Investigating Immediacy in Multiple Phase-change Single Case Experimental Designs

Using a Variational Bayesian Unknown Change-Points Model

Prathiba Natesan Batley
University of North Texas, Denton, TX, USA

Tom Minka
Microsoft Research Cambridge, Cambridge, UK

Larry Vernon Hedges
Northwestern University, Evanston, IL, USA

Abstract

Immediacy is one of the necessary criteria to show strong evidence of treatment effect in single case experimental designs (SCEDs). With the exception of Natesan and Hedges (2017) no inferential statistical tool has been used to demonstrate or quantify it until now. We investigate and quantify immediacy by treating the change-points between the baseline and treatment phases as unknown. We extend Natesan and Hedges' work to multiple phase-change (e.g. ABAB) designs using a Variational Bayesian (VB) unknown change-points model. VB was used instead of Markov chain Monte Carlo methods (MCMC) because MCMC cannot be used effectively to determine multiple change-points. Combined and individual probabilities of correctly estimating the change-points were used as indicators of accuracy of the algorithm. Unlike MCMC in the Natesan and Hedges' (2017) study, VB was able to recover the change-points with high accuracy even for short time-series and in only a fraction of the time for all time-series lengths. We illustrate the algorithm with 13 real datasets. Advantages of the unknown change-points approach, Bayesian, and Variational Bayesian estimation for SCEDs are discussed.

Investigating Immediacy in Multiple Phase-change Single Case Experimental Designs

Using a Variational Bayesian Unknown Change-Points Model

Single case experimental designs (SCEDs) are a form of interrupted time series design where observations on a single subject (i.e. a single child, patient, or sampling unit) are measured repeatedly during a baseline phase and at least one intervention or treatment phase. They are widely used 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). President Obama's State of the Union 2015 address emphasized personalized (precision) medicine initiatives. Subsequently, the National Institutes of Health established the precision medicine initiative cohort program. Based on credibility of evidence, the Oxford centre for evidence-based medicine ranked randomized *n-of-1* trial evidence as level-1 evidence for treatment decision purposes (Howick et al., 2011). Given the increased interest in SCEDs, several agencies are setting methodological standards for SCEDs to ensure high quality causal inferences (e.g. American Speech-Language-Hearing Association, 2004; Cook et al., 2014; Kratochwill et al. 2013). The goal of most analyses in SCEDs is to show that the observations in the treatment phase are only a function of the treatment and that no alternate explanations exist.

In SCEDs, no treatment is administered during the baseline phase and a treatment is administered during the treatment phase. The baseline and treatment phases are referred to as phases A and B, respectively. Visual analysis of a plot of observations over time is the most common SCED data analysis technique (Kratochwill et al. 2013). In SCEDs, evidence of treatment effect is demonstrated by establishing: (a) a stable pattern among the observations within each phase, (b) a difference in patterns of observations across phases, and (c) a change in the patterns of observations immediately following introduction or removal of treatment, which

is called immediacy. Establishing treatment effect in SCEDs also consists of showing three demonstrations of treatment effect along with absence of non-effects. Absence of non-effects is established by: (a) documenting the consistency of level, trend, and variability within phases, (b) documenting immediacy, proportion of overlap of observations across phases, and difference in the observed and predicted patterns of the observations, and (c) examining anomalies and external factors. Although establishing treatment effect using visual analysis alone may be easy for some data, this is not always the case. Difficulties may arise when: (a) immediacy is not possible due to the nature of the treatment, (b) errors are autocorrelated, and (c) the experimental effect size is not striking.

Immediacy is one of the criteria for demonstrating strong evidence of treatment effect (Kratochwill et al., 2013). A rapid change in the observations across phases indicates an immediate effect. An immediate effect provides convincing inference that change in the outcome measure was due to manipulation of the independent variable. When the treatment lacks immediacy, the researcher may not know whether the observations were solely a function of the treatment. Even though one knows when the administration of the treatment started or ended, one does not know when the treatment effect started or ended. Effect sizes computed by ignoring the lack of immediacy underestimate the treatment effect. Lack of immediacy in treatments that are expected to have immediacy compromises the internal validity of the design. However, no clear guidelines exist about decisively concluding the presence of immediacy.

Improvement over the existing approach

With the exception of Natesan and Hedges' Bayesian unknown change-point model (BUCP, 2017), immediacy is usually established by computing the change in the mean or median levels between the last 3-5 observations in a phase and the first 3-5 observations in the

subsequent phase. In contrast, BUCP uses all data points to establish immediacy. Change-points are the time-points in the SCED where there is a change that is discerned by functionally different relationships between time and the outcome variable in the difference phases. In this confirmatory approach, BUCP is used to investigate and quantify immediacy in two-phase designs. This model confirms if a change has taken place where it is designed to happen. In the presence of immediacy and a clear treatment effect, the change-point posterior mode would be accurately estimated and the posterior standard deviation would be very small. Natesan and Hedges showed how analyzing the posterior of the change-point could be used to investigate and quantify immediacy in AB and multiple baseline designs (MBDs).

However, AB designs cannot provide strong evidence of causality because they cannot be used to show three demonstrations of treatment effect, unless they are extended to multiple subjects as MBDs or replicated within a subject as multiple phase-change designs. Two-phase designs with only one participant have poor control of threats to internal validity because they do not account for other possible confounding effects that may explain treatment effect. A commonly used SCED is a multiple phase-change design (e.g. the ABAB design). Multiple phase-change designs contain at least two change-points. However, estimating multiple unknown change-points with small sample data with autocorrelation becomes quickly complicated. As a result, the popular Gibbs sampling algorithm will not compute efficiently for multiple phase-change designs. This is because the change-points in a multiple change-point model are simulated one at a time from full conditional distributions rather than from the joint distribution (for more details about issues with this approach refer Chib, 1998). Because the draws from the posterior are not iid, the uncertainty in estimating a previous change-point affects the uncertainty in estimating subsequent change-points. As a result, wide credibility intervals will be obtained

for change-points. This approach renders the logic of confirming the presence of immediacy through investigating the standard errors of change points unusable. Therefore, there is a need to extend Natesan and Hedges' confirmatory approach of quantifying immediacy to multiple phase-change designs such as the ABAB, ABA'B, and ABCD designs using an alternate method.

The ABAB design is a commonly used multiple phase-change reversal/withdrawal design. Assume that the target variable is problem behavior of a child with Autism. Effective demonstration of experimental effect would be indicated by a clear difference in the functional relationships between the scores across the phases. Specifically, the child's scores within the baseline phase (A1) would be larger than his scores in the treatment phase (B1). In an ABAB design, the treatment is withdrawn (A2) after B1 and then reintroduced (B2). Clear increase in problem behavior from B1 to A2 followed by a decrease in problem behavior from A2 to B2 are two additional distinct demonstrations of experimental effect. ABAB-type designs are referred to as withdrawal designs. In some designs, the treatment may not be completely withdrawn but replaced by another treatment, or serve as baseline for the next target behavior (changing criterion design).

However, not all treatments are aimed at reverting to baseline behavior following withdrawal. Some treatments, especially in health, may aim for an improvement in the second baseline phase rather than reversal to the original baseline's range of values. That is, studies may aim for a treatment effect that lasts even after the treatment phase is complete (Tate et al. 2013). Examples include improving communication skills, anger management techniques, and remediating gait dysfunction. Such a design would be an ABA'B design. Similarly, there are ABCD designs where the four phases may differ in their treatments.

