Accurate Models vs Accurate Estimates: A Simulation Study of Bayesian Single-Case Experimental Designs

Prathiba Natesan Batley
Brunel University London, London, UK

Larry V. Hedges
Department of Statistics, Northwestern University, Evanston, IL, USA

This paper was presented at the Texas Universities' Educational Statisticians and Psychometricians Annual meeting in 2019 and at the Annual meeting of the American Educational Research Association in 2019.

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 Behavior Research Methods 2021

Abstract

Although statistical practices to evaluate intervention effects in SCEDs have gained prominence in the recent times, models are yet to incorporate and investigate all their analytic complexities. Most of these statistical models incorporate slopes and autocorrelations both of which contribute to trend in the data. The question that arises is whether in SCED data that show trend, there is indeterminacy between estimating slope and autocorrelation because both contribute to trend and the data have limited number of observations. Using Monte Carlo simulation, we compared the performance of four Bayesian change-point models: (a) intercepts only (IO), (b) slopes but no autocorrelations (SI), (c) autocorrelations but no slopes (NS), and (d) both autocorrelations and slopes (SA). Weakly informative priors were used to remain agnostic about the parameters. Coverage rates showed that for the SA model either the slope effect size or the autocorrelation credible interval almost always erroneously contained 0 and the Type II errors were prohibitively large. Considering the 0-coverage and coverage rates of slope effect size, intercept effect size, mean relative bias, and second phase intercept relative bias, the SI model outperformed all other models. Therefore, it is recommended that researchers favour the SI model over the other three models. Research studies that develop slope effect sizes for SCEDs should consider the performance of the statistic by taking into account coverage and 0-coverage rates. These helped uncover patterns that were not realized in other simulation studies. We underline the need for investigating the use of informative priors in SCEDs.

Keywords: Single-case designs; Markov Chain Monte Carlo (MCMC); Bayesian; interrupted time-series models.

Accurate Models vs Accurate Estimates: A Simulation Study of Bayesian Single-Case Experimental Designs

Single-case experimental designs (SCEDs) involve manipulating an independent variable by applying an intervention to evaluate intervention effects by repeated, systematic measurements of an outcome variable (Horner et al. 2005; Kratochwill & Levin, 2014). Thus, SCEDs are forms of interrupted time-series designs, which are often used to evaluate intervention effects in various fields ranging from 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). The importance and necessity of SCEDs in experimental designs where randomization is often impossible or inappropriate (e.g. low incidence disabilities, rare diseases, comorbid health conditions) has been discussed at length in SCED literature (e.g. Gast & Ledford, 2014; Kratochwill et al. 2010; Kratochwill & Levin, 2014; Shadish, 2014).

Often visual analyses are conducted to analyze SCED data. These analyses are supplemented with reporting phase means, medians, percentages, and effect sizes such as standardized mean differences or indices based on the amount of data overlap between phases (Parker, Hagan-Burke, & Vannest, 2007). Although visual analysis has definite advantages with analyzing SCED data, studies have shown that the presence of autocorrelation can confound the results of visual analysis. For instance, in data with autocorrelation, it is difficult to decompose patterns due to trends (slopes) versus patterns due to autocorrelated errors. Autocorrelation is almost impossible to detect by visual analysis alone (Kazdin, 2011; Thyer & Myers, 2011). The presence of autocorrelation increases Type I errors (Matyas & Greenwood, 1990) and decreases interrater reliabilities (Brossart, Parker, Olson, & Mahadevan, 2006) in visual analysis. In fact, Jones, Weinrott, and Vaughn (1978) found that in data with moderate-high autocorrelations, visual analysis results were reduced to nearly chance levels. Therefore, there is increasing emphasis for more objective methodologies for analyzing SCED data and determining causal inferences. Many organizations (American Speech-Language-Hearing Association, 2004; Cook, Buysse, Klingner, Landrum, McWilliam, Tankersley, & Test, 2014; Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, & Shadish, 2013) have worked on reaching professional consensus on the methodological standards for SCEDs. One such standard, the U.S. Department

of Education's What Works Clearinghouse Pilot Standards for single-case designs (Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, & Shadish, 2010), advocates that researchers evaluate the difference in levels, trends, and variability across phases in order to meet evidence standards for SCEDs. Therefore, it is somewhat common to see models with intercepts and slopes for each phase and the same autocorrelation for all phases being fitted to single case experimental designs (Solomon, 2013). Multilevel model for SCEDs is an example of one such model (e.g. Baek & Ferron, 2013; Ferron et al., 2009; Ferron, Farmer, & Owens, 2010; van den Noortgate, & Onghena, 2003a, 2003b).

However, it may not be wise to fit such complex models to short time-series data. We hypothesize that this is because it is difficult to separate how much of the trend in the data is due to autocorrelation and how much is due to slope (i.e. a continuous gain or fall in the outcome variable). When fitting complex models to small sample data that are commonly found in SCEDs we still do not know which parameters will be affected and to what extent they would be affected. The purpose of this simulation study is to investigate the performance of a two-phase interrupted time-series model with first-order autocorrelation in recovering the parameters of two-phase SCED data. We fit and compare four statistical models to SCED data with slopes and autocorrelations. The first model estimates slopes (commonly known as trends in SCED literature), intercepts (levels), and autocorrelations. The second model estimates only slopes and intercepts while assuming there is no autocorrelation. The third model estimates intercepts and autocorrelations assuming that any trend displayed by the data is due to autocorrelation and not slope. The fourth is the simplest model that estimates intercepts only and assumes that no trend is present, that is, there is no pattern due to autocorrelation or slope. We investigated which model best captures the data using diagnostics such as root mean squared errors (RMSEs) of the posterior means of slopes, intercepts, autocorrelations, and standard deviations; biases of slope and intercept effect sizes; coverage rates of the credible intervals (CI) of slopes, intercepts, and autocorrelations; and 0-coverage rates (that is, the percentage of CIs that contain 0) of slopes and autocorrelations. Bayesian estimation was used for all models because of its advantages with small sample data, especially SCEDs (e.g., Natesan Batley, 2020; Natesan Batley, Contractor, & Caldas, In Press; Natesan Batley, Minka, & Hedges, 2020; Natesan Batley, Shukla Mehta, & Hitchcock, 2020; Natesan, 2019; Natesan & Hedges, 2017; Rindskopf, 2014; Shadish, 2014;

Shadish, Rindskopf, Hedges, & Sullivan, 2013). Readers are directed to the aforementioned references for further discussion of the role and advantages of Bayesian in estimating SCEDs.

