Abstract

Background

When developmental disabilities researchers use multiple-baseline designs they are encouraged to delay the start of an intervention until the baseline stabilizes or until preceding cases have responded to intervention. Using ongoing visual analyses to guide the timing of the start of the intervention can help to resolve potential ambiguities in the graphical display; however, these forms of response-guided experimentation have been criticized as a potential source of bias in treatment effect estimation and inference.

Aims and Methods

Monte Carlo simulations were used to examine the bias and precision of average treatment effect estimates obtained from multilevel models of four-case multiple-baseline studies with series lengths that varied from 19 to 49 observations per case. We varied the size of the average treatment effect, the factors used to guide intervention decisions (baseline stability, response to intervention, both, or neither), and whether the ongoing analysis was masked or not.

Results

None of the methods of responding to the data led to appreciable bias in the treatment effect estimates. Furthermore, as timing-of-intervention decisions became responsive to more factors, baselines became longer and treatment effect estimates became more precise.

Conclusions

Although the study was conducted under limited conditions, the response-guided practices did not lead to substantial bias. By extending baseline phases they reduced estimation error and thus improved the treatment effect estimates obtained from multilevel models.

What this paper adds

This Monte Carlo study contributes to the single-case design literature by addressing the concern with response-guided experimentation. The study examined the bias and precision of treatment effect estimates obtained from multilevel models under conditions with response-guided experimentation. It was found that under the simulated conditions, response-guided experimentation did not result in substantial bias of the treatment effect estimates using multilevel models.

Keywords: Single-case research, Response-guided, Multilevel modeling, Monte Carlo study

The Impact of Response-Guided Baseline Phase Extensions on Treatment Effect Estimates

1. Introduction

Single-case researchers frequently adopt a form of response-guided experimentation where decisions about the design of the study are made based on an ongoing visual analysis (e.g. Gast, 2009; Kazdin, 2010). For example, multiple-baseline researchers may delay the start of intervention until the data document a stable baseline pattern so that baseline trends can be reliably extended, or the researchers may wait for a case in a multiple-baseline design to respond to intervention prior to intervening with the next case. Using ongoing visual analyses to guide the timing of interventions can help to resolve what would be ambiguities in the graphical display and thus increase the analyst's sensitivity to detecting effects in the graphical display.

Although well intentioned, these response-guided experimental strategies may bias the intervention effect estimates or inferences. These response-guided strategies have been compared to the strategy determining sample size in a group comparison study by repeatedly testing for differences as the sample is gathered, an approach that is known to increase Type I error rates (Allison, Franklin, & Heshka, 1992). A simulation study showed an increase in the number of false detections of effects by randomization tests when the single-case data were gathered in a response-guided manner (Ferron, Foster-Johnson, & Kromrey, 2003). Furthermore, a study of visual analysis of graphs of randomly generated observations showed that when the line separating the hypothetical baseline and treatment phases was placed after a stable set of observations, as opposed to placed randomly, that visual analysts were more likely to incorrectly conclude that there were effects (Todman & Dugard, 1999). These concerns led to the development of a method of masking graphs in an ongoing visual analysis (Ferron & Jones, 2006).

1.1 Masked Visual Analysis

Masking graphs in a visual analysis, or masked visual analysis (MVA), was first proposed in single-case research by Mawhinney and Austin (1999). In MVA, the transition points between different phases (i.e., from baseline to treatment) are purposely concealed from the visual analysts, and the visual analysts are tasked with determining when the treatment was initiated. If a single-case research design involves multiple participants such as multiple-baseline design, the visual analysts are tasked with deciding the intervention initiation points for each case. Later, response-guided or ongoing MVA was developed with more detailed procedures to ensure control over Type I error rates (Ferron & Jones, 2006; Ferron & Levin, 2014).

For the response-guided MVA method (Ferron & Jones, 2006), a single-case research team is divided into two separate groups; an intervention team and an analysis team. The intervention team is responsible for interactions with the cases and data collection, whereas the analysis team is responsible for using visual analysis of a masked graph to make decisions about baseline stability and response to intervention. The graphs are ‘masked’ because information about the independent variable (i.e., which case will enter intervention first and whether the observation is part of a baseline or treatment phase) is not marked on the graph or made explicit to the analysis team. The analysis team analyzes the stability of the masked data one session at a time and the data collection continues until all cases show a stable pattern for the baseline observations. Once stability is obtained, the analysis team directs the intervention team to randomly select a case to begin the intervention phase. The intervention team does so and the information about which case is in the intervention phase is not given to the analysis team. The collected data are still masked and sent to the analysis team, and the analysis team continues analyzing the data until there is sufficient data to demonstrate that one case has initiated the

intervention. Then the second case is randomly selected to begin the intervention. This process continues until all cases participate in the intervention. By computing the probability of selecting the intervention order for all cases when there is no true effect this method theoretically controls the Type I error rate (Ferron & Jones, 2006). In addition, a recent Monte Carlo study has shown that the response-guided MVA controls Type I errors to the nominal level (Ferron, Joo & Levin, 2017).

If single-case researchers were interested in integrating several studies, including a meta-analysis of single-case studies (Shadish, 2014), they could do so by combining probabilities (Rosenthal, 1978; Solmi & Onghena, 2014). Although probability estimates of response-guided experimentation using MVA are accurate, MVA was developed for and is limited to the estimation of probabilities. Researchers who want to estimate and synthesize effect sizes, rather than probabilities, must turn to other analyses that may or may not be negatively impacted by response-guided experimentation. One widely used approach for obtaining effect size estimates of single-case data that is of particular interest in the current study involves the use of multilevel models (Van den Noortgate & Onghena, 2003a, 2003b).

1.2 Multilevel Models

Multilevel models have been utilized for analyzing single-case data because they take variability within- and between-cases into account when estimating the treatment effect (e.g., Van den Noortgate & Onghena, 2003a, 2003b). In principle, multilevel models are specifically developed for analyzing hierarchically structured data, where lower-level units are nested in higher-level units. Hierarchical structured data are often found in behavioral and social science studies. For example, in educational settings, students are nested in classes and classes are nested

in schools. Similarly, multiple-baseline data can be considered as hierarchical because the repeated observations are nested within cases.