Purpose

The purpose of the present study is to examine the performance of a variational Bayesian unknown change-points model in investigating and quantifying immediacy in multiple phase-change SCEDs. The first part of the study uses simulation to investigate the performance of the model for various data conditions. To this end, variational Bayesian (VB) method was used to estimate the parameters of an unknown change-point model with autocorrelated errors and 4 phases for commonly occurring data conditions in SCED research. The number of time-points per phase, autocorrelations, and standardized mean difference effect sizes were varied to simulate different data conditions that would mimic real life SCED data (Shadish & Sullivan, 2011). The goal of the present study was two-fold: (1) to test the feasibility of variational Bayesian in estimating the parameters of the 4-phase unknown change-point model with autocorrelations and (2) to identify the conditions that would be required to estimate the parameters of the said model so that SCED researchers can use the model to investigate experimental effects in SCEDs.

In the second part, the feasibility of these models to real data is tested by fitting the model to 13 datasets from 6 ABAB studies published in the last five years. These allow researchers and practitioners to understand the efficiency of the algorithm under various data conditions and illustrate how the method can be applied to real data. We use VB estimation to overcome the challenges of the Gibbs sampler. The proposed method extends and differs from the one presented by Natesan and Hedges (2017) in three ways: (a) it estimates multiple change-points instead of 1 change-point, (b) it evaluates the accuracy of the change-points using probabilities rather than width of credibility intervals, and (c) it uses VB instead of MCMC to estimate the parameters. Although the width of credibility intervals can be obtained from the probabilities, we

chose to use probabilities of obtaining the correct combination of change-points as a diagnostic for our study. This is because we are interested in the correct estimate of the combination of change-points and not a single change-point alone. To our knowledge, this is the first inferential statistical method that can confirm immediacy in multiple phase-change designs. This study is also the first of its kind to apply VB to SCEDs.

Quantitative Challenges in SCEDS

SCED data are often autocorrelated and have only a few observations per case. For instance, 45.3% of the studies reviewed by Shadish and Sullivan (2011) had 5 or fewer points per phase. The presence of autocorrelation in SCEDs: (a) is impossible to detect through visual analysis (Kazdin, 2011), (b) increases Type I errors (Matyas & Greenwood, 1990), and (c) is associated with low interrater reliabilities (Brossart, Parker, Olson, & Mahadevan, 2006). Therefore, quantitative methods for SCEDs have been gaining momentum in the recent years (e.g. Hedges, Pustejovsky, & Shadish, 2012, 2013; Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2013; Shadish, et al, 2015). However, autocorrelated errors violate the independence of observations assumption of most parametric and non-parametric statistics and result in biased estimates. Frequentist estimates of autocorrelation are negatively biased and have larger sampling errors for samples with less than 50 observations (Shadish, Rindskopf, Hedges, & Sullivan, 2013). Therefore, the researcher has to depend on large sample methods such as the maximum likelihood (ML). However, SCED sample sizes are too small to work well with ML.

With the exception of Natesan and Hedges (2017), all quantitative developments in SCEDs assume that the observed variable truly belongs to the phase it is designed to belong. However, this may not always be the case, especially when latency is expected. Latency happens when a treatment takes time to take effect when administered or to stop taking effect when

removed. Although latency is not desirable in SCEDs, it may be expected in some cases due to the nature of the treatment and the outcome variable. For instance, a child diagnosed with Autism may not respond immediately to a certain therapy or a drug may take some time to excrete completely out of the human body. In such cases, one must ascertain gradual and/or delayed effects or account for these with appropriate analytic strategies to untangle these from long-term treatment effects (Duan, Kravitz, & Schmid, 2013). Our method can be used to evaluate immediacy, which is one aspect of the quantifying SCED findings via transparent, objective, and replicable procedures.

Need for the Present Study

Several types of SCEDs are used in practice. Maggin et al. (2011) reported that withdrawal designs such as ABAB are the most commonly used designs in SCEDs. The Institute of Education Sciences (IES) What Works Clearinghouse (WWC) Pilot Standards for single-case designs (Kratochwill et al., 2013) recommended showing three different demonstrations of the experimental effect at three different points in time within a single case or across different cases for successful demonstration of experimental effect. The single-case reporting guideline in behavioural interventions (SCRIBE) SCED reporting guidelines (Tate et al., 2016) also reiterated the advantages of multiple phase-change designs over two-phase designs. We can expect to see an increase in the use of multiple phase-change designs. Therefore, there is an immediate need for an inferential tool to investigate and quantify immediacy in multiple phase-change designs.

The proposed method can be applied in all cases of four phase designs because it assumes that the four phases are not related to each other. Subsequently, the parameters of the four phases (e.g. intercepts, slopes) are estimated independent of each other in the proposed algorithm. Therefore, the model in the present study can be applied to any SCED with four phases.

Establishing immediacy is an important criterion for demonstrating strong evidence of treatment effect in SCEDs (Kratochwill et al., 2013). This is because the researcher has more evidence to make a case that a change in the dependent variable is initiated by a change in the treatment condition. In addition to confirming immediacy, there is a need to quantify the evidence that confirms immediacy. According to AERA guidelines, the presence of an effect must be accompanied by an index of uncertainty of that effect (AERA, 2006). The proposed method can be a valuable addition to the SCED researchers' toolkit because it both identifies and quantifies immediacy. Additionally, this quantification of immediacy sheds light on how reliable the slope and intercept estimates of the phases are. For instance, when there is lack of clear immediacy, slopes and intercepts will have wider credible intervals.

The Unknown Change-points Model

In an SCED with clear treatment effect, the dependent variable is expected to be a function of the phase to which it belongs (Natesan & Hedges, 2017). Therefore, in the presence of immediacy and treatment effect, a change in phase is reflected by the change in the function that maps the observation to its corresponding phase. In the present study, an intercept-only model is fitted to each phase where the boundary between the phases is assumed unknown a priori. The data define the change-points between the phases. Treatment effect is indicated based on the proximity between the estimated and true values of the change-points. By allowing the data to speak for themselves, this confirmatory approach investigates and quantifies the presence of immediacy or delayed effects in SCEDs.

Several approaches to change-point models have been proposed in the last few decades. A least squares estimation approach (Bai, 1994, 1997), Bayesian analysis of Poisson distributed data (Raftery & Akman, 1986), Bayesian online change-point detection (Adams & McKay,

2007) extensions of BUCP to multiple unknown change-points using hidden Markov models, genetic algorithms, and annealing stochastic approximation, respectively by Chib (1998), Jann (2000), Jeong and Kim (2013), and Kim and Cheon (2011) are some examples. BUCPs have been applied in ecology (Thomson et al., 2010), marine biology (Durban & Pitman, 2011), hydrometeorology (Perreault, Bernier, Bobée, & Parent, 2000), signal processing (Punskaya et al., 2002), 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. Barry and Hartigan (1993), and Carlin, Gelfand, and Smith (1992) provide good background materials for Bayesian change-point problems.

Bayesian Estimation

Bayesian methods often outperform classical methods in particularly small samples, allow more probabilistic interpretation of statistics than classical methods, and can more easily accommodate model complexities such as using distributions that reflect the scale of the observed variable, modeling autocorrelations, and representing hierarchical data structure. Bayesian estimation works well with small sample data because it does not depend on asymptotic or large sample theory (Ansari & Jedidi, 2000; Gelman, Carlin, Stern, & Rubin, 2004). This makes Bayesian particularly advantageous for SCEDs. Bayesian estimates a probability distribution for each parameter. This posterior distribution can be used to compute any summary statistic for the parameter of interest and its 95% highest density interval (HDI) can be directly interpreted as having 95% probability of containing the true value (Lynch, 2007). Posteriors of change-points with high probability mass at several time-points indicate weak evidence of treatment effect. Bayesian estimates of autocorrelation have the advantage of being more accurate than frequentist estimates. Frequentist confidence intervals of autocorrelation have

11

undercoverage (Shadish et al., 2013). Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin (2013) provide a comprehensive discussion of Bayesian methods.