Literature review

In SCEDs, the intervention effect can manifest itself as change in level (Crosbie, 1995; Tryon, 1982) or change in trend (Crosbie, 1995; van den Noortgate & Onghena, 2003a, 2003b) or a combination of both (Baek & Ferron, 2013). Thus, there are many ways of detecting intervention effects in SCEDs. These may utilize single level or multilevel models. In the single level model framework, Campbell and Stanley (1966) and Mood (1950) recommended testing whether the first observation of the intervention phase lay in the confidence interval of the predicted or extrapolated value at that time-point assuming no intervention effect. If the true value of the first observation of the intervention phase lay in the confidence interval of the predicted value, a researcher may conclude that there was no intervention effect whereas intervention effect may be tentatively inferred if otherwise. However, this procedure is weak because it does not make use of all data-points (Campbell & Stanley, 1966). Therefore, another option is to compare the intercepts and slopes of the regression lines of both phases. If the intercepts and slopes are the same, the null hypothesis that the treatment is not effective cannot be rejected (Campbell, 1967). Algina and Swaminathan (1977) showed that the test statistic for testing the intervention effect in single-group quasi-experimental time-series designs for linear trends follows the F-distribution. However, this is confounded by autocorrelation because the measurements are obtained on an individual across time. Ignoring autocorrelations will lead to biased parameter and standard error estimates, which in turn, hinders the validity of statistical inferences (Pankratz, 1983).

All the aforementioned procedures ignore autocorrelation. Trend in data with autocorrelations leads to under or overestimated treatment effect sizes (West & Hepworth, 1991). The presence of autocorrelations biases error variances, confidence intervals, *t* values, and Type I error rates (Glass et al., 1975; Gottman, 1980; Gottman & Glass, 1978; McCain & McCleary, 1979). Gottman and Glass (1978) showed that the Type I error of a *t* test with alpha level = 0.05, when the autocorrelation is 0.5 is 0.2584. Similarly, Hibbs (1974) concluded that the Type I error rate is inflated by 265% when the autocorrelation is 0.7. Huitema, McKean, and McKnight (1999) showed that ordinary least squares estimates of slopes have higher Type I errors for larger

values of positive autocorrelation especially for large sample sizes. This is because the variance of the slope is underestimated in the presence of positive autocorrelations. This in turn, affects the slope change parameters whose error rates were unacceptably high for autocorrelations greater than 0.20. Finally, positive autocorrelations were associated with higher Type I error rates for estimates of slope change than for estimates of level change. Therefore, Huitema, McKean, and McKnight (1999) concluded that large sample theory overestimates the harmful effects of autocorrelation of Type I error in small samples.

Glass, Wilson, and Gottman (1972) adopted autoregressive (AR) and autoregressive integrated moving average (ARIMA) processes for testing intervention effects in time-series data. Simonton (1977) outlined a procedure for comparing the regression lines of an interrupted time-series model assuming first order autocorrelations. However, this procedure requires the number of individuals to be greater than the number of measurement occasions. Obviously, this requirement is almost impossible to fulfill in single-case experimental designs. Other researchers have suggested that a minimum of 50 observations is required to obtain sufficiently accurate estimates for a first-order autoregressive model (Box & Pierce, 1970; Glass, Willson, & Gottman, 1975; Ljung & Box, 1978).

Huitema and McKean (2000) studied the two-phase interrupted time-series model and recommended that individual slopes and intercepts be estimated for each phase. However, this study was conducted in the absence of autocorrelation. McKnight, McKean, and Huitema's (2000) double bootstrap method had a bias in autocorrelation estimate ranging from 0.018 to 0.2 for a time-series length of 20. The bias decreased with increase in time-series length. However, SCED data are often even shorter in length. In fact, in a systematic literature review by Shadish and Sullivan (2011) of the SCED articles published in 2008, excluding alternating treatments design, only 54.7% of the 563 articles had more than 5 data-points in the baseline phase. The median number of total data-points in 809 studies was 20 and 90.6% had fewer than 50 data-points in total. This leaves less than 25 data-points per phase if the designs only had two phases and even fewer data-points in designs with more than two phases. Approximately 70% of the studies had fewer than 30 data-points in all.

There has been considerable effort in developing methods and effect sizes for SCED data with trend (e.g. Allison & Gorman, 1993; Center, Skiba, & Casey, 1985-86; Gorman & Allison,

1997; White, Rusch, Kazdin, & Hartmann, 1989). Researchers van den Noortgate and Onghena (2003a) discussed procedures for meta-analyses with linear trends. Parker, Vannest, and Davis (2011) developed a method to control positive baseline trend within data non-overlap. However, this procedure ignores the presence of autocorrelation and capitalizes on overlap indices, which are known to have many drawbacks in addition to ignoring the distances between data-points and being sensitive to outliers. Parker, Vannest, Davis, and Sauber (2011) developed an effect size that combined trend and autocorrelation in SCED data. Cobb and Shadish (2015) and Sullivan, Shadish, and Steiner (2015) used semi-parametric regression models to analyze SCED data with linear and non-linear trends. Beretvas and Chung (2008) used the difference in R^2 (ΔR^2) as an index of effect size, which is an indicator of change in both intercepts and slopes for single case designs with trends. Their results showed that ΔR^2 has acceptable statistical properties only in the absence of autocorrelation and has poor performance in the presence of autocorrelation especially for few cases and few time-points. Specifically, for large values of autocorrelations, the Type I error rate, that is, rejecting the null that there is no intervention effect based on ΔR^2 was high. Whether this is because the autocorrelations ended up being estimated as slopes is unknown. Solanas, Manolov, and Onghenas' (2010) and Manolov and Solanas' (2009) model with slope and autocorrelations eliminates baseline trend from SCED data to estimate slope and level changes. However, they diagnosed the performance of the models only using biases, which do not shed light on whether the trend in the data was appropriately decomposed into autocorrelations and slopes. Without examining the interval estimates of slopes and autocorrelations, it is impossible to tell if there is an indeterminacy problem, that is, whether sometimes slope is estimated as autocorrelation and vice versa. This is important because autocorrelation is not considered as part of intervention effect whereas change in slopes is usually attributed to the intervention effect in SCEDs. In sum, although parametric approaches based on regression have great promise for meta-analysis of SCEDs, we are yet to know the full extent of their weaknesses and strengths.