Multilevel models for multiple-baseline studies are also advantageous over multiple single-level models. For example, multilevel models provide not only individual cases' treatment effect estimates, but also the average treatment effect estimate across cases. Although multilevel models were developed based on the assumption of a relatively large number of second-level units, a number of simulation studies have shown that multilevel models with small sample size adjustments produce unbiased treatment effect estimates and reliable statistical inferences with as few as four cases (e.g., Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Ferron, Farmer, & Owens, 2010).

In general, two multilevel models are widely considered in single-case research: 1) a change in level model with two fixed effects, one for the across case average baseline level and one for the across case average treatment effect (Van den Noortgate & Onghena, 2003a), and 2) a change in level and trend model with four fixed effects, corresponding to the across case average baseline intercept, baseline slope, treatment effect on level, and treatment effect on slope (Van den Noortgate & Onghena, 2003b). For more detailed description of the two multilevel models, including the mathematical equations and notations of the models, see the Appendix.

In previous single-case research studies, a concern with respect to response-guided experimentation has been specifically raised in the context of estimating treatment effects through multilevel models of multiple-baseline data (Ferron, Moeyaert, Van den Noortgate, & Beretvas, 2014). To illustrate, consider a study where the intervention was designed to increase a behavior that in the absence of the intervention would not trend up or down. Although the true baseline slope parameter is zero, it is unlikely that the slope of the observed values would be

exactly zero. If by chance the slope happened to be close to zero or negative after gathering the minimum number of baseline observations the researcher would intervene, but if the slope were notably positive the researcher would gather another observation, or two, or three, each time looking for a flat or downward trend. When the more acceptable baseline emerged, the researcher would intervene. If this study were to be replicated many times, the average baseline slope would be negative, not zero, and as a consequence the intervention effect estimate may be positively biased.

1.3 Purpose

Meta-analyses of single-case studies are likely to include studies that used response-guided experimentation because many researchers are taught to make intervention decisions based on an ongoing visual analysis (e.g., Ferron & Levin, 2014, Gast, 2009; Kazdin, 2010) and thus it is important to understand the degree to which response-guided experimentation may bias treatment effect estimates. If the biases are minimal then it may be reasonable to ignore the response-guided nature of the data collection in the meta-analysis of single-case studies. If the biases are substantial, however, then it is important to determine if there are ways of responding to the data that meet the needs of researchers but create fewer problems for effect estimation and meta-analyses; or if there are ways to adjust the estimation of effects to compensate for any bias. Thus, the purpose of this Monte Carlo simulation study was to evaluate which response-guided practices do and do not bias the across case average effect estimates in multiple-baseline studies using multilevel models.

2. Methods

2.1 Response-Guided Algorithm Implementation

To investigate the accuracy of the treatment effect estimate from studies that used response-guided experimentation, an algorithm was written to simulate a variety of response-guided practices, including extending phases while waiting for baseline stability, waiting for cases to respond to intervention, or both. In addition, the response-guided algorithm was developed for ongoing analyses that are masked as to the order in which cases were entering intervention, and for analyses that were not masked. Writing an algorithm to simulate ongoing visual analysis had the advantage of making it practical for us to simulate the number of datasets necessary for a Monte Carlo study of response-guided experimentation. However, the finer grain criteria used to make decisions in visual analysis are not fully defined or agreed upon, and as a consequence not all visual analysts make the same decisions (e.g., Brossart, Parker, Olson, & Mahadevan, 2006). Thus, the algorithm that makes the study tractable also limits us to the study of a single operationalization of ongoing visual analysis.

We used the literature on the visual analysis of single-case data to identify the criteria to operationalize and to guide us in determining a range of potential parameter values. For example, we chose parameter values that indicated changes of about two standard deviations of the baseline observations, either within a phase to evidence instability or across phases to evidence response to intervention, because changes of this magnitude tend to be reliably detected visually (Knapp, 1983), and are somewhat typical (e.g., Parker & Vanest, 2009, found that across 200 AB panels from published single-case studies the average R^2 was .42, which corresponds to an average shift of about 1.7 *SD*s).

We then conducted a preliminary study examining graphs of the data generated by the algorithm for different combinations of parameter values. We selected values that: 1) led to decisions that mimicked the decisions we would make based on our ongoing visual analysis of

the graphs, and 2) balanced the aim of high detection rates of true effects through MVA with the aim of gathering the minimum number of baseline observations that is sufficient for detecting effects. The development of the algorithm is reported more fully elsewhere (Ferron et al., 2017), as are the estimates of power and Type I error for MVA. We now give more detail on each of the response-guided practices included in the algorithm.

2.1.1 Baseline Stability

The algorithm is based on the assumption that the treatment will raise the level of behavior. It extends phases if baseline phase observations are unstable with evidence of any of the following four criteria: 1) an increasing trend of observations throughout the baseline phase, 2) an increasing trend of observations at the end of the baseline phase, 3) an improved outlying observation at the end of the baseline phase, or 4) an increasing level shift of observations during the baseline phase.

The response-guided algorithm considers baseline observations as unstable if 1) the ordinary least squares (OLS) regression slope of baseline observations is greater than 0.5 times the standard deviation (SD) of baseline observations, or 2) the OLS slope of the final 3 baseline observations is greater than 0.5*SD of baseline observations, or 3) a final baseline observation is greater than the mean of baseline observations plus 2*SD of baseline observations, or 4) the difference between the mean of the last half of the baseline observations and the mean of the first half of the baseline observations is greater than 1.5*SD of baseline observations.