Variational Bayesian

Variational Bayesian (VB) is a type of Bayesian inference that is computationally more efficient than MCMC (e.g. Natesan, Nandakumar, Minka, & Rubright, 2016). For models having smooth and unimodal posterior distributions, the cost of Bayesian inference can be significantly reduced by making analytic approximations. VB (Beal & Ghahramani, 2003) inference is one such approach that consists of fitting a simple approximating family (such as Gaussian or Gamma) to the posterior distribution by minimizing Kullback-Leibler (KL) divergence. KL divergence is a non-symmetric measure of the difference between the distributions (Kullback & Leibler, 1951). Computationally, the VB procedure visits each random variable in the model and incrementally improves its posterior approximation, repeating this for several iterations. Each step is similar in complexity to Gibbs sampling but instead of drawing thousands of samples, variational inference typically sweeps through the model a few dozen times before convergence. Beal and Ghahramani (2003) and Bishop (2006) are some good background materials in VB. The Infer.NET software program (Minka, Winn, Guiver, & Knowles, 2012) provides several options for performing KL divergence minimization. When compared to MCMC, which provides a technique to approximate the exact posterior using a set of samples, VB provides a locally-optimal, exact analytical solution to the posterior.

Model and Notation

A continuous, normally distributed dependent variable with no trend (slope) is considered in the simulation study. This can be extended to models with trend by estimating the slopes as

shown in the study 2 section which illustrates the algorithm with real data. This framework can be adapted for different variable and distribution types by modifying equations 1, 7, and 8. Consider the observed value at time y_t, which follows a normal distribution with the mean of μ_t which is the expected value of the target behavior at time t and standard deviation of σ_ε as shown in equation 1.

$$y_t \sim norm(\mu_t, \sigma_\varepsilon^2). \tag{1}$$

In SCEDs the errors of the data are typically considered to be lag-1 autocorrelated (Huitema, 1985; Huitema & McKean, 1998, 2000). It is of interest to note that although autocorrelations of errors are modeled in SCED data they are never interpreted and are considered only as a nuisance parameter. Consider an ABAB design with 4 phases, that is, 3 phase changes at times $i_1, i_2, and\ i_3$, and a total of T time-points. The predicted value $\hat{y}_t$ at time point t can be modeled as,

$$\hat{y}_t = \rho r_{t-1} + \beta_{0p}, \tag{2}$$

where ρ is the autocorrelation, r_{t-1} is the residual at time $t-1$, and β_{0p} is the intercept of the linear regression model for phase p the time-point t belongs to. If phc_t represents the phase change at a given time-point t, that is, it equals 1 if t denotes the start of a new phase and 0 otherwise, the residual is modeled for P phases as:

$$r_{t-1} = \begin{cases} g_p, if\ phc_t = 1 \\ y_{t-1} - \hat{y}_{t-1}, if\ phc_t = 0 \end{cases}, where \tag{3}$$

$$g_p \sim norm\left(0, \frac{\sigma_\varepsilon^2}{(1-\rho^2)}\right); p = 1, .., P. \tag{4}$$

Equation 4 gives the relationship between autocorrelation, the variance of the random error (σ_ε^2) and the white noise created by the combination of autocorrelation and random error (e.g. Natesan & Hedges, 2017). The intercepts β_{0p} at time t are modeled as:

$$\beta_{0p} = \begin{cases} \beta_{01}, if \ t \leq i_1 \\ \beta_{0(j+1)}, if \ i_j < t \ \leq i_{j+1}; j = 1,2 \\ \beta_{0P}, otherwise \end{cases} \tag{5}$$

The following priors were used:

$$\sigma_\varepsilon \sim gamma(1,1) \tag{6}$$

$$\rho \sim norm(0,100); bounded \ by \ (-1,1) \tag{7}$$

$$\beta_{0p} \sim norm(0,100); p = 1, \dots, P \tag{8}$$

The priors are weakly informative. All change-points were specified to be sampled from discrete uniform distributions ranging from time-point 1 to T and then ordered. In equation 9, cat stands for categorical distribution. The term i_p would indicate one of the P-1 change-points based on the probabilities given in equation 10. In this case, there is equal probability of the time-point being any of the change-points.

$$i_p \sim cat(pi); p = 1, \dots, P - 1 \tag{9}$$

$$pi = \left(\frac{1}{T}, \frac{1}{T}, \dots, \frac{1}{T}\right); length(pi) = T \tag{10}$$

$$i_1 < i_2 < i_3 \tag{11}$$

The ordered constraints in equation 11 are modeled as loops in Infer.NET. When the estimated values of $i_1, i_2, and \ i_3$ are the same as the true value, that is, when the correct combination of these parameters has the highest probability in the marginal posterior, we can say that we have some evidence of immediacy. Note that immediacy can only be indicated and not confirmed.

Method

To perform Bayesian inference in this model, the method starts by enumerating all possible combinations of change-point locations. The number of change-points is assumed known. For each combination, variational inference was run to get approximate posteriors for the parameters $(\rho, \sigma_\varepsilon^2, \beta_{0p})$ as well as an approximate value for the marginal likelihood. The Infer.NET software library (Minka et al., 2014) was used to perform these computations. VB implementation in Infer.NET is similar to the implementation of syntax in popular Bayesian software programs such as JAGS. Because the prior is uniform, the marginal likelihood is proportional to the marginal probability of the change-point locations. Therefore, we take the set of marginal likelihoods just computed and normalize them to get the marginal posterior over the change-point locations. These probabilities of the combinations of change-points obtained and the marginal posteriors can be used to make inferences about immediacy. If the correct combination of change-points is estimated as having the maximum probability, sufficient evidence can be deemed for immediacy.

Study 1: Simulation

Three phase lengths ($l = 5, 8, and\ 10$ time-points per phase), four standardized mean difference values ($d = 1, 2, 3, 5$), and four autocorrelation values ($\rho = 0, 0.2, 0.5, 0.8$) were simulated based on commonly occurring values in single case studies (Maggin, O'Keefe, & Johnson, 2011; Shadish & Sullivan, 2011). Although some autocorrelation is always present in SCED data, the autocorrelation value of 0 was added as a reference point to see how presence of autocorrelation affects accuracy of estimates. One hundred datasets were generated for each combination of conditions resulting in 4800 datasets in a fully crossed $3 \times 4 \times 4$ design. The

intercepts of the baseline phases were set to 0 and the treatment phases computed as σd, where the standard deviation within each phase $\sigma = 0.2$.

Diagnostics and Interpretation

The accuracy of the algorithm was examined using average probability weighted norm (APWN), average probability of obtaining the accurate combination ($\bar{p}$), and the combinations with highest average probabilities. If in the rth replication $\boldsymbol{p}_r$ is the vector of true values of change-points and $\boldsymbol{q}_r$ is the vector of change-point estimates, and $\Pr(\boldsymbol{q}_r)$ is the probability of $\boldsymbol{q}_r$ where within a given replication $\sum_{q_r} \Pr(\boldsymbol{q}_r) = 1$, then the APWN over R replications is given by:

$$APWN = \frac{1}{R}\sum_{r=1}^{R}\sum_{q_r} \Pr(\boldsymbol{q}_r)\|\boldsymbol{q}_r - \boldsymbol{p}_r\| \tag{12}$$

In formula 12, $\|\boldsymbol{q}_r - \boldsymbol{p}_r\|$ refers to the norm or the positive length between the two vectors. APWN weights the distance between the estimated and true values with the probability of obtaining them, while $\bar{p}$ (pbar) considers only the probability of correctly estimating the true values. APWN is an unbounded statistic, is a combination of distance and probability, and depends on the phase length, in addition to the other quantities in formula 12. The probability of the estimates along with the distance between the true and the estimated values are both important in determining the accuracy of the estimate, which is the reason for using APWN. If the probability of a combination is high and the discrepancy between the combination of estimates and the true values ($\|\boldsymbol{q}_r - \boldsymbol{p}_r\|$) is high, APWN will be large indicating poor estimates. If either the probability or the discrepancy is low, APWN will be smaller indicating better estimates. Thus, smaller values of APWN are more desirable, indicating more accurate estimates. The fraction of datasets in a given condition where the true values of the change-points were

estimated was also computed as a diagnostic. Root mean squared errors and average standard errors were also computed for the autocorrelations and intercept estimates. Thirteen ANOVAs were conducted with APWN, $\bar{p}$, and fraction correct, and RMSEs and average standard errors of intercepts and autocorrelations as the dependent variables, respectively. The independent variables in the ANOVAs were $d, l, and \rho$. The eta-squared effect sizes from the ANOVAs were used to understand the pattern of the accuracy of the estimates with change in the data conditions. Coverage rates could not be computed because VB estimates standard errors and not the credible intervals.

