is typically used in SCD where the intercepts and/or slopes are compared across phases cannot provide evidence of this. First, as outlined in the WWC standards (Kratochwill et al. 2013) immediacy is necessary to show strong evidence of causality in SCDs. By using the proposed model where the change-point is assumed as unknown and later confirmed, SCD researchers will have a statistical tool that can confirm the immediacy of treatment effect using all the observations. This is the first known inferential approach that provides evidence of immediacy in SCDs. If there is lack of immediacy, researchers can investigate threats to internal validity such as implementation, and

presence of delayed or gradual effects. They can also decide how the effect sizes can be adjusted to take into account gradual and delayed effects.

Second, SCD typically has fewer time-points and many models of good fit may be possible in such a case with only slight variations in treatment time points. That is, one may be able to find significant differences in trends if one were to fit a model with baseline running up to time point $t+1$ or $t-1$. For instance, consider the popular example of an ABAB design from Lambert et al. (2006) shown in Figure 2. The middle panel overlays the line of best fit on the measurements. The dotted vertical lines show the end of each phase. If the researcher were not aware of the time-points of treatment, he/she may have fitted the model shown in the right panel to the data and still have found reasonable model fit for the data. The rightmost panel was formed by moving the first time point of the A2 phase to the B1 phase. In both fit scenarios the slopes and intercepts seem significantly different from their adjacent phases. For instance, in Figure 2 it is unclear if the dependent variable at the 15th time-point contains some amount of washout effect from the B1 phase or if it truly belongs to the A2 phase. The method proposed in the present study is recommended to determine and confirm if the estimated change-point is the same as the expected change-point. Only after this confirmation would further estimation of effect sizes between phases be valid.

INSERT FIGURE 2 ABOUT HERE

Problems with Commonly Used Statistical Analyses in SCDs

Unfortunately, commonly used statistical analyses cannot be directly applied to SCD data. Statistical methods for SCDs are hampered by the relatively small amounts of data available on each case. In fact, inference is usually based on dozens rather than hundreds of observations. In their survey of 809 single-case designs published in 113 studies, Shadish and

Sullivan (2008) found that 45.3% of the studies had 5 or less observations per phase. Linear change trajectories estimated from so few observations have considerable sampling uncertainty, even if the observations are independent. Statistical estimation is made more complex by the fact that observations from the same case are not independent but are likely to exhibit an autocorrelation structure that has to be taken into account in the statistical analysis (Huitema, 1985; Huitema & McKean, 1998, 2000). Errors of SCD data are typically lag-1 autocorrelated or serially dependent over time with a lag-1 (Huitema, 1985; Huitema & McKean, 1998, 2000). That is, the error at time point $t + 1$ is correlated with the error at time point t. This violates the assumption of most parametric and non-parametric statistics and results in biased estimates and inflated Type-I error rates (Shadish, Rindskopf, Hedges, & Sullivan, 2013). For time series with fewer than 50 points, the estimates of autocorrelation are negatively biased. Additionally, fewer time points are also plagued by larger sampling error problems. Autocorrelated errors means that there is some pattern between consecutive error terms, which manifests itself as a pattern between observed values. What makes it visually difficult to analyze is whether this pattern is because of a slope (i.e. change in the observed variable with time) or due to autocorrelated errors.

When the observations have an autocorrelation structure (as in SCDs), the within-phase trajectories of change are even more uncertain (Brossart, Parker, Olson, & Mahadevan, 2006; Gorman & Allison, 1996). Complexities such as autocorrelation structure usually mean that exact small-sample frequentist methods are intractable and we are forced to rely on large sample methods (such as maximum likelihood). Moreover, there are many different types of single case designs, each of which poses somewhat different analytic challenges (see, e.g., Shadish, et al, 2015).

The present study proposes a different analytic strategy to analyze SCDs. Rather than fitting separate models to each phase of the design with the phases defined a priori, we fit linear models to each phase separately, but do *not* assume that the boundary between phases is known a priori. We use a Bayesian model to let the data define the point where the change between phases occurs. Such models have been applied successfully to study inflation (Koop & Potter, 2004), water flooding (Zaman, Rahman, & Haddad, 2012), Alzheimer's disease (Li, Dowling, & Chappell, 2015), menstrual cycle (Huang, Elliott, & Harlow, 2014), and climate variations (Beaulieu, Chen, & Sarmiento, 2012). In this paper we will discuss how the Bayesian unknown change-point model can be used to provide inferential statistical evidence of immediacy to supplement visual analysis and identify delayed effects.

Unknown change-point models can be particularly advantageous in SCDs because it provides an objective procedure for dealing with a latent feature of the data: the actual point at which the trajectory of observations change. This may be consistent with data that show delayed effects or may reflect no obvious change in systematic behavior related to the phases of the design. Either way, by treating the change-point as an unknown we *allow the data to speak for itself*. It is important to note that Bayesian methods are exact, small sample statistical procedures. Thus Bayesian methods obviate the ambiguity accompanying the use of large sample methods (such as maximum likelihood estimation) on the usually small SCD datasets. Based on the nature of the design, multiple unknown change-points could be modeled and additional evidence of criterion that supports causality obtained every time the change point is estimated close to the true value. Variables such as participant identifier, treatment identifier, unknown number of change-points, etc. can be included in the unknown change-point model to accommodate all types of SCDs.

Bayesian Statistical Methods

We use Bayesian method to estimate the unknown change-point model. Bayesian estimation works well with small sample data because it does not depend on asymptotic or large sample theory (Ansari & Jedidi, 2000; Ansari, Jedidi, & Jagpal, 2000; Dunson, 2000; Scheines, Hoijtink, & Boomsma, 1999). Bayesian estimation provides complete distributional information about the parameter along with the credibility of each value the parameter can take (Kruschke, 2013). That is, a parameter is not estimated as a single point estimate with an associated standard error of the estimate, but rather as a distribution with a probability value associated with each possible value for the parameter. This posterior density is proportional to the product of the likelihood of the data and the prior information about the parameters (Bayes' theorem). The likelihood is the information contained in the data. Prior information can be systematically incorporated by examining results from other studies or through previous knowledge or be least informative or can be anywhere on this continuum.

The posterior distribution can be used to compute any summary statistic for the parameter of interest. For instance, the standard deviation of the posterior for the change-point, to some extent, indicates the certainty of change in functional relationships between phases (more about this is discussed in the illustration section). Posteriors with high probability densities at several time-points do not support immediacy.

The credibility interval or the highest density interval surrounding a Bayesian parameter estimate is more straightforward to interpret because it is part of a probability density function and has a shape associated with it (Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2013; Lynch, 2007). That is, a 95% credibility interval means that the probability that the true parameter value lies in this interval is 95%. For instance, consider the posterior density of

standardized mean difference between phases d obtained from Bayesian unknown change-point model which is given in the bottom panel of figure 2 for the delayed effect illustration. First, the 95% credibility interval of d has a truncated bell-shaped curve and is fairly wide $[5.40, 10.21]$, but is nonetheless a large effect. This credibility interval can be used to test the region of practical equivalence (ROPE). Let us assume that the researcher specifies a ROPE of the effect size to be between 5 and 10, perhaps because these values are considered as "large" effects in the particular substantive area. Given that most of the values in the credibility interval of d fall within this region, the hypothesis that the effect is large can be accepted. Of particular interest to SCD researchers is Shadish et al.'s (2013) finding that Bayesian estimates of autocorrelation were more accurate than frequentist estimates. They also showed that Bayesian credibility interval estimates of autocorrelation are more accurate than frequentist confidence intervals which have undercoverage. That is, fewer than the expected number of frequentist confidence intervals contain the true value of autocorrelation.

Readers may refer to Kruschke (2011a, 2011b, 2013) for more details on ROPE and accepting the null hypothesis in Bayesian statistics. The Bayesian method is computationally intensive (involving large integrals). Therefore modern sampling methods such as the Gibbs sampler are used to estimate parameter values to circumvent tricky approximations to or sometimes even unsolvable integrals.

Change-point Models

There is vast literature surrounding developing approaches to change-point models. Bai (1994, 1997) presented a least squares estimation approach for change-point problems. Barry and Hartigan (1993) and Carlin, Gelfand, and Smith (1992) are good background materials for Bayesian change-point problems. Chib (1998), Jann (2000), Jeong and Kim (2013), and Kim and

Cheon (2011) extended the Bayesian unknown change-point model to multiple unknown change-points using hidden Markov models, genetic algorithms, and annealing stochastic approximation, respectively. Raftery and Akman (1986) presented an algorithm for Bayesian analysis of Poisson distributed data. Recent developments in this area include newer algorithms, extensions to the models, and applications. Adams and McKay (2007) extended change-point models to Bayesian online change-point detection. Bayesian unknown change-point models have been applied in various fields such as ecology (Thomson et al., 2010), marine biology (Durban & Pitman, 2011), hydrometeorology (Perreault, Bernier, Bobée, & Parent, 2000), and stock data (Lin, Chen, & Li, 2012). Recently Kim and Jeong (2016) developed an approach to change-point modeling in autocorrelated time-series where the number of change-points is unknown. The current study is the first of its kind to apply Bayesian change-point models to single case designs.