Although the performance of models that estimate trends and autocorrelations have been conducted using the multilevel modeling (MLM) framework, the present study is a development over these studies because they had some limitations. For instance, Ferron, Farmer, and Owens (2010) studied MLM for multiple baseline designs and compared different approaches to show that estimates and coverage rates improved with phase length and effect size. Similarly, Ferron,

Bell, Hess, Rendina-Gobioff, and Hibbard (2009) showed that although the treatment effect estimates were relatively accurate in the presence of autocorrelation, the point estimates were biased. However, the aforementioned studies (i.e. Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Ferron, Farmer, & Owens 2010) are applicable to only to multiple baseline designs and required at least 8 participants for robust estimation of parameters. Moeyaert, Ugille, Ferron, Onghena, Heyvaert, Beretvas, & Van den Noortgate (2015) demonstrated multilevel meta-analysis of results from various types of SCEDs. Petit-Bois, Baek, Van den Noortgate, Beretvas, and Ferron (2016) conducted a simulation of meta-analysis of 10 or 30 studies and used sample sizes of 4 and 7. Thus, they had much larger data. Although Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate (2012) showed that MLM can be applied to datasets with as few as 4 participants, and a time-series of length 10 series length per SCED study, this still places greater burden on the researcher in terms of data collection. This is because a minimum of 3 participants and 5 data-points per phase are required to meet the WWC design standards. Fewer than 63% of the studies reviewed by Shadish and Sullivan (2011) had 20 or more data-points in total. What the present study solves is a much more basic problem when considering any multiphase design for one participant using the simpler, single level model where MLM is not applicable. The advantage of our approach is that the findings from our study are applicable to a wider set of SCEDs such as the ABAB design, changing criterion design, the multiple baseline design, or the alternating treatments design. Moreover, although some of the above-mentioned studies examined the coverage rates of autocorrelations, none of them examined 0-coverage of autocorrelation. Examining 0-coverage is important because it tells us if the estimated value of autocorrelation is incorrectly computed as 0 (i.e., being non-existent). On the contrary coverage rates only tell us if the true value is contained in the interval estimate.

In data that exhibit both slopes and autocorrelations, a model that neglects slope is expected to produce strongly autocorrelated residuals (Shadish, Rindskopf, & Hedges, 2008). This may be because the pattern in the data due to the slopes is estimated as the pattern in the data due to autocorrelation. Thus, Simonton (1979) questioned the specific advantages that accrue from augmented complexity in short time-series data. Huitema, McKean, and McKnight (1999) also seconded this opinion and asked if complex approaches are necessary while modeling dependency structure of observations in time-series designs. Specifically, the question remains as to whether it is prudent to fit models with slopes and intercepts that vary by phase and

an autocorrelation that is common to all phases for SCED data, which are short time-series data, let alone develop effect sizes, and multi-level models. Simpler models generally have greater statistical power and are simpler to interpret. However, the sensitivity of these models to violation of assumptions such as independence of observations needs to be studied further before they can be recommended for general use. This forms the impetus for the present study.

Models

A continuous, normally distributed dependent variable with slope and autocorrelation was considered as the outcome variable in the present study. Four single level Bayesian models were fitted to the data as shown in Table 1. These models varied based on whether slopes and autocorrelations were estimated in the model or not.

INSERT TABLE 1 ABOUT HERE

Model 1 (Intercepts, slopes and autocorrelations - SA model): The SA model estimates intercepts, slopes, and autocorrelations. Consider a SCED with two phases: baseline and treatment. Let the time-points in the baseline phase be $1, 2, \dots, t_b$ and in the treatment phase be $t_{b+1}, \dots, t_n$. 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 SA model and the serial dependency of the residual (e_t) can be expressed as,

$$\hat{y}_{pt} = \begin{cases} \beta_{11} + \beta_{21}t + \varepsilon + \rho e_{pt-1}, if\ t \le t_b \\ \beta_{12} + \beta_{22}(t - t_b) + \varepsilon + \rho e_{pt-1}, otherwise \end{cases} \text{ and} \quad (3)$$

$$e_{pt-1} = \rho e_{pt-2} + \varepsilon. \quad (4)$$

In equation 3, $\hat{y}_{pt}$ is the probability of the predicted value of the dependent variable at time t in phase p; β_{11} and β_{21} are the intercept and slope of the linear regression model for phase 1, respectively; β_{12} and β_{22} are the intercept and slope of the linear regression model for phase 2,

respectively; e_{pt} is the error at time t in phase p; ρ is the autocorrelation coefficient which stays the same across both phases; and ε is the independently distributed error. The standard deviations of $e, \varepsilon, and\ \rho$ are related as shown in equation 5.

$$\sigma_e = \frac{\sigma_\varepsilon}{\sqrt{1-\rho^2}}. \quad (5)$$

The intercept and slope β_{1p} and β_{2p} can be modeled as:

$$\beta_{ip} = \begin{cases} \beta_{i1}, if\ t \leq t_b \\ \beta_{i2}, otherwise, \end{cases} \quad (6)$$

where the terms refer to intercepts when $i = 1$ and slopes when $i = 2$. Intercept effect size ES_1 was defined as the standardized mean difference between the two phases as given in equation 7. Slope effect size ES_2 was defined as the difference between the estimated value at the mid-point of the intervention phase assuming and not assuming an intervention effect as shown in equation 8. If t_i is the number of time-points in the intervention phase, then:

$$ES_1 = \frac{\beta_{12} - \beta_{11}}{\sigma_\varepsilon} \text{ and} \quad (7)$$

$$ES_2 = \left(\beta_{12} + (t_b + \frac{t_i}{2})\beta_{22}\right) - \left(\beta_{11} + (t_b + \frac{t_i}{2})\beta_{21}\right). \quad (8)$$

Model 2 (Intercepts and autocorrelations but no slopes – NS model): The NS model assumes that any trend in the data is due to autocorrelation and not a slope parameter. Thus, the β_{21} and β_{22}terms are dropped or equal zero in equations 3 and 6 and only intercepts and autocorrelations are estimated. Thus, equation 3 becomes

$$\hat{y}_{pt} = \begin{cases} \beta_{11} + \varepsilon\ + \rho e_{pt-1}, if\ t \leq t_b \\ \beta_{12} + \varepsilon + \rho e_{pt-1}, otherwise \end{cases} \text{ and} \quad (9)$$