Results

The results of the simulation and the eta-squared effect sizes from independent ANOVAs for the change-point diagnostics are reported in Tables 1 and 2, respectively. The correct estimates of change-points had the maximum average probability for all conditions except when the standardized mean difference $d = 1$ for autocorrelations less than 0.8. Even in cases where an incorrect combination was estimated with maximum average probability, the probability of the correct combination was very close to the maximum average probability ($difference <$ .005). In all other cases, the correct change-point combinations had the highest average probability of being estimated. APWN decreased very slightly with $\rho \ and \ l$ but decreased substantially with increase in d. The average probability $\bar{p}$ and fraction of datasets with correct change-point estimates both increased with increase in d (i.e. standardized mean difference) but decreased with ρ (i.e. autocorrelation). Although phase length did not have a high effect size on APWN, $\bar{p}$, and fraction correct, there seems to be an interaction between phase length and d. For phase lengths of 8 or more, $\bar{p}$ and fraction correct increased drastically when d increased from 3 to 5 while APWN decreased with increase in d. Standardized mean difference explained the

17

maximum variation in APWN, $\bar{p}$, and fraction correct as seen in Table 2. Variation in autocorrelation explained medium variance in APWN, while all other main and two-way interaction effects were low-medium on APWN and almost negligible on $\bar{p}$ and fraction correct. These trends are clear from the interaction plots in Figure 1. Although the model seems to require a large effect size, the correct change-point combination was identified for most data conditions indicating that immediacy can be identified even for datasets with effect sizes as small as 2 standard deviations apart.

<div align="center">TABLES 1-4 AND FIGURE 1</div>

RMSEs and average standard errors of the autocorrelation and intercept estimates by condition are shown in Table 3. The corresponding effect sizes from the ANOVAs are shown in Table 4. Variance in RMSE of autocorrelations was explained the most by the true value of autocorrelation, phase length, and their interaction as shown in Figure 1. Autocorrelation RMSE decreased with increase in phase length and decrease in true autocorrelation value. Average standard error of autocorrelation, and of all intercept estimates decreased with increase in phase length as expected. Although RMSEs of the intercepts was explained by most of the data conditions, all these RMSEs varied only to the second decimal place. Therefore, these are not interpreted.

Study 2: Real data applications

We analyzed 13 datasets from 6 ABAB design studies published between the years 2012 and 2015 (Allen, Baker, Nuernberger, & Vargo, 2013; Lin & Chang, 2014; Neely, Rispoli, Camargo, Davis, & Boles, 2013; Shih, et al., 2014; Shih, C. -H., Chiang, & Shih, C. T., 2015; Shih, C.-H., Wang, Chang, & Kung, 2012). The dependent variables in these studies included: challenging behavior, percentage of intervals of academic engagement, number of occurrences of

correct moves, rate of collaborative walking, number of correct responses per session, and rate of collaborative pointing for children with Autism, and problem behavior for a woman with dual diagnosis of moderate intellectual disability and schizoaffective disorder. The number of time-points per phase ranged from 3 to 21 across these studies. Of the 13 datasets analyzed, 11 sets of change-point estimates had 98-100% probability of being estimated accurately. The estimation took less than 3 seconds for each dataset. The set of change-point estimates from one dataset (Shih et al., 2012) had 92.98% accuracy, while a different set of change-points were estimated to be 99.4% accurate for another dataset (Nealy et al., 2013). We discuss these probabilities, their interpretation, and the reasons for the inaccurate estimates. We fit the same unknown change-points model with intercepts for all datasets considering the dependent variable to be intervally-scaled. Upon visual inspection, we decided that this is a reasonable model to be fit for all data for illustration purposes.

Consider the data for subject 1 from Shih et al, (2012) as shown in Figure 2. The true values of change-points are at 8, 26, and 33. The corresponding estimates for all parameters are shown in Table 5. The only probable values for change-points were estimated to be at 8, 26, and 33. The pattern is also clear from the data. The standard error of the intercepts range from 0.289 to 0.498. The posterior densities of the change-points had only a single probable value. Both change-point estimates and their accuracy (probability) show support for immediacy.

FIGURE 2, TABLE 5 HERE

Consider the estimates of subject 2 from the same study whose plot and posteriors are shown in Figures 3a and 3b, respectively. Although the probability value for the change-point estimates and the individual probability for 8 being the first change-point is only 0.93, the probability of every other change-point estimate being the first change-point is less than 2%.

Similarly, the probability of 26 and 33 being the second and third change-points is greater than 99.5% as seen in Table 5. This shows strong evidence of immediacy and that the changes indeed occurred at the time-points they were designed to take place. The standard errors of intercepts range from 0.22-0.37 and that of autocorrelation is 0.11.

Let us consider the change-point estimates that were inaccurate (Table 5). The data and the posterior plots are given in Figures 4a and 4b, respectively. The probability of the set of change-points and each change-point is more than 99%. The first and the third change-points were accurately estimated but the second data-point was estimated to be 13 instead of 12. Closer inspection of the data plot sheds light on why this is the case. The observed values at time points 12 and 13 are much closer (3% difference) than the values at 13 and 14 (15% difference). There is a possible washout effect from the treatment in this case and therefore indication of lack of immediacy. That is, the treatment effect possibly lingered on even after the treatment was stopped.

FIGURE 4 HERE

Discussion

The present study demonstrates how the variational Bayesian unknown change-points model could be used to investigate and quantify immediacy in SCED data for multiple phase-change designs. VB estimates in the present study were better than the MCMC estimates in Natesan and Hedges' because unlike in the latter study phase length had a very small effect on the accuracy of the estimates in our study. This means even conservatively researchers need only five data-points per phase to estimate the change-points with accuracy as long as the standardized mean difference effect size was at least 2. However, accuracy does improve with increase in effect size. This is a major improvement over the Natesan and Hedges (2017) study

which required eight data points per phase but for a simpler model with only 2 phases (AB model). There are several advantages to the method we presented: (a) It is an inferential statistical tool that can be used to investigate and quantify immediacy in multiple phase-change designs, which are commonly-used SCEDs; (b) Bayesian methods allow the user to examine the shape of the posterior distribution of the change-points. These help provide a clearer evaluation of the quality of the estimates because the researcher has more information about the probability of each value the change-point can take; (c) By treating the change-points as unknown, the researcher can remain objective, allow the data to speak for themselves, and confirm the presence of treatment effect. This is also useful to identify presence and length of delayed effects; (d) The model can be modified to accommodate other distribution types, data types, functional relationships between time and the dependent variable, and additional explanatory variables; (e) By using VB, the amount of time taken to estimation is reduced to a fraction of what would be taken by commonly used Bayesian software programs such as BUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009) and JAGS (Plummer, 2003); and (f) The model is a one-stop shop for simultaneously estimating all the associated parameters in the model such as the intercepts and slopes of the phases, effect sizes, and autocorrelations.