Model and Notation

We investigate the performance of the Bayesian unknown change-point model in a simple case of AB design (i.e., one unknown change-point) using Monte Carlo simulation. We then extend it to multiple baseline design in the illustration section that follows. A continuous, normally distributed dependent variable with no trend (slope) is considered in the simulation study. This can be easily extended to models with trend by modelling slopes in the equations below. For count and proportion data, the model would have to be modified to reflect their distributions more accurately (e.g., Rindskopf, 2014). But distributional assumptions will not change the basic framework presented here and this framework can be adapted for different types of variables and distributions by modifying equations 1, 7, and 8. For instance, for count data equation 1 may be modified as a generalized linear model with logit function. The dependent

variable may follow a binomial or a Poisson distribution depending on whether the count data has limits or not, respectively.

Let us assume that the observed value at the first time point (y_{p1}) in phase p follows a normal distribution with the mean of $\hat{y}_{p1}$ and standard deviation of 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:

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

In equation 2, H_{pt-1} is the past history and Θ is the vector of parameters, 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, σ_e, and σ_ε 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; Hedges, Pustejovsky, & Shadish, 2012, 2013; Swaminathan, Rogers, & Horner, 2014; Velicer & Molenaar, 2012). The linear regression model 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. 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:

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

Equation 6 can be rewritten as:

$$\beta_{0p} = \beta_{01} * dummy + \beta_{02} * (1 - dummy); \quad (7)$$

$$\text{where } dummy = step(t_b - t) .$$

The step function returns the value 0 if the argument is negative and 1 otherwise. Therefore the indicator variable (called dummy here) is assigned a value of one if the time-point is in the baseline phase and a value of zero otherwise. In a regular change-point model, t_b is known while in the unknown change-point model t_b is an estimated parameter.

Usually a complex method is used to work around the if-else executable statement with varying intercept values depending on the phases, which themselves are unknown (e.g. Harring, Cudeck, & du Toit, 2006). This requires that the researcher knows the order of the relationship between the slopes in the two phases (i.e. whether the slope of phase 1 is greater than the slope of phase 2). This order will inform whether the intersection point of the two lines can be classified as a maxima or a minima of the values. Obviously the method will not work when there are only level differences but no slopes modeled such as the one we consider. There is a more straightforward solution using the step function in software programs such as JAGS (Just another Gibbs sampler, Plummer, 2003) and BUGS (Bayesian using Gibbs sampler, Lunn, Spiegelhalter, Thomas, & Best, 2009). BUGS and JAGS parametrize the normal distribution in terms of the precision $\tau = 1/\sigma^2$, rather than variance. We will parametrize using variance instead of precision in order to keep the statistical notation common to Bayesians and non-

Bayesians. Appendix A contains a sample dataset and code in R that calls JAGS to fit a Bayesian unknown change-point model.

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). Assuming that the researcher is unaware of the phase the observation belongs to, let us consider a time-series SCD data $Y = (y_1, y_2, \ldots, y_n)$ such that the functional relationships between the dependent and the independent variables differ based on the phase. In other words,

$$y_t = \begin{cases} \theta_1 \; if \; t \leq t_b, \\ \theta_2 \; otherwise, \end{cases} \tag{8}$$

where the parameters θ_1 and θ_2 are the mean levels of the dependent variable y_t in phases 1 and 2, respectively. Parameters θ_1, θ_2, t_b need to be estimated. In equation 8, $\theta_1 = g(\beta_{01}, \sigma_\varepsilon, \rho)$ and $\theta_2 = g(\beta_{02}, \sigma_\varepsilon, \rho)$. 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 the parameter vector $(\beta_{01}, \beta_{02}, \sigma_\varepsilon, \rho, t_b)$. 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.

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.

Method

Simulation

A simulation study was conducted to test the performance of the unknown change-point model with uninformative to relatively less informative priors. The generating values of the parameters (effect size d, standard deviation σ, and autocorrelation ρ) were selected based on the most commonly occurring parameters in SCD studies (Maggin, O'Keefe, & Johnson, 2011; Shadish & Sullivan, 2008). Maggin et al.'s (2011) study was based on 68 single-case design studies published between 1985 and 2009. Shadish and Sullivan's (2008) study was based on 809 single-case designs published in 113 studies in the year 2008 in 21 journals. Data was generated based on equations 1-8. The intercepts of the two phases were 1, 2, 3, or 5 standard deviations (σ) apart. This standardized mean difference will be referred to as d. Two autocorrelation values ($\rho = .2, .5$) and three phase lengths ($l = 5, 8, 10$) were considered. Both the treatment and the baseline phases had an equal number of time points. The standard deviation within each phase was $\sigma = 0.2$. The simulation study followed a fully crossed 2 × 3 × 4 factorial design.

One hundred datasets were generated for each condition resulting in 2400 datasets. The sufficiency of the number of replications was determined based on several procedures including ones described by Carsey and Harden (2014), and Koehler, Brown, and Haneuse (2009). First,

the estimates converged to stationarity 100% of the time in the simulated and in the real datasets. Coverage rate of change-point credibility intervals was computed for the first 50 replications and compared against coverage for all the 100 replications. Coverage rate is the rate of credibility intervals that contain the true parameter value. They ranged from 94% to 100% and differed only up to 2% in 5 of the 24 conditions. Otherwise they stayed the same across 50 and 100 replications. We tested the Monte Carlo percent bias of the posterior mode estimates of the change-points ($\hat{\phi}_R^b$) across the replications as defined in equation 9. If $\widehat{M_r}$ and M are the estimated posterior mode and true value of the change-point for the rth replication, respectively, then for a total of R replications,

$$\hat{\phi}_R^b = \frac{1}{R}\sum_{r=1}^{R}\frac{\widehat{M_r} - M}{M} \times 100\,. \qquad (9)$$

The Monte Carlo percent biases of the posterior modes were stable after 50 replications for most conditions and after 75 replications for all conditions. The average of the posterior SDs of the change-point differed from the standard deviation of the modes of the change-point from -1.78 to 0.44 with no particular direction of bias. The similarity of the values suggest the sufficiency of the estimator. Similarly, the average posterior SD of the intercept estimates ranged from 0.08 to 0.20 which was close to the population SD.

Priors

Generating values and prior distributions for the parameters are given in Table 1. Given the small sample size nature of SCDs, the choice of prior plays an important role in the posterior density estimate. A hierarchical prior distribution was used for the intercepts (β_{01}, β_{02} in equations 4 and 6). This allows the parameters of the priors to be estimated from the data rather than specifying them to have subjective information (Efron & Morris, 1975; Gelman, 2006;

James & Stein, 1960). The intercepts of both phases were independent of each other. The intercepts were drawn from normal distributions with means simulated from a normal distribution with mean 0 and standard deviation 100, and variances simulated from an inverse gamma distribution with shape parameter 1. Gelman and Hill (2007) advocated using a noninformative uniform prior on the standard deviation with the upper limit sufficiently large relative to data. Therefore the upper limit on the standard deviation of the intercepts was 500 times the true population standard deviation. The relatively informative inverse gamma (1, 1) distribution was used for variance because Gelman (2006) cautioned against the use of very low values such as .01 and .001 for the gamma prior which lead to improper posteriors. The limited sample size in SCD data will only further worsen this situation.

de Vries and Morey (2013) suggested a beta prior on autocorrelation with shape parameters $\alpha = 0$ and $\beta > 1$, preferably $\beta = 5$ in order to keep the prior sufficiently vague. However, this places higher probability density towards 0. In order to reflect the generating autocorrelation values, a uniform distribution with limits from -1 to 1 was used as the prior for autocorrelation. This is vaguer than the prior suggested by de Vries and Morey (2013) and captures all possible mathematical values of autocorrelation. The change-point had a relatively uninformative uniform discrete prior distribution with equal probability of falling between 3 and $(T - 2)$. This range was chosen because at least 3 data points are required to identify a pattern in each phase. This also reflects the standards for single case designs which requires at least 3 data points per phase (Kratochwill et al., 2013). Following Swaminathan, Rogers, and Horner (2014) and de Vries and Morey (2013), error variance and autocorrelation was assumed equal across phases.