2.1.2 Response to Intervention

The ongoing analysis algorithm also extends phases as needed so that the intervention start point for each successive case occurred only after the preceding case exhibited a treatment effect. The response-guided algorithm considers the treatment effect is not evident if 1) the mean difference

between the treatment and baseline observations is less than 2 times the standard deviation (SD_p) pooled across the two phases and 2) the mean difference between the final three treatment observations and the baseline observations is less than 2*SD_p.*2.1.3 Minimum and Maximum Series Length*

In the response-guided algorithm, we set a minimum series length to secure that a multiple-baseline study would be considered "acceptable" in accordance with current single-case intervention design standards (Kratochwill, et al., 2013). For example, both baseline and intervention phases should include at least five observations, respectively, for each case to determine the treatment effect. In addition, each successive case should start the intervention phase at least three observations after the preceding case starts the intervention phase. More specifically, suppose there is a multiple-baseline study where four cases are involved and each case enters the intervention phase one at a time. If none of the phases were extended during the ongoing analysis, then the series length for each case would be 5 + 3 + 3 + 3 + 5 = 19. The baseline phase length for each case would be 5, 8, 11, and 14, respectively.

We also included a maximum series length that the response-guided algorithm is allowed to extend the phases. By setting maxima on the number of extensions, the algorithm would not be stuck on extending the phases unnecessarily for the circumstances where there was no effect. The algorithm is allowed to extend the baseline phase up to three additional observations and the stagger between intervention start points for successive cases are allowed to be extended for up to five additional observations. The final intervention phase is allowed to be extended up to three additional observations. Thus, if four cases were involved in a multiple-baseline study and all phases were extended, then the maximum series length for each case would be 8 + 8 + 8 + 8 + 8 = 40. Note that series lengths between 19 and 40 are often observed in single-case studies as

shown in a review of single-case studies conducted by Shadish and Sullivan (2011). They showed that the median series length was 20 and 90% of the series lengths were less than 50.

2.1.4 Masked and Unmasked Analyses

Lastly, two types of the response-guided algorithm were implemented separately for ongoing analyses, one where the order that cases started the intervention was masked and one where the order was not masked. For the masked algorithm, because the order of cases entering intervention was unknown, the phase extension algorithm was applied to all cases. For the unmasked algorithm, because the order of cases entering intervention was predetermined, the phase extension algorithm was applied to one case at a time. Note that the approach of masked response-guided experimentation mimics the method described in Ferron and Jones (2006).

2.2 Study Design

A Monte Carlo simulation study was conducted to explore which forms of response-guided experimentation possibly bias effect estimates of multilevel models of multiple-baseline studies. The Monte Carlo study was conducted to examine multiple-baseline studies with four cases (n = 4). We chose four as the number of participants because it is a commonly observed number in multiple-baseline studies as evidenced by surveys of the single-case literature (e.g., Farmer et al., 2010; Shadish & Sullivan, 2012). We examined six different response-guided experimentation conditions: 1) baseline stability based on unmasked analysis, 2) baseline stability based on masked analysis, 3) response to intervention based on unmasked analysis, 4) response to intervention based on masked analysis, 5) baseline stability and response to intervention based on unmasked analysis, and 6) baseline stability and response to intervention based on masked analysis. A fixed baseline length condition (i.e., no extensions and thus a total

of 19 observations per case) was additionally compared to each of the six response-guided conditions.

We defined the effect size as the mean difference in the outcome during baseline versus intervention phases, divided by the standard deviation of the baseline observations (i.e., standardized effect size), and examined conditions where the treatment effect size used for data generation was varied from 0 to 4 in increments of 1 to cover the range of effects typically encountered in single-case studies (Parker & Vannest, 2009). Conditions with different effect sizes were considered so that we could examine the probability of incorrectly concluding there was a treatment effect when the true effect size was zero (i.e., Type I error) and the probability of correctly identifying true effects when the effect size was nonzero (i.e., power).

In addition, we generated data based on two different types of level-2 error variances for the multilevel model: simple vs. complex. For the simple condition, only the baseline levels (i.e., intercepts) were generated to vary randomly across cases. For the complex condition, both the baseline levels and treatment effects were generated to vary randomly across cases. Note that for both simple and complex conditions, level-2 error covariances were fixed to zero based on the previous single-case multilevel studies (e.g., Moeyaert, Ugille, Ferron, Beretvas & Van den Noortgate, 2016). Crossing the levels of simulation factors, the total number of conditions was 7 x 5 x 2 = 70. The number of replications per condition was set to 3,000.

2.3 Data Generation

All data were generated using the IML Procedure in SAS (SAS Institute, 2014). The data generation coupled the form of experimentation (fixed phase lengths or a specific form of response-guided experimentation) with an underlying multilevel model that was based on the assumption of a shift in level between baseline and intervention phases (i.e., the model in

Equations 1, 2.1, and 2.2 in the Appendix). The generating value for average baseline level, γ_{00}, was zero and those for, the average treatment effect, γ_{11}, were varied depending on the effect sizes. The values for the level-1 errors were randomly generated from a standard normal distribution using the RANNOR function implemented in SAS/IML. The level-2 errors were generated from a normal distribution. When data were generated with the simple level-2 error variance, the intercept variance, σ_{r0}^2, was set to 0.5 and the treatment effect variance, σ_{r1}^2, was fixed at 0. Similarly, when data were generated with the complex level-2 error variance, both intercept and treatment effect variances were set to 0.5. The values for the dependent variable, Y_{ij}, were created using these parameter values and Equations 1 to 2.2 in the Appendix.

2.4 Fitted Models

For each simulated data set, two separate multilevel analyses were run. In the first model, the change in level model was fitted based on the assumption of no trends. In this model, the level-1 error variance was assumed to be independent and homogeneous across phases and cases and the level-2 error variances were modeled for both intercept and treatment effect and assumed to be independent (i.e., Equations 2.1 and 2.2 in the Appendix). In the second model, the change in level and trend model was fitted. Note that the second model does not assume that there are trends, but rather models possible trends. The same level-1 error variance assumptions were made and all fixed effects for the change in level and trend model were allowed to vary (i.e., Equations 4.1 – 4.4 in the Appendix). As used in traditional multilevel models, parameters of both models were estimated with restricted maximum likelihood (REML) estimation. In addition, the Kenward-Roger inference method (Kenward & Roger, 1997) was used to compute adjusted standard error and degrees of freedom estimates because of the small sample size. When a relatively small number of cases is available, including multiple-baseline studies, the Kenward-

Roger inference method outperforms the other inference methods yielding reliable statistical inference for fixed effects (Ferron et al., 2009). Finally, the SAS MIXED Procedure was used to estimate the parameters of the models.