Although the simulation results show high accuracy when the data have standardized difference between the phases of 5 or higher, we believe this model shows promise in terms of still estimating the correct combination of change-points even for effect sizes of 2 with a phase length of 5. We acknowledge that an effect size of 5 places a heavy demand on the data. However, requiring only 5 data-points per phase for a complex model such as this is still a significant advantage.

Strong treatment effect is indicated when the change-point posterior mode is accurately estimated, and the distribution is narrow and clearly unimodal. Inaccurate estimates, as in Dan's case, may indicate possible washout or lag effect. This can be a valuable tool for treatments and interventions that are effective but do not have immediacy. A posterior distribution with large variance but surrounding the true value indicates weak treatment effect and possible lack of immediacy. A posterior distribution with large variance but with most of the probability mass concentrated around the true value indicates moderate to strong treatment effect depending on the shape of the posterior.

To our knowledge, the present study is the first of its kind to apply VB estimation to SCEDs. VB generally trades in accuracy for speed. However, the approximation by VB was negligible. For two-phase models, Natesan and Hedges (2017) recommended a standardized mean difference of at least 3 to detect immediacy. For practical purposes, the correct combination is still chosen even with standardized mean difference as little as 2. This means practitioners can still be able to show some evidence of immediacy for smaller effect sizes. However, given the complexity of the model, larger effect sizes do improve accuracy. The multiple phase-change model considered in the present study has 3 latent change-point variables and is therefore more complicated than the bi-phasic model which has 1 latent change-point variable. The bi-phasic Bayesian model took about 3 minutes to run whereas the VB model took up to 3 seconds to run. These data indicate that VB is a viable and efficient method to be used in SCEDs.

However, we do not declare that the model presented in this study is the ideal solution for SCED data analysis. First, the model places heavy demand on the data such as requiring an effect size of at least 5 for high accuracy of estimation. Second, it is unclear how the priors have

affected the estimates. Although we used weakly informative priors, the estimates might be improved by using more informative priors especially in small sample cases such as these. This is an avenue for further research. Third, we were unable to compare these estimates with MCMC because of convergence issues in the latter. Therefore, we are unable to declare these estimates as being superior to MCMC, although we can note that this model is more feasible than its exact MCMC counterpart. Finally, there is a learning curve associated with implementing VB for estimating the parameters of an unknown change-points model. Still readers can access the program by downloading the zip folder in the supplemental material. The program is currently set to run 4 phase models for continuous data. Extensions of the script to accommodate these other models will be released pending investigation of the corresponding algorithms.

References

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

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.

Beal, M. J., & Ghahramani, Z. (2003). The Variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. David, D. Heckerman, A. F. M. Smith, M. West (Eds.), Bayesian statistics 7 (pp. 453-464). Oxford, UK: Oxford University Press.

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.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

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.

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

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.

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

Duan, N., Kravitz, R. L., & Schmid, C. H. (2013). Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *Journal of Clinical Epidemiology, 66,* S21-S28.

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

Ferron, J. & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education, 75,* 66-81.

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.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian data analysis* (2nd ed.). London: Chapman & Hall.

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, (2007). *Bayesian Methods: A Social and Behavioral Sciences Approach* (2nd ed.). London, UK: Chapman and Hall/CRC.

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.

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.

Howick, J., Chalmers, I., Glasziou, P., Greenhaigh, T., Heneghan, C., Liberati, A., & Thornton, H. (2011). *The 2011 Oxford CEBM Evidence Table (Introductory Document)*. Oxford: Oxford Centre for Evidence-Based Medicine. Available from: http://www.cebm.net/index.aspx?o=5653

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.

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.

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., 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_SCED.pdf.

Kratochwill, T.R., Hitchcock, J., Horner, R.H., Levin, J.R., Odom, S.L., Rindskopf, D.M, and 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.

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

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

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, C.-Y. & Chang, Y.-M. (2014). Increase in physical activities in kindergarten children with cerebral palsy by employing MaKey–MaKey-based task systems. *Research in Developmental Disabilities, 35*, 1963-69.

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.

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.

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

Minka, T., Winn, J., Guiver, J., & Knowles, D. (2012). Infer.NET 2.5 [Software]. Cambridge, UK: Microsoft Research Cambridge.

Minka, T., Winn, J., Guiver, J., Webster, S., Zaykov, Y., Yangel, B., Spengler, A. & Bronskill, J. (2014). *Infer.NET 2.6*. Microsoft Research Cambridge, UK. http://research.microsoft.com/infernet

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

Natesan, P. & Hedges, L. V. (2017). Bayesian unknown change-point models to investigate immediacy in single case designs. *Psychological Methods*.

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

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.

Punskaya, E., Andrieu, C., Doucet, A., & Fitzgerald, W. J. (2002). Bayesian curve fitting using MCMC with applications to signal segmentation. *IEEE Transactions on Signal Processing, 50*, 747–758.

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

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

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. (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.

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.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item risk of bias in n-of-1 trials (RoBiNT) scale. *Neuropsychological Rehabilitation, 23*, 619-638.

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., . . . Wilson, B. (2016). The single-case reporting guideline in behavioural interventions (SCRIBE) 2016 statement. *Archives of Scientific Psychology*, *4*, 1-9.

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.

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.

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
Diagnostics of Change-points from the Simulation

l	ρ	d	APWN	$\bar{p}$	$max\_\bar{p}$**	Highest prob locs*			f(correct)
						$e(i_1)$	$e(i_2)$	$e(i_3)$	
5	0	1	6.148	0.003◊	0.003◊	5	11	13	0.02
		2	5.127	0.04					0.21
		3	3.213	0.222					0.56
		5	0.424	0.81					0.90
	0.2	1	6.079	0.003	0.004	8	12	15	0.03
		2	5.112	0.046					0.13
		3	3.341	0.218					0.46
		5	0.507	0.798					0.89
	0.5	1	5.856	0.007	0.008	5	8	10	0.02
		2	5.011	0.054					0.10
		3	3.643	0.196					0.36
		5	0.844	0.728					0.85
	0.8	1	5.253	0.023					0.04
		2	4.825	0.08					0.14
		3	4.081	0.169					0.30
		5	1.79	0.56					0.67
8	0	1	8.407	0.002	0.002	14	15	24	0.01
		2	5.064	0.063					0.24
		3	1.766	0.328					0.58
		5	0.147	0.867					0.96
	0.2	1	8.579	0.002	0.003	3	17	24	0.01
		2	5.815	0.055					0.20
		3	2.627	0.298					0.53
		5	0.206	0.844					0.95
	0.5	1	8.56	0.002	0.008	8	16	20	0.01
		2	6.819	0.044					0.11
		3	4.232	0.225					0.47
		5	0.674	0.754					0.84
	0.8	1	7.805	0.02					0.04
		2	7.085	0.078					0.10
		3	5.841	0.188					0.27
		5	2.897	0.481					0.57
10	0	1	8.616	0.001	0.002	10	25	30	0.00
		2	4.274	0.078					0.20
		3	1.164	0.412					0.56
		5	0.087	0.921					0.94
	0.2	1	9.161	0.001	0.003	10	25	30	0.01

	2	5.755	0.059					0.20
	3	1.976	0.369					0.49
	5	0.136	0.887					0.93
0.5	1	9.212	0.004	0.006	10	20	28	0.01
	2	7.606	0.048					0.11
	3	4.522	0.25					0.39
	5	0.685	0.795					0.88
0.8	1	9.85	0.025					0.05
	2	9.16	0.076					0.12
	3	7.654	0.165					0.23
	5	3.317	0.568					0.68