Diagnostics and Interpretations

Four parallel chains were run with starting values independently generated for each chain from the prior distribution. Convergence was checked using two convergence diagnostics: the multivariate potential scale reduction factor (MPSRF, Brooks & Gelman, 1998) and Heidelberger and Welch's convergence diagnostic (1983). The package RunJAGS (Denwood, 2013) conveniently runs parallel chains and iterates the model estimates until convergence. In order to compare the performance (i.e., accuracy of estimates, time taken till convergence) of the unknown change-point model (Model 1), a piecewise model was estimated where the change-point was specified (Model 2). Both models together took between 109 and 390 seconds to run until convergence. Means, standard deviations, modes, and 95% credibility intervals of the change-point posterior distributions were obtained. Root mean square errors (RMSEs) of the posterior mean and mode, mean posterior standard deviation (MPSD), and bias of the posterior mode of the change-point were computed. Because the change-point is discrete the posterior mode was considered. The posterior means of change-points may often not be measured discrete time-points and are heavily influenced by extreme values. What is of importance here is how often the change-point was accurately estimated. Posterior mode is that quantity.

Three individual ANOVAs were run with MPSD, and RMSE of the posterior mean and posterior mode of the change-point as dependent variables. The independent variables were $d, \rho,$ and l. It should be noted that although phase length is an independent variable in the ANOVAs, the phase length was perfectly correlated with the range of the prior distribution of the change-point. That is, longer phase lengths contain more data-points, which mean more information. However this is accompanied by a less informative prior where more data-points with low but equal probabilities are possible candidates for change-point. Therefore, the increase

in the information contributed by longer phase lengths is countered by the less informative prior specification. This makes it difficult to distinguish which part of the effect was due to phase length and which part due to the range of the prior.

Results

Study 1: Simulation

Overall trends from ANOVAs. Eta-squared effect sizes from independent ANOVAs are reported in Table 2. RMSEs and MPSDs decreased with increase in both standardized mean difference between the phases and phase length (Figure 3). Standardized mean difference (d) explained the maximum variation in RMSE of means (45.54%) and modes (65.59%), and MPSDs (71.93%) of change-point estimates. MPSD was largest when $d = 1 \; and \; l = 10$. This is because smaller differences between phases makes it difficult to discern patterns clearly. In data with longer phase lengths this is further compounded by the prior that places very low but equal probabilities on several data-points as possible candidates for change-point. That is, there are 16 possible values the change-point can take in a dataset with phase length 10. This is because according to WWC standards there need to be at least 3 observations per phase. Each of these have a probability of $\frac{1}{16}$ for being the change-point. This is more spread out when compared to data with phase length 5 which will have 6 possible candidates each with probability of $\frac{1}{6}$ for being the change-point. Again, the effect of phase length on the posterior standard deviation (17.21%), and the RMSE of mean (34.13%) and mode (18.53%) cannot be separated from the effect of the range of the uniform prior specified for the change-point. It is a well-known fact that prior has larger effect on the posterior for small samples. But the lack of

information due to sample size is compensated by the information samples with large population effect size.

INSERT FIGURES 3 AND 4, AND TABLE 2 ABOUT HERE

Credibility Intervals and Posteriors. Credibility intervals of the change-points for phase length of 5 are shown in Figure 4. The credibility intervals for other phase lengths followed the same pattern. When both standardized mean difference and phase length were small, most credibility intervals spanned the entire prior distribution range. The original data plot is shown for one case with large posterior standard deviations (top left) and one case with small posterior standard deviations (top right, Figure 5). Both cases have large immediacy indices. We define the *immediacy index* as the mean difference between the first and last three observations of the dependent variable in the baseline and treatment phases, respectively. It may seem that both large immediacy index and clarity of the distinct data pattern between phases contribute to a narrow CI, that is, more certainty in the estimate. However, no clear statistical pattern emerged from using the immediacy index as a variable that explains the variation in the uncertainty of the change-point estimate. It should be noted that the immediacy index is only a function of six consecutive data-points, whereas the effect size is a summary of all data. Therefore, this result is not surprising. The posterior plots of these two cases are shown in the bottom panel of the figure.

INSERT FIGURE 5 ABOUT HERE

For case 1, the 95% credibility interval of the change-point ranged from 3 to 18 which is the range of the prior. The mode and the mean of the change-point were 10 and 10.387, respectively. In this case although the immediacy index is not small, there is overlap between the values in the baseline and treatment phases. That is, most values in the baseline phase seem candidates for the treatment phase as well. Therefore it seems reasonable that the algorithm has

larger uncertainty associated with the change-point estimate in this case. The posterior plot of the change-point sheds more light on the probabilities associated with each possible time point in this interval. In this posterior distribution the true value seems the most likely candidate for the change-point, but the distribution has two other frequently occurring values because of this overlap in values between phases. This shows that in addition to the posterior mean, mode, and standard deviation, researchers will benefit by examining the shape of the posterior. The posterior distribution of case 2shows that except for the mode, there is close-to-zero probability associated with integer values from 3 to 18 in this posterior. The mean, mode, 2.5th and 97.5th percentiles of this distribution are all 5. This is clear evidence for 5 as the change-point.

The immediacy index was highly negatively correlated with the posterior standard deviation of the change-point estimate ($r = -.67$). The posterior standard deviation was almost zero for large immediacy index values. Autocorrelation values had a very small impact on the estimates. The distance between the RMSEs of the mode and the mean decreased with increase in the standardized difference between phases. The RMSE of mode was smaller than that of the mean when $d \geq 3$.

Effects of Autocorrelation and Phase Length. The accuracy of the intercept estimates for both unknown and known change-point models and both phases increased with decrease in autocorrelation and increase in phase length. Mean posterior standard deviation of the intercept estimates for both models and both phases increased with increase in autocorrelation and decrease in phase length. The mean posterior standard deviations of the first and second intercepts between the two models were about .27 and .53 standard deviations apart, respectively. The posterior means of the intercepts for both phases and autocorrelations were less than .017 standard deviations apart between the unknown and the known change-point models.

Study 2: Application of Bayesian unknown change-point model to real data

Multiple baseline design is an extension of the AB design where the treatment is administered to different participants at different time points. This design allows at least three demonstrations of the treatment effect, which is a required criterion to meet evidence standards for SCDs (Kratochwill et al., 2010). An example data from Laski, Charlop, and Schreibman (1988) is shown in figure 6. We analyzed this multiple baseline data using the Bayesian unknown change-point model presented above. The parameter estimates are given in tables 3, 4, and 5. Except for cases 1, 3, and 7 the change-points were estimated with good accuracy as seen from their posterior distributions in figure 7a. This indicates support for immediacy effect. Visual analysis shows immediacy for all cases except case 8. The mean difference between the first and last three observations of phases 2 and 1, respectively were 19-48 points for all cases except case 8. There is considerable overlap for cases 7 and 8, some overlap for case 1, and a one point overlap for case 4. For case 1, although the mean difference is 28, there is doubt about whether change happened because of the treatment or because of autocorrelation between points 4 and 5. In the Bayesian analysis, although the mode of case 1 is correctly estimated to be 4, the probability mass around time-point 3 cannot be ignored. This is because the model takes the pattern of all the observations into account. Similarly there is considerable overlap in the values of the observations in baseline and treatment phases for case 7. This may have led to larger uncertainty in the change-point estimate. Case 8 had smaller overlap and yet had accurate estimates. This was probably due to the large immediacy index. Note that Laski at el. (1988) did not examine immediacy.

Figure 7b shows the posteriors of d for each case. The d values indicate simple standardized mean differences after removing autocorrelation, that is, the mean difference was

standardized using σ_ε. None of the 95% credibility intervals for effect sizes contained 0, but the 2.5th percentile of the effect size for case 7 was very close to 0. In addition to the considerable overlap between observations, the small effect size value in case 7 may have also impacted its change-point estimate. Laski at el.'s (1988) conclusion that the treatment effect was lowest in case 7 was based on descriptive statistics. That is, they only examined the difference in the averages between the phases and therefore could not make inferential comparisons.

If the ROPE for strong effect size is built between 3 and 6, the null hypothesis can be accepted for cases 2, 3, 4, and 5 because 95% of the posterior density falls within this region. The variance of the distribution from which the intercepts were drawn had a mean of 3185.2 with 95% credibility interval ranging from 1492 to 6073. A standard deviation of 56.4 may be considered large, but the range of observed values in the baseline phase ranged from 0 to 81.91 and in the intervention phase ranged from 27.67 to 92.15. The standard deviation had to be wide enough to accommodate such a wide range of values. The 2.5^{th} and the 97.5^{th} percentiles of the within person standard deviations ranged from 4.04 to 10 and was much smaller than the between person standard deviation.