2.5 Analysis

We examined the accuracy of the intervention effect estimates and statistical inference for each of the conditions, using each of the models. For the accuracy of parameter estimation of the change in level and the change in level and slope models, relative bias and root mean square error (RMSE) were computed using the following equations:

$$\text{Relative Bias} = \frac{\sum_{i=1}^{R} \frac{\widehat{\gamma_i} - \gamma}{\gamma}}{R} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{R} (\widehat{\gamma_i} - \gamma)^2}{R}} \quad (2)$$

where R is the number of replications, γ is the generating parameter and $\widehat{\gamma_i}$ is the parameter estimate for i^{th} replication. Relative bias was computed for each replication then averaged across replications, whereas RMSE was computed by taking the sum of squared differences between the population and estimated parameter across replications and then averaged across replication (mean square error). The final RMSE was obtained by taking the square root of the mean square error. Relative bias of magnitude less than .05 can be considered to be acceptable bias for parameter estimates (Hoogland & Boomsma, 1998).

As a measure of the performance of statistical inference, 95% confidence interval (CI) coverage, and statistical power/Type I error of the treatment effect were computed. For CI coverage rates, the CI was first calculated as noted using the Kenward-Roger method for obtaining the degrees of freedom. The CI coverage rate was then obtained by computing the proportion of replications in which the generating value for the treatment effect was inside the

computed CI. Similarly, for statistical power/Type I error, the two-tailed *p*-value of the treatment effect estimate was computed, and then the proportion of replications in which the *p*-value was less than the nominal α-level (.05) was computed to obtain empirical power/Type I error rates. When the true treatment effect was zero, the proportion of *p*-values less than .05 refers to the Type I error rate, and when the true intervention effect was non-zero, then the proportion estimates the statistical power. In addition, we also tracked the average series lengths of multiple-baseline studies across replications to investigate the effect of response-guided experimentation for each condition.

3. Results

The results are provided in Tables 1 to 4. Table 1 shows the results from the fixed series length as a basis of comparison for results in which response-guided experimentation was used. Note that series lengths for the fixed criteria condition were 19 as with the minimum series length for the response-guided algorithm. Tables 2 and 3 show the results from the conditions where phases were extended based on baseline stability and response to intervention, respectively. Table 4 represents the results from the condition where phases were extended based on both baseline stability and response to intervention.

3.1 Bias and RMSE of the Effect Estimate

Overall investigation of these tables shows that none of the methods of responding to the data led to substantial bias in the treatment effect estimates. Both change in level and change in level and trend models produced minimal bias in treatment effect estimates across conditions. The maximum relative bias we observed was .021 for the condition where the treatment effect was estimated with the change in level model and both baseline stability and response to intervention were applied in the response-guided experimentation and the ongoing analysis was

unmasked. Note that the amount of the maximum relative bias, .021, is less than 5% bias of the population values.

Furthermore, as intervention timing decisions became responsive to more factors, baselines became longer and treatment effect estimates became more precise. That is, the RMSE was smaller when the algorithm responded to both baseline stability and response to intervention rather than when responding to just one of these factors. The maximum average series length was 37.9 when both baseline stability and response to intervention were used, whereas the corresponding series lengths were 23.5 when only baseline stability was used, and 36.4 when only response to intervention was used. When phases were extended based on either baseline stability or both baseline stability and response to intervention, the series length became longer for masked analyses than unmasked analyses and, consequently the effect estimates were more precise (i.e., RMSE was smaller) for masked analyses than unmasked analyses. Interestingly, when phases were extended based on responding to intervention only, however, the phases became shorter for masked analyses than unmasked analyses. Thus, RMSE was larger for masked analyses than unmasked analyses.

Although responding to baseline stability was not affected, as the effect size increased, the timing of the stagger between intervention and baseline phases was more likely to be extended when effect sizes were smaller. As a result, an increasing pattern of RMSE was observed as the effect size increased. In addition, as expected, RMSE was smaller when the treatment effect was estimated with the change in level model than with the change in level and trend model because the model with trends is more complex and over-parameterized. It was also found that as the level-2 error became more complex, a consistently higher amount of error for the treatment effect estimate was observed.

3.2 Statistical Inferences of the Effect Estimate

As shown in Tables 1 to 4, the 95 % CI coverage rates ranged from .922 to .975 across conditions, indicating the estimated values were close to the nominal coverage rate (.95). When the data were generated with more complex level-2 error variance, CI coverage rates were consistently lower than those resulting from when data were generated with simple level-2 error variance. Similarly, CI coverage rates for the change in level and trend model were slightly lower than those for the change in level model. No substantial difference in CI coverage rates between masked and unmasked analyses were found.

Overall, the Type I error result indicated that the probability of incorrectly concluding an effect of treatment when there was no true effect was typically not too far from the nominal level (.05): the actual Type I error rates were lower than Bradley's liberal criterion for robustness of .075 (Bradley, 1978), with the exception of one estimate of .078,. Although minimal, Type I error rates were higher than the nominal level, when the data generation model included more complex level-2 error variance. The change in level and trend model, consistently showed higher Type I error rates than the change in level model.

Consistent with expectation, power increased as the effect size increased and even reached 100 % in many conditions of the study. In addition, similar to the RMSE result, power to detect the treatment effect was affected by the methods of response-guided experimentation. For example, power increased substantially when the series length of multiple-baseline studies became longer. The change in level and trend model, and more complex level-2 error variance condition showed consistently lower power than the change in level model, and simple level-2 error variance condition, respectively.

