How to Analyze Single-Case Studies in Education: an Illustration with Two Alternative Methods

Diep T. Nguyen, and John Ferron
University of South Florida, 4202 E. Fowler Avenue, Tampa, FL 33620

Abstract

In educational settings, single-case, or single-subject, research is the intensive study of individual learners through repeated measurement of an outcome while altering the conditions under which the learner is being observed. The most common single-case design is a multiple-baseline design, which is characterized by short interrupted time series data (typically 10 to 30 observations) on a small number of participants (commonly 4 to 8 persons) where the introduction of the intervention is temporally staggered across the participants. These features present some challenges for traditional analyses because small sample size at level 1 (short time series) or level 2 (few participants) makes it difficult to identify and estimate an appropriate model for the data. This paper illustrates two alternative methods for analyzing multiple-baseline studies: a variation of multilevel modeling and a non-parametric approach. A reanalysis of a previously published multiple-baseline study is presented to demonstrate the use of these two methods in single-subject educational research.

Key Words: Single-subject, Single-case, Multiple-Baseline, randomization test, multilevel model

1. Multiple-Baseline Studies

1.1 Introduction

Single-case designs are used in educational studies to examine the effects of interventions on individual students. Of the various types of single-case designs, the multiple-baseline (MB) design is the most commonly used (Shadish & Sullivan, 2011). It consists of collecting interrupted time-series data concurrently from multiple participants. Participants begin the study in a baseline phase (A) where repeated observations are made. Participants then transition to an intervention phase (B), where the transition from baseline to intervention is temporally staggered across participants (Baer, Wolf, & Risley, 1968; Christ, 2007). This temporal staggering of the intervention distinguishes the multiple-baseline design from a replicated AB design, and aids in arguments of internal validity because had an event other than treatment impacted the time-series it is not likely that such an event would create shifts in the series that coincided with the temporal staggering of the intervention.

At least five baseline observations and a minimum of five treatment phase observations are needed for each of at least three participants for the design to meet standards of the What Works Clearinghouse (Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, & Shadish, 2010). Researchers are encouraged to randomly assign participants to baseline lengths (Ferron, Moeyaert, Van den Noortgate, & Beretvas, in press; Kratochwill & Levin, 2010), to document high levels of reliability in the repeated measurements of the outcome variable (Kratochwill et al., 2010), and to graphically display and visually analyze the graphed data (Horner, Swaminathan, Sugai, & Smolkowski, 2012; Kratochwill et al., 2010).

There are several challenges to statistically analyzing the data that result from multiple-baseline studies. First, the number of participants, typically four to eight (Ferron, Farmer, & Owens, 2010), is relatively small compared to what is observed in group longitudinal studies in education. Second, the time series are relatively short for time series data (e.g., a median series length of 24 was reported by Ferron, Farmer, and Owens, 2010). The limited amount of data presents challenges because it can be difficult to identify an appropriate model and to accurately estimate the underlying parameters. We consider two approaches: 1) randomization tests, which do not require identification of a parametric model, and 2) multilevel models, which can capitalize on similarities across cases to improve parameter estimation.

1.2 Randomization Tests for MB data

Randomization tests (Edgington & Onghena, 2007) are nonparametric tests that determine statistical significance by comparing an obtained test statistic value to an empirically derived randomization distribution. The randomization distribution is constructed by repeatedly recalculating the test statistic value using the same values for the dependent variable but different values for the independent variable, where the different independent variable values reflect the possible random assignments. The randomization distribution can be constructed exhaustively so that it contains one test statistic value for each of the N possible assignments, or it can be approximated by randomly resampling with replacement a large number of assignments from the set of possible assignments.

A variety of randomization schemes have been developed for multiple-baseline studies. In one approach, researchers randomly assign participants to baselines of predetermined lengths (Wampold & Worsham, 1986). In a second approach, researchers establish a set of possible intervention start times for each time series, where the possible start times for one series do not overlap with the possible start times for another series, which ensures the baseline lengths vary across series. Researchers then randomly assign participants to baseline lengths and randomly selects a start times from within the designated sets of possible start times (Koehler & Levin, 1998). Finally, researchers may choose to gather baseline data until each series reaches stability and then randomly choose the order in which participants enter treatment (Ferron & Jones, 2006).