Note: $max\_\bar{p}$ is the average probability of the combination with the highest probability. * Only incorrect combinations are shown, blanks refer to accurately estimated combinations. **$max\_\bar{p} = \bar{p}$ for correct combinations and therefore not shown. $\Diamond\bar{p} = 0.0027; max\_\bar{p} = 0.0032$. f(correct) represents the fraction of datasets with correct change-point estimates for a given condition

Table 2

Eta-squared of Change-point Diagnostics in Percentages from the ANOVAs

Factor	APWN	$\bar{p}$	fraction(correct)
l	3.84	0.35	0.02
ρ	7.36	1.96	3.71
d	76.19	93.33	92.65
$\rho \times d$	4.01	3.59	3.06
$l \times \rho$	3.85	0.20	0.18
$l \times d$	3.64	0.33	0.14

Table 3:
RMSE and Average SE of Parameter Estimates from Simulation

True values			RMSE					Average SE				
l	ρ	d	ρ	β_{01}	β_{02}	β_{03}	β_{04}	ρ	β_{01}	β_{02}	β_{03}	β_{04}
5	0	1	0.17	0.03	0.06	0.03	0.01	0.19	0.13	0.09	0.09	0.07
		2	0.14	0.02	0.02	0.05	0.01	0.19	0.09	0.08	0.09	0.08
		3	0.14	0.01	0.02	0.01	0.02	0.19	0.09	0.08	0.09	0.08
		5	0.13	0.00	0.02	0.00	0.02	0.20	0.09	0.09	0.09	0.09
	0.2	1	0.32	0.02	0.06	0.02	0.01	0.19	0.13	0.08	0.09	0.07
		2	0.27	0.03	0.08	0.00	0.02	0.19	0.12	0.08	0.09	0.08
		3	0.28	0.02	0.02	0.01	0.02	0.20	0.09	0.08	0.08	0.08
		5	0.27	0.00	0.02	0.01	0.02	0.20	0.09	0.09	0.09	0.09
	0.5	1	0.54	0.00	0.11	0.03	0.01	0.19	0.13	0.08	0.08	0.07
		2	0.52	0.02	0.10	0.02	0.02	0.20	0.13	0.08	0.08	0.08
		3	0.50	0.01	0.06	0.02	0.03	0.20	0.10	0.08	0.09	0.08
		5	0.48	0.01	0.03	0.00	0.04	0.20	0.09	0.09	0.09	0.09
	0.8	1	0.78	0.01	0.14	0.00	0.03	0.20	0.16	0.08	0.08	0.07
		2	0.75	0.00	0.16	0.02	0.05	0.20	0.14	0.08	0.08	0.08
		3	0.74	0.01	0.13	0.05	0.04	0.20	0.13	0.08	0.09	0.08
		5	0.71	0.00	0.11	0.02	0.05	0.20	0.12	0.09	0.09	0.09
8	0	1	0.12	0.02	0.03	0.01	0.03	0.16	0.09	0.07	0.08	0.06
		2	0.09	0.02	0.03	0.00	0.03	0.16	0.07	0.07	0.07	0.07
		3	0.10	0.01	0.01	0.01	0.01	0.16	0.07	0.07	0.07	0.07
		5	0.09	0.00	0.01	0.00	0.01	0.16	0.07	0.07	0.07	0.07
	0.2	1	0.24	0.02	0.04	0.01	0.03	0.16	0.08	0.07	0.07	0.06
		2	0.20	0.01	0.05	0.00	0.03	0.16	0.08	0.07	0.07	0.07
		3	0.19	0.00	0.01	0.01	0.01	0.16	0.07	0.07	0.07	0.07
		5	0.18	0.00	0.01	0.00	0.02	0.16	0.07	0.07	0.07	0.07
	0.5	1	0.44	0.00	0.09	0.03	0.04	0.16	0.09	0.07	0.07	0.07
		2	0.38	0.03	0.08	0.03	0.05	0.16	0.09	0.07	0.07	0.07
		3	0.35	0.01	0.03	0.01	0.04	0.16	0.09	0.07	0.07	0.07
		5	0.33	0.00	0.02	0.00	0.03	0.16	0.07	0.07	0.07	0.07
	0.8	1	0.58	0.00	0.15	0.06	0.06	0.16	0.11	0.07	0.07	0.07
		2	0.55	0.01	0.15	0.06	0.06	0.16	0.10	0.07	0.07	0.07
		3	0.52	0.02	0.15	0.06	0.07	0.16	0.11	0.07	0.07	0.06
		5	0.47	0.01	0.12	0.01	0.09	0.15	0.10	0.07	0.07	0.06
10	0	1	0.11	0.02	0.01	0.01	0.01	0.15	0.09	0.07	0.07	0.06
		2	0.09	0.01	0.00	0.01	0.00	0.15	0.07	0.06	0.06	0.06
		3	0.09	0.00	0.01	0.00	0.01	0.15	0.06	0.06	0.06	0.06
		5	0.08	0.00	0.01	0.00	0.01	0.15	0.06	0.06	0.06	0.06
	0.2	1	0.21	0.02	0.01	0.00	0.01	0.15	0.08	0.07	0.07	0.06
		2	0.17	0.01	0.02	0.01	0.01	0.15	0.08	0.06	0.06	0.06
		3	0.17	0.00	0.01	0.01	0.01	0.15	0.06	0.06	0.06	0.06

	5	0.16	0.00	0.01	0.00	0.01	0.15	0.06	0.06	0.06	0.06
0.5	1	0.36	0.01	0.08	0.06	0.01	0.14	0.09	0.07	0.06	0.07
	2	0.32	0.03	0.06	0.01	0.00	0.14	0.09	0.06	0.06	0.06
	3	0.30	0.01	0.02	0.00	0.00	0.14	0.07	0.06	0.06	0.06
	5	0.27	0.00	0.02	0.01	0.02	0.14	0.06	0.06	0.06	0.06
0.8	1	0.49	0.03	0.11	0.03	0.01	0.14	0.08	0.07	0.06	0.06
	2	0.44	0.01	0.13	0.03	0.02	0.14	0.10	0.06	0.06	0.06
	3	0.44	0.02	0.14	0.02	0.00	0.14	0.10	0.06	0.06	0.06
	5	0.37	0.01	0.09	0.01	0.05	0.13	0.07	0.06	0.06	0.06

Table 4:
Eta-squared values in Percentages from ANOVAs of RMSEs and Average SEs

Factor	RMSE					Average SE				
	ρ	β_{01}	β_{02}	β_{03}	β_{04}	ρ	β_{01}	β_{02}	β_{03}	β_{04}
l	13.00		4.99		33.41	98.12	47.00	92.13	88.03	80.92
ρ	80.72		78.38	27.64	35.55		22.47			
d		34.02	9.45	16.16	3.20		16.34			5.17
$\rho \times d$		34.32	4.10	15.50	7.29		5.30			
$l \times \rho$	4.30	16.00		12.46	8.90					
$l \times d$					6.22		5.23			9.99

Note: Only values above 4% are reported.