4. Discussion

When developmental disabilities researchers use multiple-baseline designs, they are encouraged to delay the start of interventions until baselines stabilize or until the preceding cases have responded to intervention (e.g., Gast, 2009; Kazdin, 2010). Although waiting for the baseline to stabilize or intervention to have an effect can help resolve what would be ambiguities in the graphical display, these forms of response-guided experimentation have been criticized as a potential source of bias in treatment effect estimation and inference (Ferron et al., 2014). In previous multiple-baseline design research, these concerns have remained unresolved in the context of estimating treatment effects using multilevel models. Thus, this study aimed to use a Monte Carlo simulation method to empirically examine the bias and error in treatment effect estimates obtained from multilevel models under conditions with response-guided experimentation. The simulations included four-case multiple-baseline studies, which varied in the size of the average treatment effect, the complexity of level-2 error variances, and the methods used to extend the phases.

The results of the study indicated that none of the methods of responding to the data led to substantial bias in the average treatment effect estimates for the conditions examined here. The relative bias of the treatment effect estimates across conditions were within 5% of the population value. In addition, statistical inferences including CI coverage rate, Type I error and statistical power for the treatment effect estimates were acceptable for many conditions. Type I error rates were slightly inflated for the complex error conditions, but never exceeded .078. Note that Type I error of .078 is slightly above a Bradley's liberal criterion (.075; Bradley, 1978). These findings seem surprising because researchers have raised concern with regard to the impact of response-guided experimentation on treatment effect estimation (Ferron et al., 2014). Although a recent simulation study found strict control of Type I error for response-guided MVA

(Ferron et al., 2017), that study was limited to visual analysis and thus questions remained as to whether estimating the effect from the response-guided data using multilevel models would provide unbiased estimates. The findings of the current study provide empirical evidence that response-guided experimentation, as we have operationalized it, does not result in substantially biased treatment effect estimates when the data are analyzed with multilevel models. Furthermore, the current study findings suggest that practitioners who use response-guided experimentation may obtain multiple-baseline data for which the use of multilevel models is appropriate.

Although the response-guided experimentation we considered in this study did not cause substantial problems in estimation of and inferences about treatment effects using the multilevel models, it should be emphasized that the results are limited to multilevel models and only when the model was correctly specified. As noted in previous studies (e.g., Allison, Franklin, & Heshka, 1992; Ferron et al., 2003; 2017; Todman & Dugard, 1999), response-guided experimentation may inflate the probability of Type I error for visual analysis and randomization tests, and only control the Type I error rate in MVA if the graphs are masked. Thus the results of the current study should not be generalized to other analytical methods in single-case research and more research is needed to investigate the impact of response-guided experimentation on other analyses.

In addition, we recognize that the current study was conducted under limited simulation conditions and data generation models. For example, the simulation study only included a four-case multiple-baseline design. Although four is a typical number for multiple-baseline studies, the number of cases varies from study to study, and has include as many as 36 cases in applied settings (Ferron et al., 2014). Note that the number of cases in the multiple-baseline study often

affects the precision of the treatment effect estimates, especially for multilevel models. Previous studies have found larger number of cases tend to yield more precise parameter estimates for fixed effects in multilevel models (e.g., Moeyaert et al., 2016). Thus, it would be worthwhile to generalize the response-guided algorithm and the results of this study to include more cases than four.

Furthermore, the current study was also limited to the conditions where only the methods of experimentation, the size of effects, and the complexity of the level-2 errors were varied. The current study only examined multilevel models where the level-2 error structures were correctly specified and independently generated. In addition, the data were generated with a relatively simple change in level model. Effects are more complex in some single-case studies and previous studies have investigated more flexible and complex forms of multilevel models, including those with linear trend effects in the treatment phase (Moeyaert et al., 2016), nonlinear trends in the treatment phase (Hembry, Bunuan, Beretvas, Ferron, & Van den Noortgate, 2015), autocorrelation in level-1 errors (Baek & Ferron, 2013), and heterogeneous level-1 error variances (Joo, Ferron, Moeyaert, Beretvas, & Van den Noortgate, 2017), Future research should examine the performance of more complex multilevel models with response-guided experimentation.

For the response-guided algorithm we developed in the study, the criteria and parameter values were determined based on review of the visual analysis literature, and preliminary study, which included refinement based on feedback from experienced visual analysts about the appropriateness of the decisions being made. From the current study, we also found that the developed response-guided algorithm was sensible in the way that series length of the multiple-baseline studies varied in accordance with the methods of response-guided experimentation. For

example, when the phases were extended responding to intervention, the average series length over replications decreased as the treatment effect size increased, indicating fewer staggers were extended when an apparent treatment effect was present. Contrastingly, when the phases were extended waiting for the baseline to stabilize, the average series length was not affected by the size of the treatment effect. In addition, for the condition where both baseline stability and response to intervention were used in making extensions, series length became longer than those where only one of the criteria was used. This finding provides further evidence that the developed response-guided algorithm leads to data that align with the theoretical expectations.

The response-guided algorithms, however, are still not as nuanced as a visual analysis and are limited by the criteria and parameter values that were chosen. Criteria and parameter values chosen to maximize congruence of the algorithms decisions and those of a visual analyst, would likely vary across research contexts and visual analysts. In the future, alternative algorithms using different criteria and/or parameter values need to be examined to determine the degree to which the findings in this study generalize across different operationalization of ongoing visual analysis. For example, in some contexts the researchers may wish to focus directly on variability in making decisions about extending phases, or may want to increase the maximal number of extensions. Future research could determine whether these sorts of changes would impact the findings. In other contexts, researchers may choose to wait until stability is obtained and then randomly select a start point from two or three observations. Although the addition of this randomized element would not be expected to create bias, it would be interesting in future research to specifically examine such an experimental strategy. One could also examine whether choosing to start interventions after a deteriorating trend is observed, as opposed to after stability is observed, would lead to different results.