Model 3 (Slopes and intercepts but no autocorrelation – SI model): The slopes model assumes that the data are not autocorrelated. Thus, the ρ term vanishes or equals zero, thereby making equations 1-6 represent a simple piecewise regression model where only slopes and intercepts are estimated. The model becomes

$$\hat{y}_{pt} = \begin{cases} \beta_{11} + \beta_{21}t + \varepsilon, if\ t \leq t_b \\ \beta_{12} + \beta_{22}(t - t_b) + \varepsilon, otherwise. \end{cases} \quad (10)$$

Model 4 (Intercepts only and no autocorrelations or slopes – IO model): The IO model is the simplest of all models where no trend is assumed. Therefore, both slopes and autocorrelations are set to zero and only intercepts are estimated. Thus, the variability in the data is only due to random error as shown in equation 11.

$$\hat{y}_{pt} = \begin{cases} \beta_{11} + \varepsilon\,, if\ t \leq t_b \\ \beta_{12} + \varepsilon, otherwise \end{cases} \text{ and} \quad (11)$$

Priors: The priors were the same for the parameters that were common to all the models. We used weakly informative priors, which purposely include less information than what we actually have (Gelman & Jakulin, 2007). This allows the parameters of the priors to be estimated from the data rather than specifying them to have subjective information especially for small sample data like the ones in the present study (Efron & Morris, 1975; Gelman, 2006; James & Stein, 1960). The intercepts and slopes of both phases were independent of each other. The intercepts and slopes are drawn from normal distributions with hyperpriors in order to reduce the impact of the prior specification on the estimates as given in equations 12 - 15. The variances of the intercepts and slopes were independently drawn from gamma distributions with mode and standard deviations ranging uniformly between 0.01 and 1.[1] We chose the means of the intercepts to come from a distribution that uniformly ranged from zero to fifty because we assume that the mean of the dependent variable would not be outside these bounds based on the simulation design. Of course, practitioners should choose appropriate priors for their data depending on the scale of the outcome variable. For instance, the mean of an outcome variable such as the number of problem behaviors exhibited by a child with Autism during an observation period might range from zero to an upper limit that makes substantive sense. The means of the slope parameters were sampled from a unit normal distribution because this value indicated change in the outcome variable which might be positive or negative and included all plausible values for means of the slopes based on the simulation parameters.

$$\beta_{1p} \sim norm(\mu_{1p}, \sigma^2_{int}) \quad (12)$$

$$\beta_{2p} \sim norm(\mu_{2p}, \sigma^2_{slope}) \quad (13)$$

$$\mu_{1p} \sim unif(0, 50); p = 1, 2 \quad (14)$$

$$\mu_{2p} \sim norm(0, 1); p = 1, 2 \quad (15)$$

Other prior specifications were as follows:

$$\sigma_\varepsilon \sim unif(0.1, 5) \text{ and} \quad (16)$$

$$\rho \sim unif(-1, 1). \quad (17)$$

[1] Specifying gamma distributions using mode and standard deviations is simply easier to visualize (see http://doingbayesiandataanalysis.blogspot.com/2012/08/gamma-likelihood-parameterized-by-mode.html)

Method

Data were simulated for the following conditions for an interrupted time series model. Phase Length (l): 5, 8, 10, 15; standard deviation (σ_ε): 1, 2, 5; Intercept effect size (ES_1): 0.5, 1, 2, 5; Slope effect size (ES_2): 0, 0.3, 0.5, 1; and autocorrelations (ρ): 0, 0.2, 0.5, 0.8. Therefore, this was a fully crossed $3 \times 4 \times 4 \times 4 \times 4$ resulting in 768 conditions. One hundred datasets were generated for each condition yielding 76800 datasets. Some of the data conditions such as phase length, standard deviation, intercept effect size, and autocorrelations were chosen based on previous literature (Natesan & Hedges, 2017; Natesan Batley, Minka, & Hedges, 2020). The four models discussed in the models section were each fitted to each dataset. Root mean squared errors (RMSEs), mean relative biases, and coverage rates of the intercepts, slopes, intercept and slope effect sizes, autocorrelations, and standard deviations were used to compare the performance of the models. RMSEs are defined as the square root of the average squared deviation of the estimated value from the true value over all replications for a given data condition. Mean relative bias is computed as the average of the ratio of the difference between the true and the estimated value of a parameter. RMSE and relative bias for a parameter x whose estimate in the ith replication is x_i and true value is X over n replications is given in equations 18 and 19.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - X)^2}{n}} \text{ and} \tag{18}$$

$$Relative\ bias = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - X)}{X}. \tag{19}$$

Larger RMSEs and larger mean relative biases indicate less accurate estimates. According to Hoogland and Boomsma (1998), any relative bias greater than 0.05 was substantial in covariance structure models. We note here that relative bias was not computed for conditions where the true value was 0. Coverage rates are defined as the percentage of interval estimates that contain the true parameter value. 0-Coverage rates, that is, the percentage of credible intervals (CI) that contained 0 were used to examine if there was indeterminacy between estimating the slope and the autocorrelation. That is, when the credible interval of the autocorrelation contains zero when it is not expected to, the trend in the data may be incorrectly or inaccurately attributed to slope and the vice versa. This is represented as 0-coverage and represented as the parameter estimate

followed by "-0" (e.g. $\rho - 0$). However, if both slope and autocorrelation credible intervals contain zero, this signals that in some iterations autocorrelation takes more credit for the trend in the data (while the slope is estimated to be zero) and in some iterations slope takes more credit for trend in the data (while the autocorrelation is estimated to be zero).