INSERT FIGURES 6, 7a, and 7b, AND TABLES 3 - 5 ABOUT HERE

Recommendations for Interpretation

Based on the results of the present study, here are our recommendations for practitioners who wish to implement this method for their own data. The two questions the method answers are: Is there evidence of immediacy? If yes, how strong is this evidence? In addition to visual analysis, the posterior distributions of the change-points should be examined to confirm immediacy. Immediacy is supported when the change-point posterior mode is estimated to be near the change-point, and the distribution is narrow and clearly unimodal. Delayed effects or

lack of immediacy is indicated when the mode is estimated to be at any other data point with a narrow posterior. A posterior distribution with large variance but close to the true value may indicate lack of immediacy and possible gradual effect. However, the shape of the posterior needs to be examined before concluding so. Note that when using a categorical prior, it is not uncommon to obtain a large standard deviation for the posterior when the effect is moderate or low. This is because the estimates can only jump from one discrete time point to another. So jumping from time-point 7 to 8 will result in larger posterior standard deviation than jumping from 7 to 7.5.

We have illustrated building a ROPE around the effect size. This can be extended to change-points as well. In cases where a delayed or gradual effect is expected, a region of practical equivalence (ROPE) may be built around the change-point to check for probabilities of credible values in this region. The credibility interval of the change-point is tested so the researcher can determine which of the credible values in this interval fall within the ROPE. When most of the values fall within the ROPE the null hypothesis that the change in the dependent variable takes place within the region of when the intervention is expected to have an effect is accepted. Immediacy of the treatment effect is indicated when the null is accepted for a ROPE that contains only one value, that is, the true change-point value. The possibility of accepting the null rather than failing to reject the null is an especially attractive feature of Bayesian statistics that SCD researchers would benefit from.

Discussion

The present study demonstrates how Bayesian unknown change-point models can be used to evaluate SCD data for a simple AB-design and multiple baseline design. There are several advantages to this method: (a) So far immediacy has only been confirmed using visual

analysis. The method presented here is the first inferential statistical method that can be used to confirm and evaluate immediacy; (b) Unlike visual analysis, the method identifies immediacy using all observations to find patterns within phases. In addition to identifying immediacy, the proposed method quantifies immediacy in cases where change between phases is unclear, gradual, or delayed; (c) The use of Bayesian methods allows the researcher to examine the shape of the posterior distribution in addition to its descriptive statistics. This gives a clearer evaluation of the quality of the estimates; (d) For treatments with predicted delays, ROPE can be systematically built around the expected immediacy effect of the treatment. The researcher has the possibility of accepting the null that the change-point indeed occurred where it was expected to occur; and (e) The model presented can be modified to accommodate other distribution types, data types, functional relationships between time and the dependent variable, and add explanatory variables.

The results of the simulation study show that a standardized mean difference of 3 or larger (computed ignoring the autocorrelation) was necessary for a reasonably accurate change-point estimate. The posterior standard deviations decreased drastically with increase in standardized mean difference. In fact, the entire 95% credibility interval for the change-point of most datasets with large standardized mean difference and phase length ≥ 8 was a single time-point, which was the true value of the change-point. Although requiring at least 8 data points per phase may seem high, it is not unreasonable in SCD research because 54.7% of the studies reviewed by Shadish and Sullivan (2008) had more than 5 points per phase. However, only 20% of the 62 effect sizes reviewed by Shadish and Sullivan (2008) had an effect size of 3 or more. This could be countered by having longer phase lengths and specifying more conservative priors based on previous research. Most importantly, the benefits of collecting three extra points per

phase in order to facilitate the analysis presented in our study make the efforts worthwhile. We considered an intercept-only model which means the model can only detect sudden changes following the treatment effect. When treatment effects are more gradual or vary with time, slopes must be modeled. This is an avenue for future research.

The prior distribution impacted the interval widths of the estimates. In the present study we used uninformative, naïve uniform (categorical) priors for the change-point. We could not separate out how much of the variation in RMSE was due to the prior and how much was due to phase length. This is because longer phase lengths imply more data and therefore more information. But longer phase lengths also mean less informative prior. So it is unclear how much the data compensated for the 'uninformativeness' of the prior. Prior choice is especially important given the small sample nature of SCDs. Although we did not test the impact of prior choice on the estimates, we can speculate based on research (e.g. Natesan, Nandakumar, Minka, & Rubright, 2016) that using very vague priors [e.g. $N(0, 1000)$] instead of the hyperpriors we used could lead to less accurate estimates and severe shift in the scales. Given that the intercepts of the two phases are independently estimated, this may also affect the accuracy of the change-point estimate.

Depending on the treatment, the design, and other information (e.g. from other published studies, meta-analyses), priors may be specified more systematically. For instance, if previous research shows that the participant will take 5 days to begin responding to the treatment, the probability of the time point being the change-point can be specified to range from time-point 8 to $t - 2$, if there are t time points. This allows for the WWC standards that there need to be at least three time points per phase. Therefore, if the treatment were administered even as early as the third time point change would be observed only beginning the eighth time point.

Nonetheless, the naïve priors that we used yielded good estimates for the Laski et al. (1988) data. A wider discussion is needed on how to show evidence of causality when a delayed effect is expected due to the nature of the treatment. For example, it is not uncommon for a drug to take effect or a child to take time to respond to an intervention. Concluding that these interventions do not have causal effects solely based on immediacy may indicate following a very rigid framework of rules that does not acknowledge the nature of different treatments.

Rindskopf (2014) noted that Bayesian methods are likely to become the preferred method of analysis for SCDs. This is because Bayesian methods work well with small sample data, are robust to several distributional assumptions, and allow a more comprehensive understanding of the statistical estimates through posterior analyses and region of practical equivalence (ROPE) tests. For instance, when count data or proportion data are used as dependent variables, non-linear models such as generalized linear models may be used to model the data. Generalized linear models are large sample procedures—that is, their properties are guaranteed to hold only in large samples. In contrast, Bayesian methods yield exact small sample results. Therefore, investigating Bayesian unknown change-point models for count and proportion data is a natural extension. The most advantageous aspect of Bayesian statistics that is of particular interest to SCD researchers is the credibility interval, which allows direct probabilistic interpretation of a statistical estimate (Gelman, Carlin, Stern, Dunson, Vehtari, & Rubin, 2013). Bayesian methodology is a relatively new territory for most SCD researchers. Therefore, learning Bayesian methods can be challenging.

It is worth noting that the time taken for estimation in the Bayesian analysis increases with increase in model complexity. For instance, the two models together took up to 5 minutes to run. However, this is a small price to pay for the additional information obtained from Bayesian

analysis. Gast and Ledford (2014) reiterated the need for statistical techniques that are applicable to many types of SCD data, used by practitioners with little training, valid, reliable, and sensitive to change in SCD research. We believe that the method we presented in this study is one such technique.

JAGS was used to fit the model in this study. JAGS and BUGS (OpenBUGS and WinBUGS) use the same format and are more integrated into R. They involve the same ease of implementation and are easier for people already familiar with R. Stan (20016), a newer Bayesian software program requires the model to be defined in a more prescriptive manner and is supposed to be more efficient than JAGS. However, Stan would require more programming skills than JAGS and BUGS.

The presented method only considered equal phase lengths. Although logically we do not foresee unequal phase lengths to affect the accuracy of the estimates, this cannot be known for certain. We did not consider models with slopes or other types of functional relationships between the independent and the dependent variable across phases. Multiple phase change designs such as ABAB designs are frequently used in SCDs because their setup can help show three demonstrations of treatment effect, in accordance with WWC guidelines. Extending the current study to multiple phase change designs would require investigating the performance of multiple unknown change-points model. Another avenue for research is developing an effect size that takes into account the accuracy of the change-point estimate. SCD researchers would greatly benefit from a study that recommends a course of effect sizes and other measures in studies with gradual or delayed effects.

References

Adams, R.P. & MacKay, D.J. (2007). Bayesian online changepoint detection, Technical Report arXiv:0710.3742v1 [stat.ML]. University of Cambridge.

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

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.

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/.

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

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

Bai, J. (1994). Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis, 15*, 453–472.

Bai, J. (1997). Estimation of a change point in multiple regression models. *Review of Economic Statistics, 79*, 551–563.

Barry, D. & Hartigan, J.A. (1993). A Bayesian analysis of changepoint problems. *Journal of the American Statistical Association, 88*, 309-319.

Beaulieu, C., Chen, J. & Sarmiento, J. L. (2012). Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society A, 370*, 1228-1249.

Brooks, S., & Gelman, A. (1998). Some issues in monitoring convergence of iterative solutions. *Journal of Computational and Graphical Statistics*, *7*, 434-455.