We hope the empirical evidence in this study will be useful to applied researchers and meta-analysts, considering the appropriateness of effect estimation in the context of response-guided experimentation. In a meta-analytic context, we recommend coding whether or not the study was response-guided and then using this coded variable as a potential moderator in the meta-analytic model. By doing so, we would get a more complete understanding of the impact of response-guided experimentation on treatment effect estimates, which in turn would further the debate on the merits and limitations of response guided experimentation. Finally, because the effects were not substantially biased by any of the methods of responding to the data, and not more or less biased by whether or not the graphs were masked, we recommend these decisions be based on other factors. For example, if researchers were planning to conduct a MVA in addition to estimating treatment effects through a multilevel model, it would be important to use masking for the MVA. The use of masking would have the additional benefit of increasing the precision of the average treatment effect estimate of the multilevel model, as long as baseline stability was being used to extend phases because under these conditions the masked analysis would lead to longer series lengths.

Acknowledgement

We gratefully acknowledge support from the Institute of Educational Sciences, U.S. Department of Education (through Grant R305D150007). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Allison, D. B., Franklin, R. D., & Heshka, S. (1992). Reflections on visual inspection, response guided experimentation, and Type I error rate in single-case designs. *Journal of Experimental Education*, *61*, 45-51.

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.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

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.

Farmer, J., Owens, C. M., Ferron, J. M., & Allsopp, D. (2010, May). A review of social science single-case meta-analyses. Paper presented at the annual meeting of the American Education Research Association, Denver, CO.

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.

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.

Ferron, J. M., Foster-Johnson, L., & Kromrey, J. (2003). The functioning of single-case randomization tests with and without random assignment. *The Journal of Experimental Education, 71,* 267-288.

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

Ferron, J. M., Joo, S. H., & Levin, J. R. (2017). A Monte-Carlo evaluation of masked-visual analysis in response-guided versus fixed-criteria multiple-baseline designs, *Journal of Applied Behavior Analysis*, doi: 10.1002/jaba.410

Ferron, J. M., & Levin, J. R. (2014). Single-case permutation and randomization statistical tests: Present status, promising new developments. In T. R. Kratochwill & J. R. Levin (Eds.) *Single-Case Intervention Research: Methodological and Statistical Advances.* Washington, DC: American Psychological Association.

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating casual effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods, 19,* 493-510.

Gast, D. L. (2009). Single Subject Research Methodology in Behavioral Sciences. New York, NY: Routledge.

Hembry, I., Bunuan, R., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2015). Estimation of a nonlinear intervention phase trajectory for multiple baseline design data. *Journal of Experimental Education, 83,* 514-546.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26,* 329-367.

Joo, S. H., Ferron, J. M., Moeyaert, M., Beretvas, S. N., & Van den Noortgate, W. (2017). Approaches for specifying the level-1 error structure when synthesizing single-case data. *Journal of Experimental Education,* Manuscript accepted for publication.

Kazdin, A. E. (2010). *Single-Case Research Designs: Methods for Clinical and Applied Settings*. New York, NY: Oxford University Press.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53,* 983-997.

Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5,* 155-164.

Kratochwill, T. R., Hitchcock, J. H., 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.

Mawhinney, T. C., & Austin, J. (1999). Speed and accuracy of data analysts' behavior using methods of equal interval graphic data charts, standard celeration charts, and statistical control charts. *Journal of Organizational Behavior Management, 18,* 5-45.

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2016). The misspecification of the covariance structures in multilevel models for single-case data: A Monte Carlo simulation study. *The Journal of Experimental Education, 84,* 473-509.Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40,* 357-367.

Petit-Bois, M., Baek, 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.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185-193.

SAS Institute. (2014). *SAS 9.4,* Computer Software. SAS Inst.

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52,* 109-122.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43,* 971-980.

Solmi, F., & Onghena, P. (2014). Combining p-values in replicated single-case experiments with multivariate outcome. *Neuropsychological Rehabilitation, 24,* 607-633.

Solomon, B. G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification, 38,* 477-496.

Todman, J. & Dugard, P. (1999). Accessible randomization tests for single-case and small-n experimental designs in AAC research. *Augmentative and Alternative Communication, 15,* 69-82.

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18,* 325-346.

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, 35,* 1-10.

Appendix

Change in Level Model

To illustrate the change in level model in the context of single-case research design studies, consider an observation at time *i* for case *j*, Y_{ij}. Also, consider an indicator variable for the treatment phase T_{ij} such that T_{ij} = 0 if the observation Y_{ij} is in the baseline phase and T_{ij} = 1 if the observation Y_{ij} is in the treatment phase. Then, a first-level equation of the model is,

$$Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + e_{ij} \quad (1)$$

where β_{0j} is the mean baseline level for case *j*, β_{1j} is the average treatment effect across cases (i.e., the shift in the mean level of responding that occurs with intervention), and e_{ij} is the error, frequently assumed to be independently sampled from a normal distribution with a mean of 0 and variance σ_e^2. However, previous studies with multilevel models for multiple-baseline designs also investigated the efficacy of first-order autoregressive models (Petit-Bois, Baek, Van den Noortgate, Beretvas, & Ferron, 2016), as well as heterogeneous level-1 error structures across phases (Joo et al., 2017) and across cases (Baek & Ferron, 2013). Variation in level of response and treatment effect across the cases can be obtained using a second-level model,

$$\beta_{0j} = \gamma_{00} + r_{0j} \quad (2.1)$$

$$\beta_{1j} = \gamma_{10} + r_{1j} \quad (2.2)$$

where γ_{00} is the across-case average baseline level, γ_{10} is the across-case average treatment effect, and the error terms, r_{0j} and r_{1j} have variances σ_{r0}^2 and σ_{r1}^2, respectively. Although the residuals, r_{0j} and r_{1j}, can be assumed to be either independent or correlated in general multilevel models, they are often assumed independent because more complex second-level error structures have been found to increase the bias in associated variance estimates (Moeyaert, Ugille, Ferron, Moeyaert, & Van den Noortgate, 2016).