Adequacy of Iterations and Replications

R was used for simulation and data analysis (R, 2014). The package RunJAGS (Denwood, 2013) conveniently calls JAGS (Plummer, 2003), runs parallel chains and iterates the model estimates until convergence is indicated. Four parallel chains were run with starting values independently generated for each chain from the prior distribution. The first 100,000 iterations were discarded using the burn-in procedure. Convergence was checked using two convergence diagnostics: the multivariate potential scale reduction factor (MPSRF, Brooks & Gelman, 1998), and Heidelberger and Welch's (1983) convergence diagnostic. In order to determine the adequacy of 100 replications (datasets) per condition, RMSEs and coverage rates of intercepts, slopes, intercept and slope effect sizes, autocorrelations, and standard deviations of the most complex model (SA) were plotted against the number of datasets generated. This procedure is similar to the one proposed by Koehler, Brown, and Haneuse (2009). When the RMSEs and coverage rates stopped fluctuating wildly or appeared to converge, there was indication of sufficient number of replications. This indicated that running the simulation for more datasets would not contribute to better diagnostic estimates such as RMSEs. In our study, 100 replications per data condition were deemed sufficient. The cumulative RMSEs and the coverage rates appeared to stop fluctuating significantly after the first 40 replications for all parameters. The cumulative RMSEs of all parameters fluctuated less than 0.03 after the first 60 iterations as shown in figure 1. The pattern for coverage rates was similar. We also stopped at 100 replications because of the computationally intensive nature of the estimation. It took 45 days to estimate all parameters of the four models across all 76800 datasets on six computers, each with quad core processors. We used doParallel and foreach (Weston & Calway, 2017) to parallelize the replications across the processors. Independent ANOVAs were conducted to measure the effect of the various data conditions on the RMSEs and coverage rates of the parameters. The data conditions were the independent variables. Eta-squared was computed for all main and 2-way interaction effects. Plots were examined to understand the patterns in parameter recovery.

INSERT FIGURE 1 ABOUT HERE

Results

Overall trends from the ANOVAs

The eta-squared effect sizes from independent ANOVAs are given in tables 2 and 3. These values give us a general pattern of those data conditions that affected the RMSEs, mean relative biases, coverage rates, means, and mean posterior standard deviations.

Autocorrelations: Longer phase lengths yielded smaller RMSE autocorrelations indicating that longer time-series yield more accurate estimates. However, phase lengths did not affect the coverage rates of autocorrelations with the exception of high 0-coverage rates for longer phase lengths combined with high autocorrelations. The ρ RMSEs were always larger for the NS model compared to the SA model especially for larger values of intercept effect size. The interaction effect between autocorrelation and model accounted for 8.56% of the variation in ρ RMSE. This is shown in Figure 2. Estimates from datasets with longer phase lengths combined with larger autocorrelation values covered 0 less frequently than those with smaller phase lengths and smaller autocorrelation values as shown in Figure 3. NS model had lower 0-coverage rates but substantially higher coverage rates of autocorrelation than the SA model as shown in Figure 4. Even for original autocorrelation value of 0.8, 60% of the SA model's CIs contained 0. Coverage rates of autocorrelation increased with increase in true ρ value for both NS and SA models, but the increase in coverage rate was more rapid for SA model. The NS model had narrower CIs than the SA model as shown in Figure 5. The width of the CIs increased with increase in phase length and decrease in autocorrelation.

INSERT TABLE 2-3 AND FIGURES 2-5 ABOUT HERE

Slopes: SI and SA models were compared for their recovery of slopes and slope effect sizes. RMSE of the first phase slopes (β_{12}) decreased with increase in phase length and decrease in standard deviation and autocorrelation as shown in figure 6. This makes intuitive sense because longer time-series, smaller standard deviations, and lower autocorrelations all contribute to clearer patterns, and hence, smaller slope RMSEs. The SA model had slightly lower β_{12} RMSE than the SI model but this effect was very small. The RMSE of the second phase slope β_{22} was

impacted most by variation in standard deviation and phase length. β_{22} RMSE decreased with increase in phase length and decrease in standard deviation.

INSERT FIGURE 6 ABOUT HERE

Intercepts: Standard deviation, autocorrelation, and their interaction explained most of the variation in the RMSE of the intercept of the first phase β_{11}. The RMSEs of β_{11} increased with increase in standard deviation and autocorrelation which is to be expected because increase in both these data conditions leads to less clear data patterns. The RMSEs of the second phase intercept β_{21} increased with increase in standard deviation and intercept effect size. The interaction effects between intercept effect size and standard deviation and intercept effect size and model also had a substantial effect on the RMSE of β_{21}. For intercept effect sizes up to 2, β_{21} RMSE was similar for all models but rapidly increased for the IO model followed by NS, SI, and SA models as shown in figure 7. It might seem illogical that with increase in intercept effect size, β_{21} RMSE would increase because larger intercept effect size would indicate clearer pattern. To understand this result more, we computed the mean relative bias of β_{21}. Mean relative bias of β_{21}increased with increase in intercept effect size as shown in figure 8. However, the absolute value of mean relative biases were less than 0.05 only for small values of intercept effect size for only SI and NS models.

INSERT FIGURE 7 and 8 ABOUT HERE

Slope and Intercept Effect sizes: The mean bias of the slope effect size (ES_2), that is, the mean difference between the true slope effect size and the posterior mean of the estimated slope effect size increased with increase in both intercept effect size and standard deviation together except for a standard deviation value of 1. This is probably because data patterns become less clear with increase in standard deviation. Slope effect size 0-coverage was higher for the SA model than that of the SI model and its credible intervals contained 0 more than 90% of the time. Still both models had overcoverage of 0. As expected, the 0-coverage of slope effect sizes decreased with increase in autocorrelation as shown in figure 9. However, the absolute mean relative bias of slope effect size was greater than 0.5 for all conditions. This is extremely high. The slope effect size coverage rates were all above 95% for all conditions except the SI model only for an autocorrelation value of 0.8. This situation seems to be exacerbated slightly by the intercept

effect size. The mean relative bias of the intercept effect size (ES_1) was largest for the IO model and lowest for the SI model yet the absolute mean relative bias values were greater than the acceptable value of 0.05 for all conditions. For IO and NS, ES_1 increased with increase in true intercept effect size, but SA and SI models exhibited an opposite pattern as shown in figure 9.

INSERT FIGURE 9 ABOUT HERE

Standard Deviations: The RMSE of the standard deviation (sigma) only differed to the second decimal for all cases. For the case of intercept effect size of 5, the SA model had high sigma RMSE. The mean standard deviation of β_{11} was affected by standard deviation and the model. As expected, this mean standard deviation increased with increase in standard deviation and model complexity. Models that estimated autocorrelations had larger mean standard deviations than those without. The same pattern was found for the mean standard deviation of β_{21}.