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.

Carlin, B.P., Gelfand, A.E., & Smith, A.F.M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics, 41*, 389–405.

Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Carsey, T. M. & Harden, J. J. (2014). Monte Carlo Simulation and Resampling Methods for Social Science. Thousand Oaks, CA: Sage.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics, 86*, 221-241.

de Vries, R. M., & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods, 18*, 165-185.

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.

Denwood, M. J. (In Review). runjags: An R package providing interface utilities, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*. URL: http://runjags.sourceforge.net

Durban, J. W. & Pitman, R. L. (2011). Antarctic killer whales make rapid, round-trip movements to subtropical waters: evidence for physiological maintenance migrations? *Biology letters, 1-4.* DOI: 10.1098/rsbl.2011.0875

Efron, B. & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association, 70,* 311–319.

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

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.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis, 3*, 515-533.

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.

Gelman, A. & Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge University Press: NY.

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.

Harring, J. R, Cudeck, R., & du Toit, S. H. C. (2006). Fitting partially nonlinear random coefficient models as SEMs. *Multivariate Behavioral Research, 41*, 579–596.

Hedges, L. V., Pustejovsky, J. E., & Shadish,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.

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

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.

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.

Huang, X., Elliott, M. R., & Harlow, S. D. (2014). Modeling menstrual cycle length and variability at the approach of menopause using hierarchical change point models. *Journal of the Royal Statistical Society, Series C: Applied Statistics, 63*, 445-466.

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

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

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods, 3,* 104–116.

Huitema, B. E., & McKean, J. W. (2000). A simple and powerful test for autocorrelated errors in OLS intervention models. *Psychological Reports, 87,* 3–20.

Jann, A. (2000). Multiple change-point detection with a genetic algorithm, *Software Computation, 2*, 68–75.

James, W. & Stein, C. (1960). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium 1*. Berkeley: University of California Press.

Jeong, C. & Kim, J. (2013). Bayesian multiple structural change-points estimation in time series models with genetic algorithm, *Journal of Korean Statistical Society, 42,* 459–468.

Kim, J. & Cheon, S. (2011). Bayesian multiple change-point estimation with annealing stochastic approximation Monte Carlo. *Computational Statistics*, *25*, 215–239.

Kim, J. & Jeong, C. (2016). A Bayesian multiple structural change regression model with autocorrelated errors. *Journal of Applied Statistics, 43*, 1690-1705.

Koehler, E., Brown, E., & Haneuse, S. J.-P.A. (2009). On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. *The American Statistician, 63*, 155-162.

Koop, G. M. & Potter, S. M. (2004). Forecasting and estimating multiple change-point models with an unknown number of change points. *Federal Reserve Bank of New York Staff Reports, 196*.

Kratochwill, T. R. (Ed.). (1978). *Single subject research: Strategies for evaluating change.* New York, NY: Academic Press.

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

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

Kratochwill, T.R., & Levin, J.R. (Eds.). (2014). *Single-Case Intervention Research: Methodological and Statistical Advances*. Washington, D.C.: American Psychological Association.

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.

Kreft, I. & De Leeuw, J. (1998). *Introducing Multilevel Modeling*. London: Sage.

Kruschke, J. K. (2011a). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*, 299–312.

Kruschke, J. K. (2011b). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press/Elsevier.

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

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.

Laski, K. E., Charlop, M. H., & Schreibman, L. (1988). Training parents to use the natural language paradigm to increase their autistic children's speech. *Journal of Applied Behavior Analysis, 21*, 391-400.

Levin, J. R., O'Donnell, A. M., & Kratochwill, T. R. (2003). Educational/psychological intervention research. In I. B. Weiner (Series Ed.) and W. M. Reynolds & G. E. Miller (Vol. Eds.). *Handbook of psychology: Vol. 7. Educational psychology* (pp. 557–581). New York: Wiley.

Li, C., Dowling, N.M., & Chappell, R. (2015).Quantile regression with a change-point model for longitudinal data: An application to the study of cognitive changes in preclinical Alzheimer's disease. *Biometrics, 71*, 625-635.

Lin, J.-G., Chen, J., & Li, Y. (2012). Bayesian analysis of student t linear regression with unknown change-point and application to stock data analysis. *Computational Economics, 40,* 203-217.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique, and future directions. Statistics in Medicine, 28, 3049–3067.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York: Springer.

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.

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

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. doi: 10.1080/00220973.2012.745470

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *Journal of the American Statistical Association, 78*, 47–65.

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. (2016). Bayesian prior choice in IRT estimation using MCMC and variational Bayes. *Frontiers in Psychology, 7*, 1422. doi:10.3389/fpsyg.2016.01422

Neely, L., Rispoli, M., Camargo, S., Davis, H., & Boles, M. (2013). The effect of instructional use of an iPad® on challenging behavior and academic engagement for two students with autism. *Research in Autism Spectrum Disorders, 7*, 509-516.

Perreault, L., Bernier, J., Bobée, B., & Parent, E. (2000). Bayesian change-point analysis in hydrometeorological time series Part 1. The normal model revisited. *Journal of Hydrology, 235*, 221-241.

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 Core Team, (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Raftery, A. E. & Akman, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika, 73*, 85-89.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models* (2nd ed.). Thousand Oaks, Calif.: Sage.

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

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*, 37–52.

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

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The Role of Between-Case Effect Size in Conducting, Interpreting, and Summarizing Single-Case Research (NCER 2015-002)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. URL: http://ies.ed.gov/.

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.

Shadish, W.R., & Sullivan, K.J. (2008). 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.

Shih, C.-H., Chang, M.-L., Wang, S.-H., & Tseng, C.-L. (2014). Assisting students with autism to actively perform collaborative walking activity with their peers using dance pads combined with preferred environmental stimulation. *Research in Autism Spectrum Disorders, 8,* 1591-96.

Shih, C.-H., Chiang, M.-S., & Shih, C.-T. (2015). Assisting students with autism to cooperate with their peers to perform computer mouse collaborative pointing operation on a single display simultaneously. *Research in Autism Spectrum Disorders, 10*, 15-21.

Shih, C.-H., Wang, S.-H., Chang, M.-L., & Kung, S.-Y. (2012). Assisting patients with disabilities to actively perform occupational activities using battery-free wireless mice to control environmental stimulation. *Research in Developmental Disabilities, 33*, 2221-2227.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel Analysis*. London: Sage.

Stan Development Team (2016). *Stan Modeling Language Users Guide and Reference Manual, Version 2.9.0.* URL: http://mc-stan.org/

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology, 52*, 213-230.

Thomson, J. R., Kimmerer, W. J., Brown, L. R., Newman, K. B., Nally, R. M., Bennett, W. A., Feyrer, F., & Fleishman, E. (2010). Bayesian change point analysis of abundance trends for pelagic fishes in the upper San Francisco Estuary. *Ecological Applications, 20*, 1431-1448.

Zaman, M. A., Rahman, A., & Haddad, K. (2012). Detection of change point in annual maximum flood series over Eastern Australia using a Bayesian approach. Paper presented at the *Hydrology and Water Resources Symposium 2012*.

Table 1

Generating Values and Prior Distributions for the Parameters

Parameters	Meaning	Prior Distributions	Generating values
β_{01}	Intercept of baseline phase	$norm(mu1, sd1)$ $mu1 \sim norm(0, 100)$ $sd1^2 \sim gamma(1, 1)$	1
β_{02}	Intercept of treatment phase	$norm(mu2, sd2)$ $mu2 \sim norm(0, 100)$ $sd2^2 \sim gamma(1, 1)$	$1 - \sigma d$
σ	Standard deviation of y within a phase	$unif(0.1, 5)$	0.2
t_b	Change-point where baseline phase ends	$categorical(c)$ $c = (0, 0, \frac{1}{T-4}, \frac{1}{T-4}, \dots, 0, 0)$	5, 8, 10
ρ	Autocorrelation	$unif(-1, 1)$	0.2, 0.5

Table 2

Eta-squared in Percentages

Sources	RMSE-mean	RMSE-mode	Mean.SD
d	45.54	65.59	71.93
rho	2.92	2.74	0.23
length	34.13	18.53	17.21
d × length	15.37	11.70	9.65
rho × length	1.23	1.21	0.65
d × rho	0.08	0.09	0.15

Table 3

Parameter Estimates for Laski et al. (1988)