It is worthwhile to note that the change in level model in Equations 1, 2.1 and 2.2 is based on the assumption of no trend in baseline and no effect of intervention on trend. However, in practical settings, the assumption of a lack of trends may be violated depending on the characteristics of behavior being measured and the treatment. Also, the presence of trend in single-case data has been evident as shown in the previous studies (e.g., Solomon, 2014). It is possible that observations in the treatment phase show a linear increasing (or decreasing) trend due to a delay in the treatment effect or show a nonlinear increasing (or decreasing) trend due to floor/ceiling effect. For these circumstances, the assumption should be relaxed, for instance by including more predictor variables in the first-level equation of the multilevel model to model trends in addition to a treatment effect.

Change in Level and Trend Model

In the change in level and trend model it is assumed that there may be linear trends that may differ across baseline and treatment phases. Consequently, more parameters should be specified in the first-level and second-level equations. Consider a new time variable ($Time_{ij}$) specifying the i^{th} observation for case j. Furthermore, by combining the time variable with the phase variable, T_{ij}, an interaction term ($T_{ij} * Time_{ij}$) can be created. $Time_{ij}$ is centered such that 0 corresponds to the first intervention observation. Then, the first-level equation can be formed,

$$Y_{ij} = \beta_{0j} + \beta_{1j} T_{ij} + \beta_{2j} Time_{ij} + \beta_{3j} T_{ij} * Time_{ij} + e_{ij} \quad (3)$$

where β_{0j} is the projected baseline level at the first intervention observation, β_{1j} is the average treatment effect across cases on the level of response at the first intervention observation (i.e., the immediate effect of intervention), β_{2j} is the trend effect in the baseline phase, and β_{3j} is the effect of intervention on the trend (i.e., the change in trend between baseline and treatment

phases). Similar to the change in level model, e_{ij} indicates error randomly sampled from a normal distribution with a mean of 0 and variance σ_e^2. The second-level equations could be

$$\beta_{0j} = \gamma_{00} + r_{0j} \quad (4.1)$$

$$\beta_{1j} = \gamma_{10} + r_{1j} \quad (4.2)$$

$$\beta_{2j} = \gamma_{20} + r_{2j} \quad (4.3)$$

$$\beta_{3j} = \gamma_{30} + r_{3j} \quad (4.4)$$

where γ_{00} is the across-case average projected baseline level, γ_{10} is the across-case average immediate treatment effect, γ_{20} is the across-case average baseline slope, and γ_{30} is the across-case average treatment effect on slope. The error terms r_{0j}, r_{1j}, r_{2j} and r_{3j} have variances σ_{r0}^2, σ_{r1}^2, σ_{r2}^2 and σ_{r3}^2, respectively, and again the residuals are typically assumed to be independent.

Tables

Table 1
Simulation Results for Fixed Series Length Condition

Simple Level-2 Error Condition									
		Change in Level Model				Change in Level and Slope Model			
Effect	*N*	Bias	RMSE	CI Cov	TI/Pwr	Bias	RMSE	CI Cov	TI/Pwr
0.0	19.0	-.001	.241	.971	.029	-.006	.419	.957	.043
1.0	19.0	.001	.239	.975	.805	-.005	.419	.958	.585
2.0	19.0	.001	.247	.968	.999	.000	.424	.963	.981
3.0	19.0	.008	.247	.971	1.000	.014	.420	.960	.998
4.0	19.0	.000	.243	.973	1.000	.014	.420	.964	1.000
Complex Level-2 Error Condition									
		Change in Level Model				Change in Level and Slope Model			
Effect	*N*	Bias	RMSE	CI Cov	TI/Pwr	Bias	RMSE	CI Cov	TI/Pwr
0.0	19.0	-.004	.435	.945	.055	-.005	.562	.932	.068
1.0	19.0	.015	.434	.938	.395	.003	.549	.941	.377
2.0	19.0	.011	.435	.939	.873	.014	.555	.938	.847
3.0	19.0	-.004	.431	.932	.994	-.002	.550	.941	.981
4.0	19.0	-.008	.439	.935	1.000	-.006	.559	.930	.996

Note. Effect = effect size, *N* = series lengths, Bias = relative bias, RMSE = root mean square error, CI Cov = confidence interval coverage, Pwr = power, TI = type I error. When effect size equals to zero, Type I error is computed and power is computed, otherwise.

Table 2

Simulation Results for the Baseline Stability Condition

Simple Level-2 Error Condition										
			Change in Level Model				Change in Level and Slope Model			
Analysis	Effect	Avg *N*	Bias	RMSE	CI Cov	TI/Pwr	Bias	RMSE	CI Cov	TI/Pwr
Masked	0.0	23.4	.001	.220	.973	.027	.012	.379	.961	.039
	1.0	23.5	.006	.223	.969	.850	.022	.383	.962	.690
	2.0	23.5	.001	.221	.968	.999	.016	.390	.957	.993
	3.0	23.5	.001	.223	.971	1.000	.013	.386	.963	.999
	4.0	23.5	-.003	.222	.967	1.000	.010	.383	.962	1.000
Unmasked	0.0	21.1	.001	.237	.968	.032	.003	.399	.954	.046
	1.0	21.1	-.002	.232	.974	.823	.007	.396	.959	.628
	2.0	21.1	.005	.239	.966	.998	.007	.391	.965	.991
	3.0	21.1	-.006	.241	.967	1.000	-.008	.403	.965	.998
	4.0	21.1	-.002	.242	.968	1.000	.000	.403	.959	1.000
Complex Level-2 Error Condition										
			Change in Level Model				Change in Level and Slope Model			
Analysis	Effect	Avg *N*	Bias	RMSE	CI Cov	TI/Pwr	Bias	RMSE	CI Cov	TI/Pwr
Masked	0.0	23.5	-.008	.412	.943	.057	-.001	.511	.940	.060
	1.0	23.5	-.003	.409	.943	.397	.010	.509	.940	.396
	2.0	23.5	.007	.430	.936	.874	.020	.548	.925	.861
	3.0	23.5	-.003	.426	.935	.994	.010	.531	.936	.984
	4.0	23.5	.021	.415	.947	1.000	.026	.523	.936	.997
Unmasked	0.0	21.1	.001	.430	.940	.060	-.001	.536	.938	.062
	1.0	21.1	-.006	.420	.943	.389	-.006	.533	.939	.389
	2.0	21.1	-.005	.425	.943	.866	-.005	.532	.940	.849
	3.0	21.1	-.018	.422	.940	.991	-.025	.532	.940	.978
	4.0	21.0	.008	.425	.940	1.000	.012	.537	.936	.998

Note. Analysis = ongoing analysis, Effect = effect size, Avg *N* = average series lengths across replications, Bias = relative bias, RMSE = root mean square error, CI Cov = confidence interval coverage, Pwr = power, TI = type I error. When effect size equals to zero, Type I error is computed and power is computed, otherwise.