0-Credible Intervals for Autocorrelation and Slope

In order to better understand the behavior of the SA model with respect to 0-coverage, we investigated how many credible intervals of autocorrelation and slopes both contained the value of zero. If both autocorrelation and slopes of the second phase contained zero in their credible interval when they should not, this indicates a possible indeterminacy problem. That is, some of the patterns in the data are sometimes interpreted as only slopes with no credit given to autocorrelation and sometimes as only autocorrelation with no credit given to slope. The issue with this indeterminacy is that such an estimation would lead to increased Type II errors. That is, concluding that there is no autocorrelation or slope when there truly is.

We investigated 0-coverage in datasets where neither the true autocorrelation nor the true β_{22} values were zero. These were 43,200 in total. Figure 10 presents the histograms for the number of datasets whose autocorrelation CIs and β_{22} CIs that contain zero and the histograms for the number of datasets where both, either, or neither CIs contain zero. The histogram shows that 66.24% of the datasets' estimates contained zero in CIs of both parameters, 10.79% of the datasets' estimates contained 0 in only the autocorrelation CI, 20.9% of the datasets' estimates contained 0 in only the β_{22} CI. Only 2.02% of the datasets' estimates did not contain 0 in both autocorrelation and β_{22} CIs. Similarly, 66.9% of the datasets' second phase slope CIs contained zero 90-100% of the time. In over half of the data conditions, more than 80% of the datasets'

estimates showed that both CIs contained zero when they should not as shown in the histograms. This was the most prevalent case. That is, the probability of Type II error for both autocorrelation and slope of the second phase was over 0.8 in more than half the data conditions.

INSERT FIGURE 10 ABOUT HERE

Phase length (52.46%), intercept effect size (11.64%), autocorrelation (19.11%), and the interaction between phase length and autocorrelation (7.27%) explained variation in credible intervals of both second phase slope and autocorrelation containing zero. Data with longer phase lengths and larger true autocorrelation values had fewer cases where both credible interval estimates contained zero. Next, we considered cases where the second phase slope credible intervals contained zero, but the autocorrelation credible intervals did not. Data with longer phase lengths and higher true autocorrelation values had more cases where the autocorrelations CIs were estimated to contain zero. Only phase length (31.72%) explained cases where the second phase slope CIs contained zero, but the autocorrelation CIs did not. It was more common to see autocorrelation CIs contain 0 and second phase slope not contain zero for longer time-series and higher autocorrelations. In cases where autocorrelation CIs contained zero, but second slope phase CIs did not, intercept effect size (56.86%) explained most of the variance followed by the interaction of intercept effect size and autocorrelation (7.24%), and phase length and autocorrelation (6.65%). More autocorrelation CIs contained zero when the intercept effect sizes were large.

Finally, we considered only cases where the CI estimates of both second phase and autocorrelation did not contain zero when they should not. Phase length (17.17%), intercept effect size (19.6%), autocorrelation (7.27%), and the interaction effects between length and autocorrelation (6.34%), length and intercept effect size (19.77%), and autocorrelation and effect size (5.33%) explained most of the variation in these estimates. Cases with longer time-series, large intercept effect sizes, and large autocorrelations had more correct CIs, that is, those where neither autocorrelation CI nor the second phase slope CI contained 0. This shows that in general, clearer data patterns, that is, longer time-series with larger autocorrelations and intercept effect sizes yield more power to identify slope effect size and autocorrelation.

Conclusion

The question of which model needs to be fitted to data, in general, and SCED data, in particular, has long been a problem of interest for researchers. Often then the question is whether we need to mimic the true model, that is, the model from which the data are generated or whether we need to find the simplest model that best explains our data. Statisticians have tended to favor the Occam's razor approach by leaning towards selecting the simplest model, which is evident in many model fit indices such as the AIC and the BIC, which penalize models for complexity. We revisit the question posed in the title of this study as to whether we should favor accurate models or accurate estimates. This study tends towards the latter because by selecting the "correct" model, that is, the model that was used to generate the data, we obtain not only incorrect estimates but also reach incorrect decisions and potentially make Type II errors. Additionally, when it comes to whether the practitioner would be concerned more with obtaining the accurate model or arriving at proper inferences and conclusions, we would always favor the latter. Thus, our recommendation is to lean toward simpler models that we can expect to yield better estimates.

The mean relative bias of intercepts and intercept effect sizes show that the intercepts only (IO) model may not be the best-suited model to estimate the parameters of a two-phase SCED model with slopes and autocorrelations. This is perhaps because there is a pattern in the datasets that is unaccounted for when using the IO model. In fact, none of the models had desirable mean relative bias for intercept effect size and slope effect size. Although the slopes and autocorrelations (SA) model had lower RMSE for autocorrelation than the no slopes but autocorrelations (NS) model, it also had substantially higher 0-coverage rates and lower coverage rates for autocorrelations with wider credible intervals and high probability of Type II error rates. This indicates that the precision of the autocorrelation estimates obtained from the SA model is smaller than that of the NS model. The NS model had slightly higher second phase intercept and intercept effect size mean relative biases than the SA model.

The slopes and intercepts but no autocorrelations (SI) model had fewer 0-coverage rates for slope effect size than the SA model although the RMSE of the second intercept for the SA model was slightly better than that of the SI model. It also had the lowest intercept effect size mean relative bias, lower 0-coverage of slope effect size and lowest second phase intercept mean relative bias of all models. The main disadvantage of the SI model include that it does not

estimate autocorrelations. However, practitioners are not interested in estimating autocorrelations other than to get rid of their effects while computing effect sizes for interventions. Rather, practitioners are most interested in computing and interpreting the accuracy of intercept and slope effect sizes and their credible intervals. This shows that researchers are better off choosing the slopes and intercepts model without estimating autocorrelations, rather than use a model that includes intercepts, autocorrelations, and slopes. This of course, comes with the caveat that none of the models had desirable 0-coverage rates of slope effect sizes. The best model of the four, the SI model still had 0-coverage rates ranging from 59% to 95%. This overcoverage of 0 value however was accompanied by adequate coverage of the true value. The 0-coverage has implications for false decisions about the slope effect size even when the effect size is large, however these same credible intervals also contained the true value of the slope effect size. This implies that the credible intervals were much wider than desired and is an avenue for further research. Perhaps more informative priors could lead to narrower credible intervals. Still these findings only further make the case for future simulation studies to include 0-coverage rates because this diagnostic is very rarely reported in simulation studies. Therefore, we do not know how many studies that have adequate coverage rates of true values still might have undesirable 0-coverage rates.