		Baseline Phase					Treatment Phase				
Sub	Param	L95	Mdn	U95	M	SD	L95	Mdn	U95	M	SD
1	β_{0p}	17.16	30.22	40.16	29.39	6.37	55.17	59.79	64.17	59.74	2.29
2		27.48	35.04	42.65	34.99	3.85	72.64	78.75	84.38	78.64	3.25
3		-6.01	2.99	12.55	3.26	4.67	39.05	49.56	58.46	49.22	4.90
4		14.23	20.12	25.88	20.13	2.97	52.05	59.10	66.51	59.22	3.66
5		28.99	38.11	46.10	37.90	4.36	73.05	81.77	89.98	81.62	4.31
6		54.45	59.27	64.25	59.24	2.46	78.55	83.07	87.46	83.07	2.26
7		49.88	62.11	72.67	61.69	6.15	67.49	74.14	81.65	74.37	3.58
8		36.67	50.04	59.63	49.23	6.04	68.82	80.05	99.09	81.64	7.95
1	t_b	3	4	4	3.58	0.51					
2		5	5	5	4.99	0.12					
3		7	8	9	7.88	0.82					
4		10	10	10	10.00	0.02					
5		7	7	7	7.00	0.03					
6		7	7	7	7.04	0.23					
7		3	4	14	6.16	3.91					
8		11	11	11	11.01	0.16					
1	σ_e	6.99	8.83	10.00	8.69	0.91					
2		7.36	9.05	10.00	8.90	0.81					
3		8.33	9.50	10.00	9.37	0.52					
4		8.95	9.72	10.00	9.62	0.34					
5		8.02	9.36	10.00	9.22	0.61					
6		3.98	6.19	9.14	6.37	1.36					
7		7.86	9.28	10.00	9.14	0.66					
8		8.75	9.65	10.00	9.55	0.39					
1	ρ	-0.89	-0.47	0.02	-0.45	0.23					
2		-0.46	-0.15	0.18	-0.14	0.17					
3		-0.62	0.11	0.69	0.06	0.36					
4		-0.39	-0.06	0.27	-0.06	0.17					
5		-0.08	0.24	0.59	0.25	0.17					
6		-0.41	-0.17	0.08	-0.17	0.13					
7		-0.27	0.10	0.49	0.10	0.19					
8		-0.08	0.35	0.87	0.37	0.25					

Note. Sub = Subject, Param = Parameter, Mdn = Median, M = Mean; L95 = 2.5th percentile, U95 = 97.5th percentile

Table 4

Parameter estimates for Laski et al. (1988) that varied between phases

Sub	Param	Baseline Phase					Treatment Phase				
		L95	Mdn	U95	M	SD	L95	Mdn	U95	M	SD
1	β_{0p}	17.41	30.64	40.84	29.71	6.36	55.35	59.87	64.27	59.80	2.28
2		27.19	35.04	42.32	35.00	3.82	72.90	78.70	84.72	78.53	3.69
3		-5.75	3.10	12.93	3.39	4.68	39.55	49.57	58.93	49.26	4.93
4		14.11	20.16	25.73	20.13	2.95	52.30	59.16	66.66	59.28	3.65
5		29.59	38.25	46.48	38.06	4.30	73.39	81.72	90.03	81.64	4.23
6		54.47	59.29	64.24	59.25	2.44	78.52	83.10	87.38	83.07	2.23
7		50.11	62.00	72.71	61.62	6.08	67.60	74.10	81.80	74.31	3.63
8		36.41	49.73	59.75	48.91	6.16	68.99	80.27	101.04	82.09	8.51
1	μ	16.78	30.16	41.37	29.52	6.84	53.39	59.70	65.48	59.58	3.61
2		26.27	35.02	43.70	34.91	4.67	71.30	78.67	85.52	78.48	4.23
3		-6.17	3.03	13.89	3.29	5.26	38.59	49.47	60.11	49.14	5.64
4		13.03	20.13	27.13	20.07	3.93	51.04	59.09	67.58	59.12	4.64
5		28.24	37.88	47.20	37.72	5.13	72.50	81.59	90.87	81.37	5.16
6		52.79	59.25	65.50	59.09	3.75	76.57	82.98	88.70	82.85	3.79
7		49.16	61.95	73.32	61.54	6.67	65.94	73.77	82.39	73.92	4.75
8		35.47	50.13	60.59	49.15	6.75	67.65	80.28	98.77	81.38	8.14

Note. Sub = Subject, Param = Parameter, Mdn = Median, M = Mean; L95 = 2.5th percentile, U95 = 97.5th percentile

Table 5

Parameter estimates for Laski et al. (1988) that were fixed between phases

Sub	d estimates						σ_ε estimates					
	L95	Mdn	U95	M	SD	Mo	L95	Mdn	U95	M	SD	Mo
1	2.05	3.46	5.11	3.52	0.79	3.36	6.98	8.84	10.00	8.69	0.91	9.43
2	3.55	4.92	6.49	4.96	0.77	4.85	7.36	9.05	10.00	8.90	0.80	9.55
3	3.44	4.96	6.37	4.91	0.73	4.98	8.32	9.49	10.00	9.36	0.53	9.77
4	3.08	4.06	5.10	4.07	0.51	4.01	8.94	9.71	10.00	9.62	0.34	9.87
5	3.46	4.73	6.20	4.76	0.70	4.71	8.02	9.36	10.00	9.22	0.62	9.70
6	2.08	3.87	5.79	3.91	0.97	3.77	4.04	6.19	9.18	6.36	1.35	5.88
7	0.03	1.37	2.82	1.39	0.71	1.35	7.87	9.28	10.00	9.14	0.66	9.66
8	1.37	3.16	6.11	3.41	1.27	2.82	8.78	9.66	10.00	9.55	0.39	9.85
	ρ estimates						t_b estimates					
1	-0.89	-0.47	0.01	-0.45	0.23	-0.48	3	4	4	3.59	0.51	4
2	-0.47	-0.15	0.16	-0.15	0.16	-0.15	5	5	5	4.99	0.15	5
3	-0.64	0.12	0.70	0.08	0.36	0.22	7	8	9	7.89	0.83	7
4	-0.38	-0.06	0.28	-0.06	0.17	-0.07	10	10	10	10.00	0.02	10
5	-0.09	0.24	0.58	0.24	0.17	0.23	7	7	7	7.00	0.03	7
6	-0.41	-0.17	0.09	-0.16	0.13	-0.17	7	7	7	7.04	0.24	7
7	-0.27	0.09	0.49	0.10	0.19	0.09	3	4	14	6.14	3.90	3
8	-0.07	0.36	0.87	0.38	0.25	0.32	11	11	11	11.01	0.15	11
	σ_μ^2 estimates											
all	1492.00	3185.20	6073.90	3456.40	1328.00	2843.50						

Note. σ_μ^2 is the between-person variance, Sub = Subject, Param = Parameter, Mdn = Median, M = Mean, Mo = Mode, L95 = 2.5th percentile, U95 = 97.5th percentile,