Table 3.

Simulation Results for the Response to Intervention Condition

Simple Level-2 Error Condition										
			Change in Level Model				Change in Level and Slope Model			
Analysis	Effect	Avg *N*	Bias	RMSE	CI Cov	TI/Pwr	Bias	RMSE	CI Cov	TI/Pwr
Masked	0.0	34.3	.000	.199	.970	.030	-.012	.308	.959	.041
	1.0	29.6	.008	.215	.963	.865	.019	.343	.952	.795
	2.0	22.8	.017	.230	.966	.998	.023	.388	.961	.991
	3.0	19.8	.004	.240	.974	1.000	.006	.417	.955	.998
	4.0	19.1	-.001	.249	.967	1.000	.002	.426	.957	1.000
Unmasked	0.0	36.4	.004	.194	.971	.029	.003	.296	.961	.039
	1.0	31.4	.012	.207	.966	.877	.012	.332	.958	.810
	2.0	23.5	.018	.228	.971	1.000	.029	.380	.963	.993
	3.0	19.9	.008	.241	.968	1.000	.016	.415	.956	.998
	4.0	19.1	-.007	.248	.965	1.000	-.012	.429	.955	.999
Complex Level-2 Error Condition										
			Change in Level Model				Change in Level and Slope Model			
Analysis	Effect	Avg *N*	Bias	RMSE	CI Cov	TI/Pwr	Bias	RMSE	CI Cov	TI/Pwr
Masked	0.0	34.3	-.006	.413	.933	.067	-.016	.476	.930	.070
	1.0	29.8	-.002	.410	.947	.387	-.010	.479	.940	.423
	2.0	22.9	.022	.418	.929	.884	.026	.521	.931	.868
	3.0	19.8	.014	.427	.947	.993	.016	.534	.947	.980
	4.0	19.1	.007	.422	.944	1.000	.007	.534	.946	.996
Unmasked	0.0	36.3	.007	.396	.945	.055	.005	.455	.940	.060
	1.0	31.2	.014	.417	.936	.412	.021	.494	.925	.443
	2.0	23.6	.007	.419	.942	.879	.011	.520	.935	.862
	3.0	19.9	-.002	.420	.944	.992	.010	.534	.938	.978
	4.0	19.1	-.003	.427	.937	1.000	.000	.536	.940	.998

Note. Analysis = ongoing analysis, Effect = effect size, Avg *N* = average series lengths across replications, Bias = relative bias, RMSE = root mean square error, CI Cov = confidence interval coverage, Pwr = power, TI = type I error. When effect size equals to zero, Type I error is computed and power is computed, otherwise.

Table 4.

Simulation Results for Both Baseline Stability and Response to Intervention Condition

Simple Level-2 Error Condition										
			Change in Level Model				Change in Level and Slope Model			
Analysis	Effect	Avg *N*	Bias	RMSE	CI Cov	TI/Pwr	Bias	RMSE	CI Cov	TI/Pwr
Masked	0.0	37.9	.002	.185	.974	.026	-.003	.299	.959	.041
	1.0	35.8	.001	.190	.970	.908	-.002	.303	.962	.839
	2.0	31.2	.004	.200	.968	1.000	.011	.330	.958	.997
	3.0	28.5	.000	.206	.969	1.000	.008	.341	.961	1.000
	4.0	27.9	-.004	.210	.967	1.000	.010	.351	.961	.999
Unmasked	0.0	36.9	-.002	.191	.971	.029	-.004	.296	.963	.037
	1.0	32.7	.009	.206	.965	.879	.016	.329	.958	.812
	2.0	25.6	.011	.227	.968	.999	.017	.367	.961	.993
	3.0	22.0	.010	.230	.971	1.000	.025	.386	.961	.999
	4.0	21.2	.002	.239	.970	1.000	.005	.410	.955	.999
Complex Level-2 Error Condition										
			Change in Level Model				Change in Level and Slope Model			
Analysis	Effect	Avg *N*	Bias	RMSE	CI Cov	TI/Pwr	Bias	RMSE	CI Cov	TI/Pwr
Masked	0.0	37.9	-.003	.405	.935	.065	-.003	.460	.928	.072
	1.0	35.8	.001	.404	.939	.419	-.002	.469	.930	.443
	2.0	31.3	.009	.406	.944	.889	.016	.492	.937	.875
	3.0	28.5	-.004	.409	.939	.995	.008	.492	.937	.986
	4.0	27.8	.001	.416	.934	1.000	.009	.503	.931	.997
Unmasked	0.0	37.0	-.006	.408	.938	.062	-.011	.471	.922	.078
	1.0	32.8	.006	.401	.937	.409	.011	.478	.935	.440
	2.0	25.5	.021	.424	.937	.890	.028	.517	.930	.871
	3.0	22.0	.009	.415	.937	.991	.017	.528	.934	.982
	4.0	21.2	.001	.427	.929	1.000	.004	.535	.937	.998

Note. Analysis = ongoing analysis, Effect = effect size, Avg *N* = average series lengths across replications, Bias = relative bias, RMSE = root mean square error, CI Cov = confidence interval coverage, Pwr = power, TI = type I error. When effect size equals to zero, Type I error is computed and power is computed, otherwise.