Finally, the question is whether it is better to fit a simpler model such as the SI model even though it is not the "true" model. The advantages of fitting a simpler model to yield estimates that are more powerful outweigh the need to fit the more accurate model (SA) as our results show. Although Harrington and Velicer (2015, p. 176) noted that in single case designs, any analysis that ignores autocorrelations is "indefensible," Allison and Gorman (1993) suggested that failure to address and properly model trend can result in biased parameter estimates and inflated standard errors. On the other hand, Shadish, Rindskopf and Hedges (2008) reported that modelling the trajectory of the data might reduce the inflation of autocorrelation based on model misspecification. Our results shed additional light on these viewpoints mainly because we consider credible intervals, coverage rates, and 0-coverage of CIs. We have shown that researchers may want to choose only one of these sources of trend, that is, slope in favour of autocorrelation, in order to reduce 0-coverage and reduce model complexity.

Our results also emphasize that in simulation studies it is not adequate to observe only RMSE, standard errors, and biases as is common practice. Interval estimates and their coverage and 0-coverage rates have a more complete and sometimes even a different story to tell when evaluating the accuracy of parameters (Jennings, 1986, 1987; Natesan, 2015). Coverage rates have nominal values against which the performance of a model can be checked unlike RMSE and biases which are unbounded statistics. We have also showed that in addition to examining coverage rates, examining 0-coverage rates is important because excessively incorrect 0-coverage rates lead to incorrectly failing to reject the null hypothesis about the parameter. Adequate coverage rates along with excessively incorrect 0-coverage rates indicate wider than necessary interval estimates.

RMSEs can only be used to compare one criterion against another to conclude which criterion had lower RMSE. Whether this low RMSE is desirable or substantially above desirable is unknown unless the value is zero. We have shown that investigating the performance of credible intervals of two variables in tandem can be helpful in evaluating model performance. In fact, in the present study, the best model in terms of RMSE (SA) is not the best in terms of coverage.

We used weakly informative priors for the study. This allows us to stay agnostic about the parameters and try to estimate them. Of course, with small sample sizes such as the ones encountered in SCEDs, researchers may find it helpful to use informative priors based on previous research. Using informative priors may yield better estimates. Again, the use of informative priors in Bayesian estimation of SCEDs needs more investigation.

The implications of our study are multi-fold: First, our study informs authors of standards such as the WWC standards that estimating slopes and autocorrelations for SCED data often yields inaccurate estimates and is not recommended. Researchers may incorrectly infer that their intervention did not have a statistically significant intervention effect as shown by the confidence intervals of the trend of the data. Second, there has been much effort spent on developing effect sizes for slopes for SCEDs. The present study indicates that any effect size that is a function of the difference between the slopes severely underestimates slope effect size by often containing 0 in its credible interval unless it ignores autocorrelations. Therefore, future studies that develop slope effect sizes for SCEDs should take 0-coverage as an important diagnostic for testing the

performance of these effect sizes. Given that there is a need for statistics to be used in the domain of SCED analysis, the present study is of interest because it informs standards that should be developed that are standardized for SCED researchers to use.

Open practices statement: Software codes used to generate the data and evaluate the models are available, and preregistration is not applicable.

References

Algina, J. & Swaminathan, H. A. (1977). A procedure for the analysis of time-series designs. *Journal of Experimental Education, 45,* 56-60. https://doi.org/10.1080/00220973.1977.11011588

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research & Therapy, 31*, 621-631. https://doi.org/10.1016/0005-7967(93)90115-B

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/.

Baek, E. K., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across-participant variation in autocorrelation and residual variance. *Behavior Research Methods*, *45*, 65–74. https://doi.org/10.3758/s13428-012-0231-z

Baek, E., Moeyaert, M., Petit-Bois, M., Beretvas, S. N., Van den Noortgate, W., & Ferron, J. M. (2014). The use of multilevel analysis for integrating single-case experimental design results within a study and across studies. *Neuropsychological Rehabilitation, 24*, 590–606. https://doi.org/10.1080/09602011.2013.835740

Beretvas, S. N. & Chung, H. (2008). An evaluation of modified R2-change effect sizes for single-subject experimental designs. *Evidence-based Communication Assessment and Intervention, 2*, 120-128. https://doi.org/10.1080/17489530802446328

Box, G. E. P., & Pierce, W. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association, 65*, 1509–1526. DOI: 10.1080/01621459.1970.10481180

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. https://doi.org/10.1177/0145445503261167

Campbell, D. T. (1967). From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), *Problems in Measuring Change.* Madison: University of Wisconsin Press.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19,* 387-400. https://doi.org/10.1177/002246698501900404

Cobb, P., & Shadish, W. (2015). Assessing trend in single-case designs using generalized additive models. *Multivariate Behavioral Research*, *50*, 131-131. DOI:10.1080/00273171.2014.988991

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. doi:10.1177/0741932514557271

Crosbie, J. (1995). Interrupted time-series analysis with short series: Why it is problematic; how it can be improved. In J. M. Gottman (Ed.), *The analysis of change* (pp. 361–395). Mahwah, NJ: Erlbaum.

Denwood, M. J. (2013). 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

Efron, B. & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association, 70,* 311–319. https://doi.org/10.2307/2285814

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41*, 372-384. doi:10.3758/BRM.41.2.372

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

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

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis, 3*, 515-533. doi:10.1214/06-BA117A

Gelman, A. & Jakulin, A. (2007). Weakly informative priors. Retrieved from http://www.stat.columbia.edu/~gelman/presentations/weakpriorstalk.pdf

Glass, G. V., Willson, V. L., & Gottman, J. M. (1972). *Design and Analysis of Time-series Experiments*. Boulder, CO: Laboratory of Clinical Research.

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time series experiments.* Boulder: Colorado Associate University Press.

Gorman, B. S., & Allison, D. B. (1997). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159-214). Mahwah, NJ: Erlbaum.