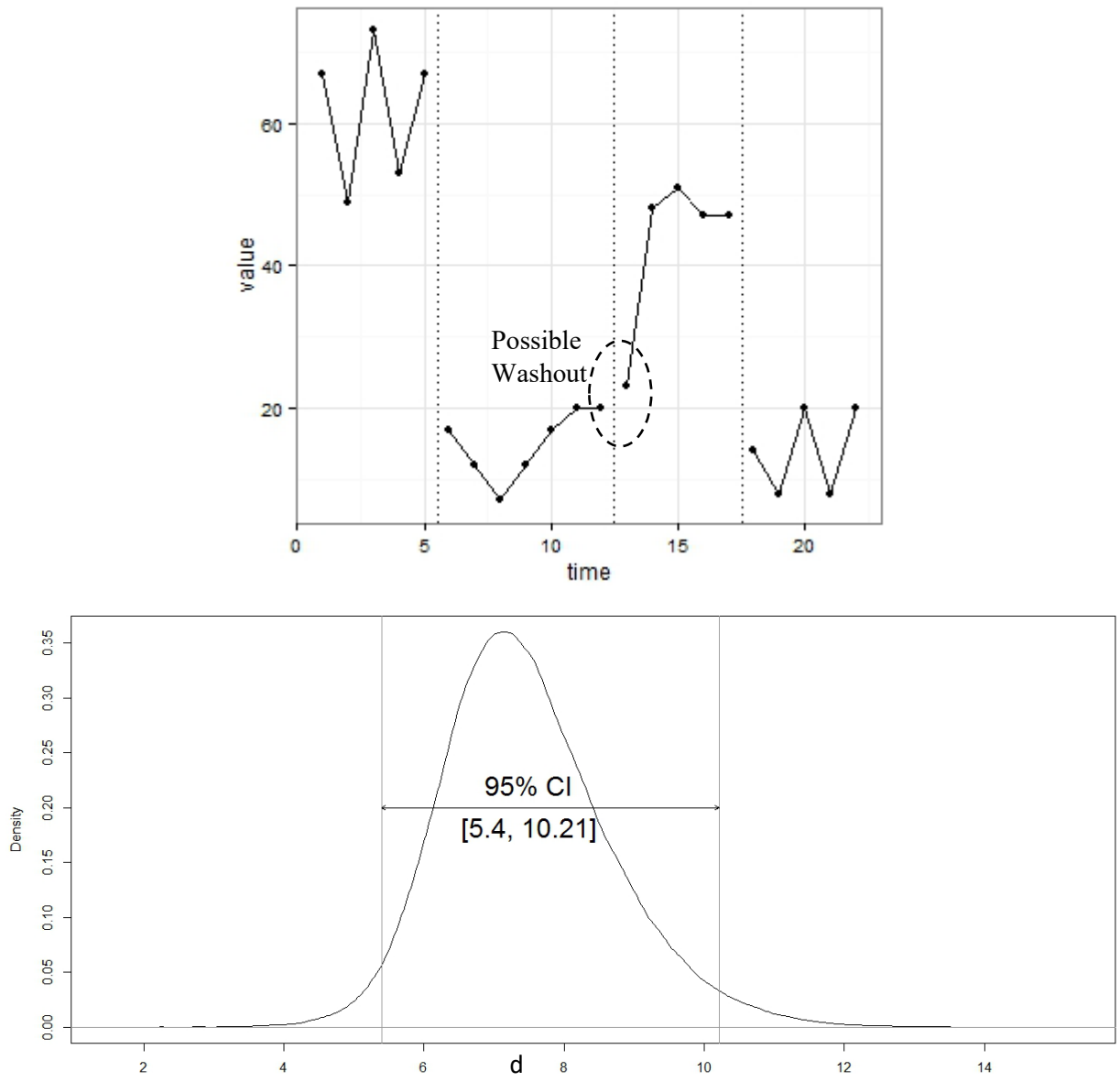


Figure 1. Dan's challenging behavior data and posterior distribution of effect size from Neely et al. (2013)

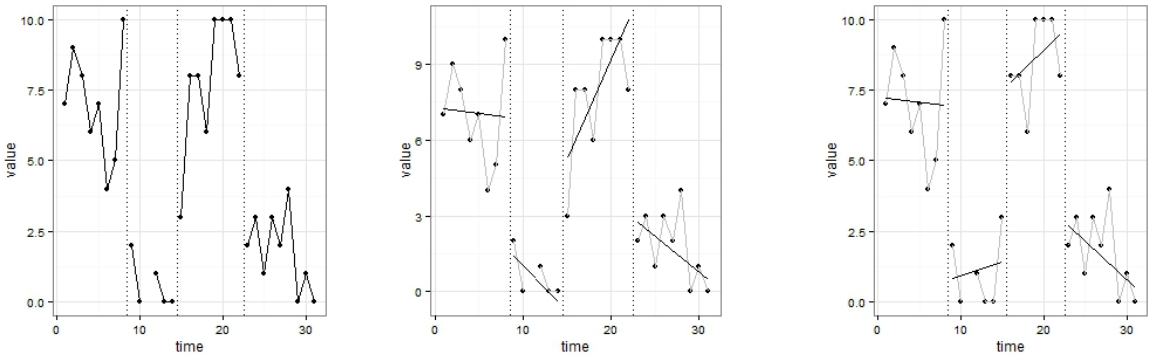
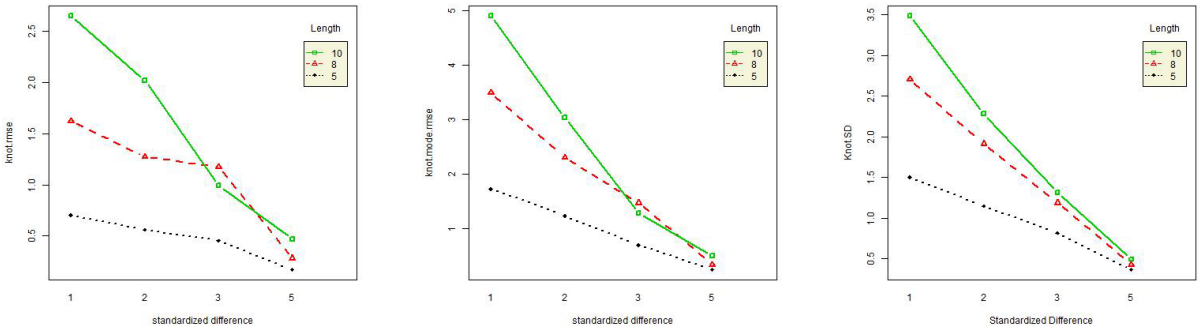


Figure 2. ABAB design data from Lambert et al. (2006)



RMSE vs Mean RMSE vs Mode RMSE vs MPSD

Figure 3. Interaction plots of RMSEs of mean and mode, and MPSD

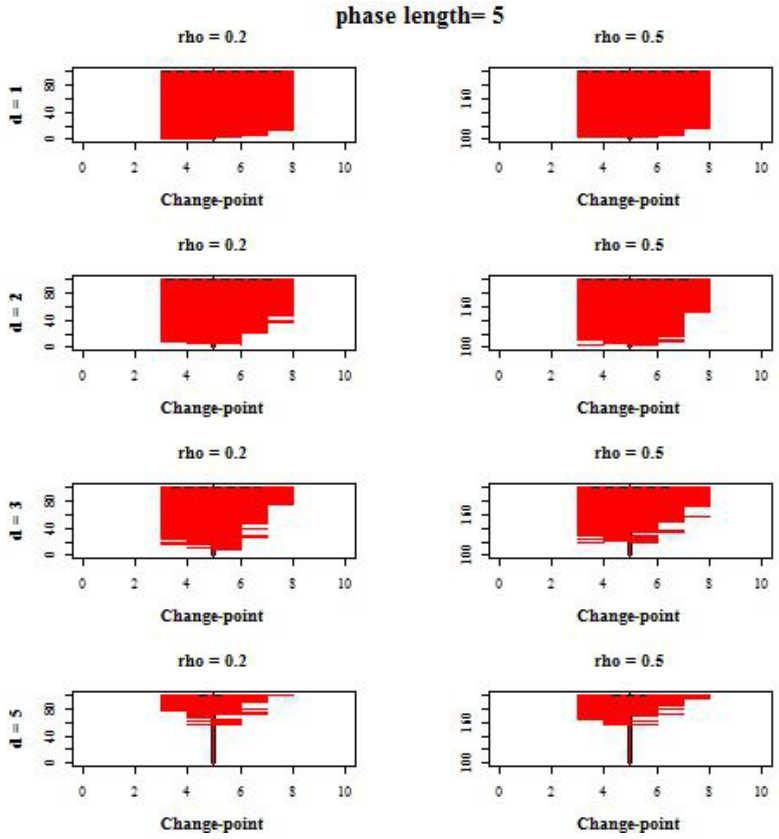
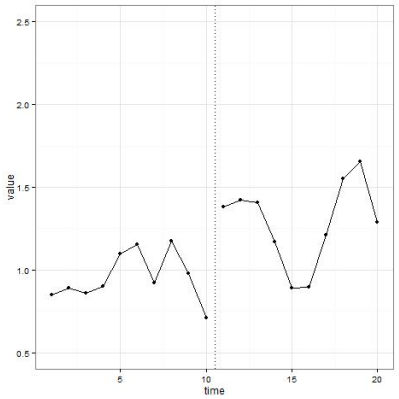
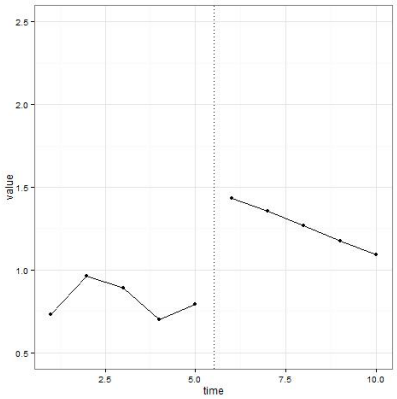


Figure 4. 95% credibility intervals for phase length = 5 as a function of effect size d and autocorrelation ρ