Table 5

Estimates from Shih et al. (2012) and Neely et al. (2013)

|	Change-points			Prob	Autocorrelation		nP	Intercept 1		Intercept 2		Intercept 3		Intercept 4	
	1	2	3		Mean	SE		Mean	SE	Mean	SE	Mean	SE	Mean	SE
Sub 1	**8**	**26**	**33**	**1.000**	**0.789**	**0.089**	**0.375**	**0.174**	**0.464**	**16.720**	**0.305**	**1.312**	**0.498**	**18.820**	**0.289**
Sub 2	0	8	26	0.001	0.753	0.086	0.553	0.000	100.000	0.089	0.388	5.452	0.256	4.358	0.208
	0	8	33	0.000	0.694	0.095	0.525	0.000	100.000	0.094	0.409	3.991	0.228	5.911	0.255
	0	26	33	0.002	0.794	0.082	0.591	0.000	100.000	3.737	0.201	0.721	0.396	5.914	0.230
	7	26	33	0.000	0.505	0.110	0.623	0.114	0.433	5.048	0.260	0.659	0.433	5.908	0.254
	8	26	32	0.000	0.429	0.114	0.603	0.108	0.422	5.494	0.280	0.505	0.489	5.532	0.259
	8	**26**	**33**	**0.930**	**0.471**	**0.112**	**0.845**	**0.105**	**0.351**	**5.493**	**0.233**	**0.650**	**0.376**	**5.907**	**0.221**
	8	26	34	0.000	0.471	0.123	0.608	0.105	0.414	5.492	0.274	1.586	0.414	5.824	0.267
	8	26	35	0.001	0.491	0.113	0.641	0.104	0.400	5.492	0.265	1.689	0.377	6.000	0.265
	8	26	42	0.000	0.608	0.101	0.637	0.097	0.385	5.482	0.254	2.917	0.270	6.696	0.327
	8	27	33	0.000	0.356	0.127	0.584	0.111	0.437	5.128	0.282	0.572	0.506	5.905	0.275
	9	26	33	0.005	0.426	0.122	0.678	0.726	0.375	5.538	0.271	0.639	0.426	5.906	0.250
	10	26	33	0.008	0.414	0.118	0.691	0.871	0.353	5.688	0.278	0.637	0.423	5.906	0.249
	11	26	33	0.020	0.404	0.118	0.717	1.044	0.331	5.895	0.283	0.636	0.417	5.906	0.245
	12	26	33	0.007	0.418	0.119	0.690	1.314	0.321	6.017	0.297	0.638	0.423	5.906	0.248
	13	26	33	0.015	0.461	0.115	0.719	1.507	0.298	6.255	0.298	0.648	0.409	5.907	0.240
	14	26	33	0.001	0.547	0.115	0.650	1.906	0.293	6.128	0.317	0.665	0.418	5.908	0.244
	15	26	33	0.001	0.520	0.109	0.654	1.929	0.285	6.304	0.333	0.661	0.420	5.908	0.246
	16	26	33	0.002	0.516	0.109	0.671	2.046	0.273	6.582	0.346	0.660	0.416	5.908	0.243
	17	26	33	0.005	0.578	0.107	0.707	2.246	0.251	7.015	0.348	0.673	0.396	5.909	0.232
CB-Dan	5	12	17	0.005	-0.298	0.169	0.018	61.230	3.160	14.740	2.663	43.950	3.160	13.680	3.160
	5	13	16	0.000	-0.189	0.171	0.012	61.370	3.953	15.810	3.119	48.710	5.122	18.730	3.606
	5	**13**	**17**	**0.994**	**-0.352**	**0.171**	**0.032**	**61.170**	**2.361**	**15.710**	**1.857**	**48.340**	**2.648**	**13.630**	**2.361**
	5	13	18	0.000	-0.200	0.205	0.012	61.340	3.941	15.800	3.107	42.230	3.941	13.930	4.413
	5	14	17	0.000	-0.074	0.206	0.012	61.530	4.066	19.230	3.025	48.200	5.262	13.850	4.066
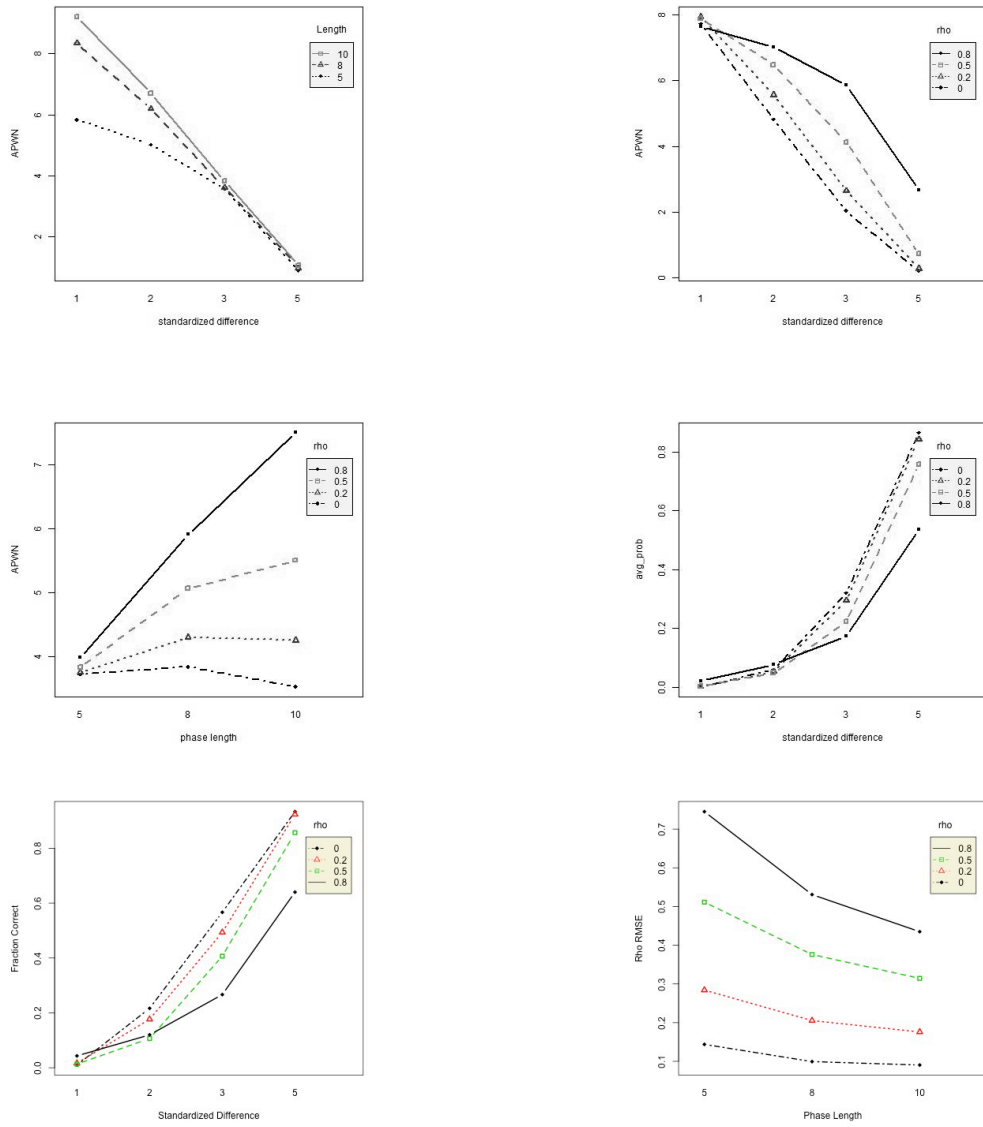
Figure 1: Interaction Plots of APWN, fraction correct, $\bar{p}$(avg_prob), and rho RMSE with $d, l, and \rho$
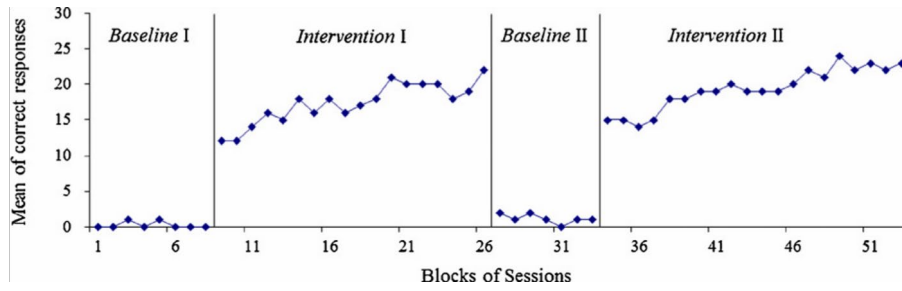
Figure 2: Data for Subject-1 from Shih et al. (2012)
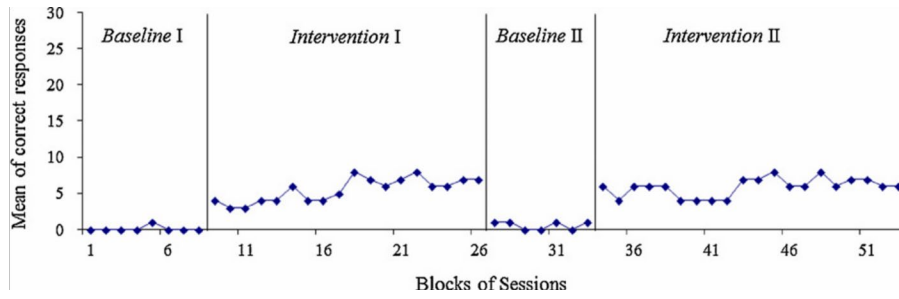
Figure 3a: Data from Subject-2 from Shih et al. (2012)
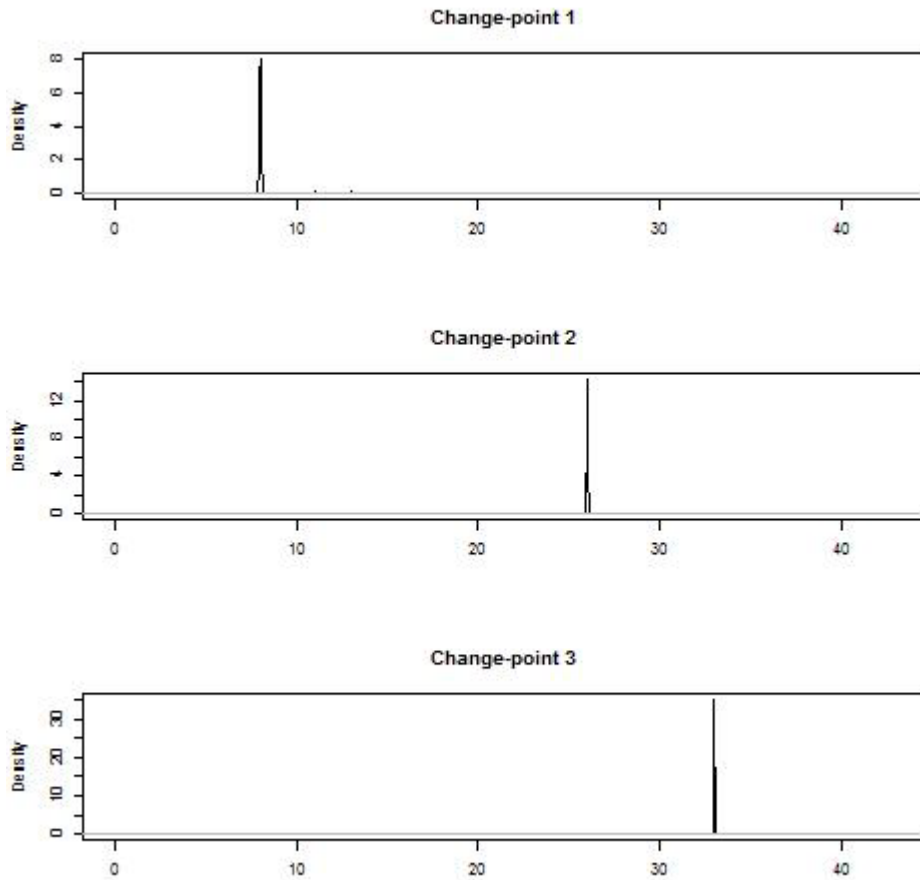
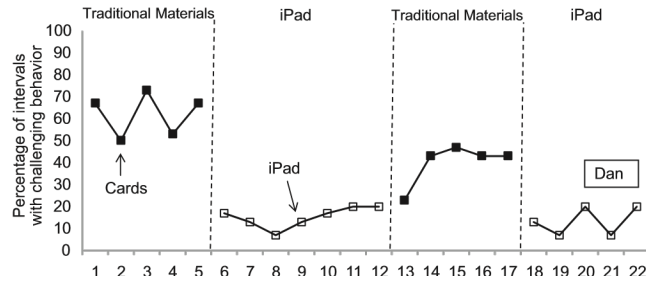Figure 3b: Posteriors of the change-points for Subject-2