Gottman, J.M. (1980). Analyzing for sequential connection and assessing inter-observer reliability for the sequential analysis of observational data. *Behavioral Assessment, 2*, 361-368. https://doi.org/10.1007/978-1-4612-3516-3_5

Gottman, J. M. & Glass, G. (1978). *Time-series analysis of interrupted time-series experiments*. In T. Kratochwill (Ed.), Strategies to Evaluate Change in Single Subject Research, NY:Academic Press

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

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

Hibbs, D. A. (1974). *Problems of statistical estimation and causal inference in time-series regression models*. In H. L. Costner (Ed.), Sociological Methodology (pp. 252–308). San Francisco: Jossey-Bass.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. An overview and a meta-analysis. *Sociological Methods and Research, 26*, 329–367. https://doi-org.ezproxy.brunel.ac.uk/10.1177/0049124198026003003

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

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

Huitema, B. E. & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60,* 38-58. https://doi.org/10.1177/00131640021970358

Huitema, B. E., McKean, J. W., & McKnight, S. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement, 59*, 767–786. https://doi.org/10.1177/00131649921970134

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

Jennings D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association, 81*, 471–476. DOI: 10.1080/01621459.1986.10478292

Jennings D. E. (1987). How do we judge confidence-interval adequacy? *American Statistician, 41*, 335–337. DOI: 10.1080/00031305.1987.10475509

Jones, R. R., Weinrott, M., & Vaughn, R. S. (1978). Time-series analysis in operant research. *Journal of Applied Behavior Analysis, 10,* 151-166. doi: 10.1901/jaba.1977.10-151

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

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. doi: 10.1198/tast.2009.0030

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., & 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. https://doi.org/10.1177/0741932512452794

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. https://doi.org/10.1177/10983007060080020701

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*, 297–303. https://doi.org/10.1093/biomet/65.2.297

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods*, *41*, 1262–1271. https://doi.org/10.3758/BRM.41.4.1262

Matyas, T. A. & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied*

Behavior Analysis, 23, 341–351. doi: 10.1901/jaba.1990.23-341

McCain, L. J. & McCleary, R. (1979). *Analysis of interrupted time series data*. In T. D. Cook and D. T. Campbell (eds.) The Design and Analysis of Quasi-Experiments. Chicago, IL: Rand-McNally.

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double boot strap method to analyze linear models with autoregressive error terms. *Psychological Methods, 5,* 87-101. https://doi.org/10.1037/1082-989X.5.1.87

Moeyaert, M., Ugille, M., Ferron, J., Onghena, P., Heyvaert, M., Beretvas, S. N., & Van den Noortgate, W. (2015). Estimating intervention effects across different types of single-subject. *School Psychology Quarterly, 30*, 50–63. https://doi.org/10.1037/spq0000068

Mood, J. L. (1950). *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

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

Natesan, P. (2015). Comparing interval estimates for small sample ordinal CFA models. *Frontiers in Psychology, 6*, 1599. https://doi.org/10.3389/fpsyg.2015.01599

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

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

Natesan Batley, P., Contractor, A., & Caldas, S. (2020). Bayesian time-series models in single case experimental designs: A tutorial for trauma researchers. *Journal of Traumatic Stress*. https://doi.org/10.1002/jts.22614

Natesan Batley, P., Minka, T., & Hedges, L. V. (2020). Investigating immediacy in multiple phase-change single case experimental designs using a Bayesian unknown change-points model. *Behavioral Research Methods.* doi: 10.3758/s13428-020-01345-z

Natesan Batley, P., Shukla Mehta, S. & Hitchcock, J. (2020). A Bayesian Rate Ratio Effect Size to Quantify Intervention Effects for Count Data in Single Case Experimental Research. *Behavioral Disorders.* https://doi.org/10.1177/0198742920930704

Pankratz, A. (1983). *Forecasting with univariate Box–Jenkins models: Concepts and cases.* New York: Wiley.

Parker, R. I., Hagan-Burke, S., & Vannest, K. J. (2007). Percent of all nonoverlapping data PAND: An alternative to PND. *Journal of Special Education, 40*, 194-204. https://doi.org/10.1177/00224669070400040101

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

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining non-overlap and trend for single-case research: Tau-U, *Behavior Therapy, 2*, 284-299. DOI: 10.1016/j.beth.2010.08.006

Petit-Bois, M., Bark, E. K., Van den Noortgate, W., Beretvas, S. N., & Ferron, J. M. (2016). The consequences of modeling autocorrelation when synthesizing single-case studies using a three-level model. *Behavior Research Methods, 48*, 803-812. DOI: 10.3758/s13428-015-0612-1

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/.

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

Shadish, W. R. (2014). Statistical analysis of single-case designs: The shape of things to come. *Current Directions in Psychological Science, 23*, 139-146. https://doi.org/10.1177/0963721414524773

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 188-196. https://doi.org/10.1080/17489530802581603

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

Shadish, W.R., & Sullivan, K.J. (2013). Characteristics of Single-Case Designs Used to Assess Treatment effects in 2008. *Behavior Research Methods, 43*, 971-980. https://doi.org/10.3758/s13428-011-0111-y

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. DOI:10.1016/J.RASD.2014.08.011

Simonton, D. K. (1977). Cross sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin, 84,* 489-502. https://doi.org/10.1037/0033-2909.84.3.489

Simonton, D. K. (1979). Reply to Algina and Swaminathan. *Psychological Bulletin, 86,* 927-928. https://doi.org/10.1037/0033-2909.86.5.927

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in N = 1 designs. *Behavior Modification*, *34*, 195–218. https://doi.org/10.1177/0145445510363306

Solomon, B. G. (2013). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification, 38,* 477–496. https://doi.org/10.1177/0145445513510931

Sullivan, K. J., Shadish, W. R., & Steiner, P. M. (2015). An introduction to modeling longitudinal data with generalized additive models: Applications to single-case designs. *Psychological methods*, *20*, 26-42. https://doi.org/10.1037/met0000020

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. DOI: 10.1177/1468017310381812

Tryon, W. W. (1982). A simplified time-series analysis for evaluating treatment interventions. *Journal of Applied Behavior Analysis, 15*, 423–429. doi: 10.1901/jaba.1982.15-423

Ugille, M., Moeyaert, M., Beretvas, T., Ferron, J., & Van den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods, 44*, 1244-1254. https://doi.org/10.3758/s13428-012-0213-1

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

van den Noortgate, W., & Onghena, P. (2003b). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, *18*, 325–346. https://doi.org/10.1521/scpq.18.3.325.22577

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: daily experiences. *Journal of Personality*, *59*, 609–662. https://doi.org/10.1111/j.1467-6494.1991.tb00261.x

Weston, S. & Calway, R. (2017). Getting started with doParallel and foreach. Retrieved from https://cran.r-project.org/web/packages/doParallel/vignettes/gettingstartedParallel.pdf on March 5, 2018.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta analysis in individual-subject research. *Behavioral Assessment, 11,* 281-296.