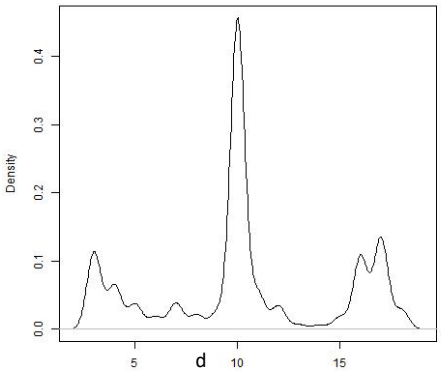
Large posterior SD



Small posterior SD



Case 1: $d = 1, \rho = 0.5$



Case 2: $d = 2, \rho = 0.5$

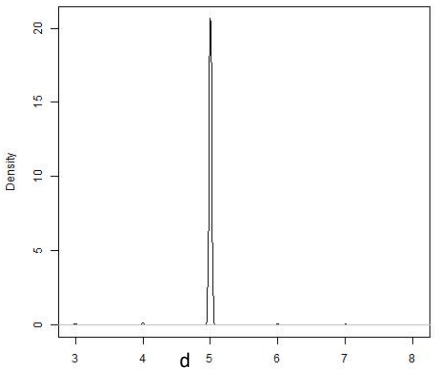


Figure 5. Plots and change-point posteriors of two cases

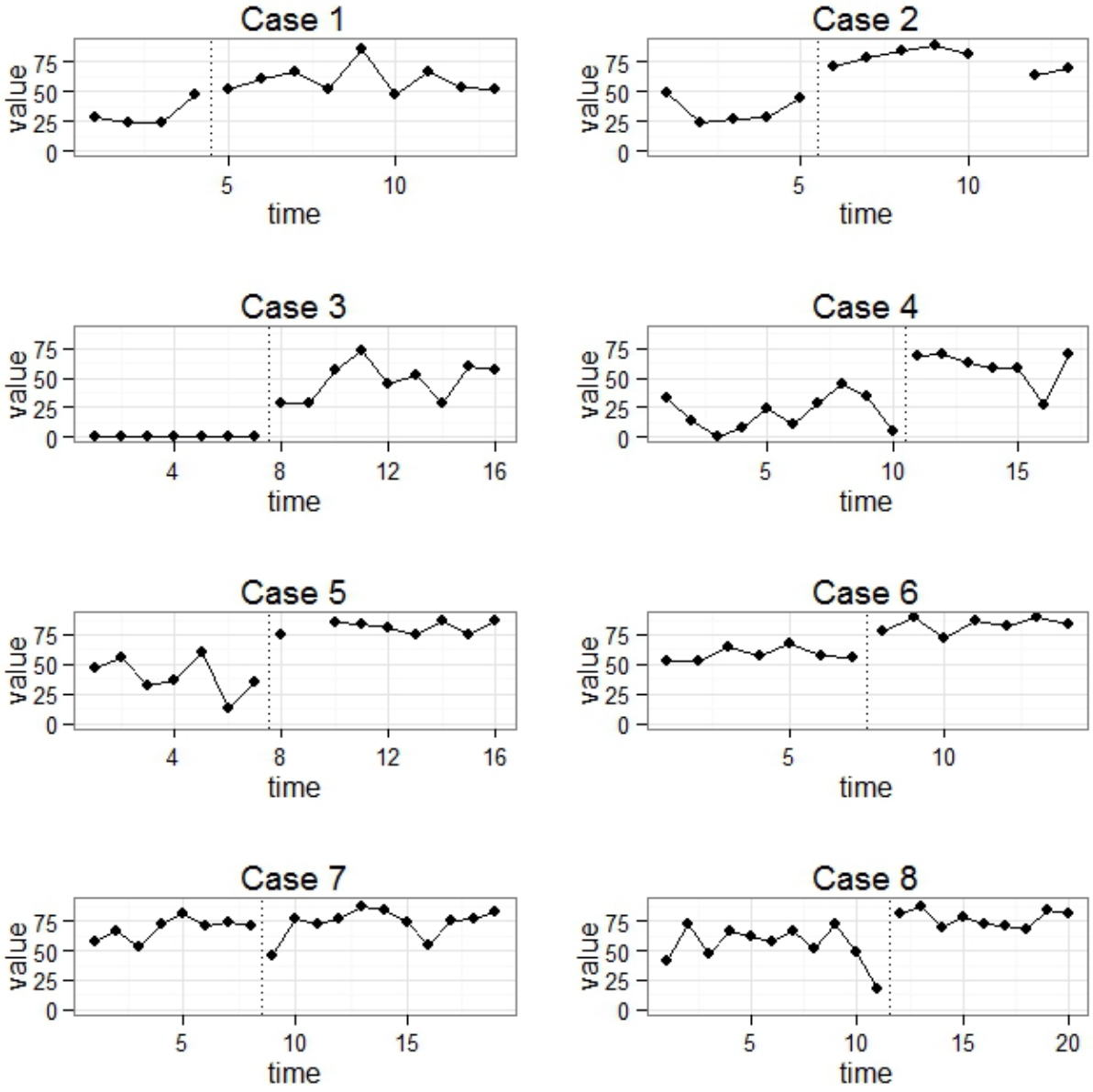


Figure 6. Multiple baseline data from Laski et al. (1988)

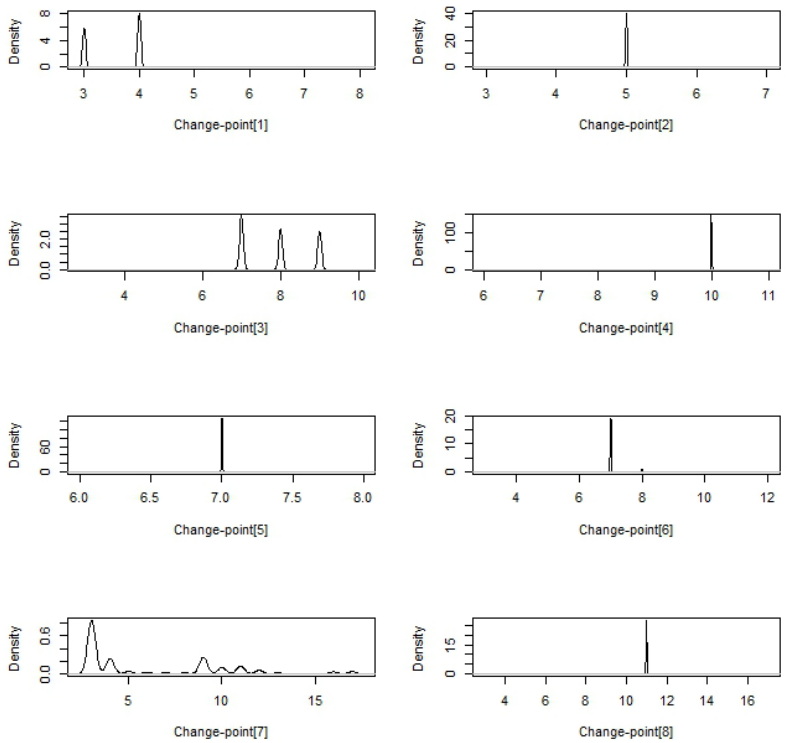


Figure 7a. Posterior densities of change-points for Laski data

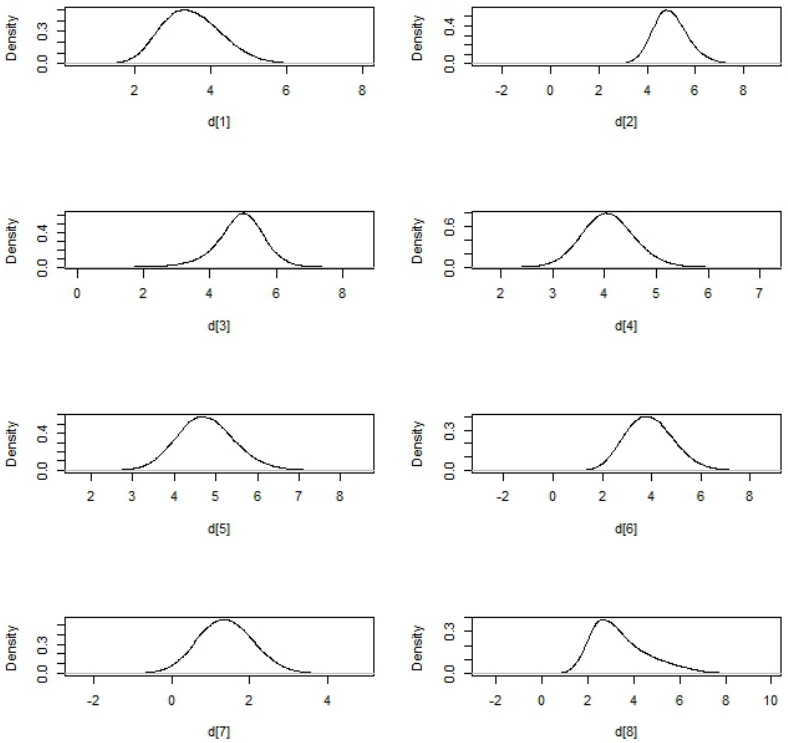


Figure 7b. Posterior densities of *d* for Laski data