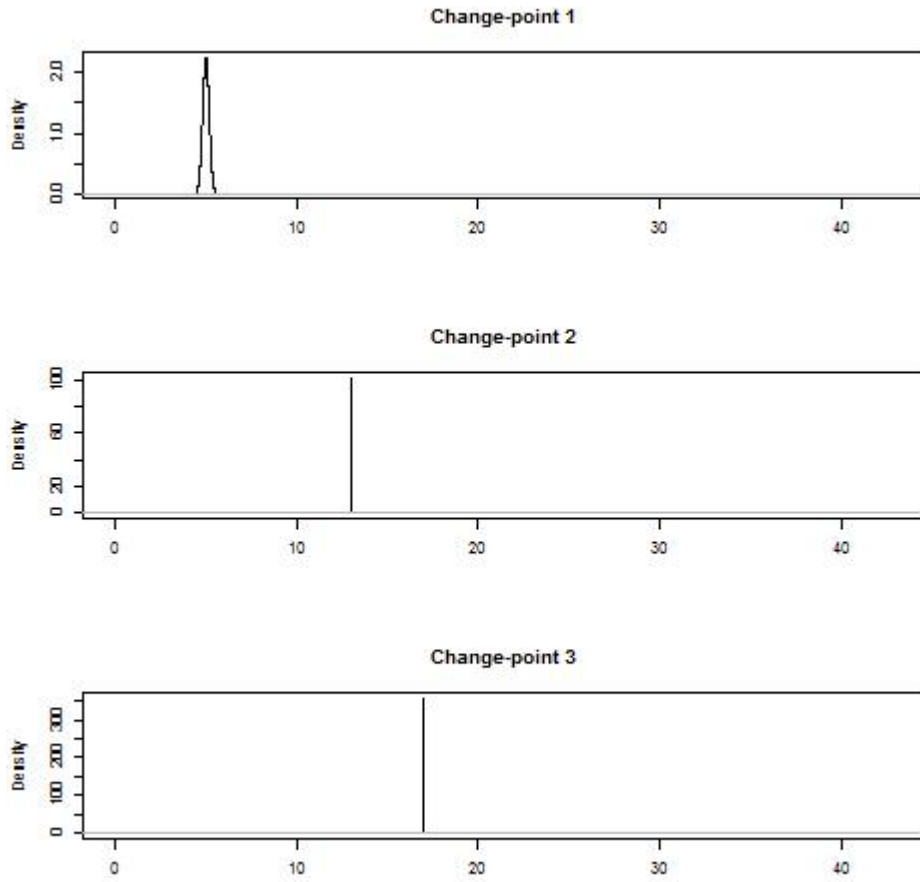Figure 4a: Data for Dan's challenging behavior from Neely et al. (2013)

Figure 4b: Posteriors of change points for Dan's challenging behavior from Neely et al. (2013)