Randomization tests are statistically valid (i.e., they maintain a Type I error rate at or below the nominal alpha) as long as the design incorporates randomization, the randomization distribution is consistent with the randomization method, and the test statistic is chosen without knowledge of the chosen assignment (Edgington, 1980). The ability to control the Type I error rate with time series that are potentially autocorrelated and non-stationary (due to effects other than the treatment that impact the time series) is a

substantial motivation for using randomization tests with multiple-baseline data. Concerns with the application of randomization tests include: 1) the feasibility of randomization in some contexts, 2) statistical power, where the power depends on the size of the effect, the number of participants, the series length, the autocorrelation in the time series, and type of randomization that is chosen (Ferron & Sentovich, 2002), and 3) the lack of a parameter that indexes the size of the effect.

1.3 Multilevel Models for MB data

Statistical models have been proposed for multiple-baseline data where each series is modeled separately (e.g., Maggin et al., 2011; McKnight, McKean, & Huitema, 2000) and multilevel models have been proposed to simultaneously model all of the time series (Shadish, Kyse, & Rindskopf, 2013; Shadish & Rindskopf, 2007, Van den Noortgate & Onghena, 2003a; 2007).

In the simplest multilevel model the variation in outcome for the j^{th} participant is modeled as a function of a single dichotomous predictor that indicates the phase to which an observation belongs (0 = baseline, 1 = intervention). More specifically,

$$Y_{ij} = \beta_{0j} + \beta_{1j} Phase_{ij} + e_{ij} \qquad e_{ij} \sim N(0, \sigma_e^2) \tag{1}$$

Y_{ij} is the observed value at the i^{th} point in time for the j^{th} participant, β_{0j} is the baseline level for the j^{th} participant, β_{1j} is the treatment effect for the j^{th} participant (i.e., the shift in level between baseline and intervention phases), and e_{ij} is the error term. The regression coefficients are assumed to vary across participants,

$$\begin{aligned} \beta_{0j} &= \theta_{00} + u_{0j} \\ \beta_{1j} &= \theta_{10} + u_{1j} \end{aligned} \qquad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \\ \sigma_{u_0 u_1} & \sigma^2_{u_1} \end{bmatrix} \right) \tag{2}$$

where θ_{00} indicates the average baseline level across participants, θ_{10} refers to the average treatment effect across participants, and u_{0j} and u_{1j} are the second-level errors.

A variety of extensions of this model are possible, including the modeling of autocorrelation in the level-1 errors (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009), allowing the level-1 error variance-covariance parameters to vary across participants (Baek & Ferron, 2013), modeling linear trends within the baseline and treatment phases (Van den Noortgate & Onghena, 2003b), modeling non-linear trends within the treatment phases (Hembry, Bunuan, Beretvas, Ferron, & Van den Noortgate, in press), modeling outcome variables that are counts (Shadish et al., 2013; Shadish & Rindskopf, 2007), adding a third level to model variation between sites or studies (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2014; Van den Noortgate & Onghena, 2008), and treating differences between cases as fixed effects as opposed to random effects (e.g., Rindskopf & Ferron, 2014).

The use of multilevel models allows researchers to estimate parameters that can answer research questions about the size of the average treatment effect and the variation in the treatment effect across participants. Concerns with the application of multilevel

modeling to multiple-baseline data stem from uncertainties about whether an estimated model is correctly specified, and the small sample sizes that are characteristic of multiple-baseline studies. When models are correctly specified accurate inferences about the average treatment effect can be made with as few as four participants if the Kenward-Roger (1997) approach to estimating standard errors and degrees of freedom is used, but the between person variance estimates tend to be biased by the small sample size (Ferron et al., 2009).

2. Application

2.1 Study Description

2.1.1 Description of Application.

Data for this study were a subset obtained from the Ingersoll and Wainer (2013) investigation, which used a multiple-baseline design to examine the effect of a treatment on mothers' fidelity of implementing a technique designed to increase speech in children with Autism Spectrum Disorder. Data included in our analyses were for the fidelity outcome measure and the subset of five mothers who received treatment two times a week. These mothers (and their children) were randomly assigned to pre-determined baseline periods. The series lengths ranged from three to 11 for the baseline phase and 20 to 24 for the treatment phase. Our purpose was to reanalyze this data using: (1) a randomization test to determine if there is evidence of a treatment effect and (2) multilevel modeling to estimate the size of the treatment effect and the degree to which it varies across participants.

2.1.2 Graph of Data.

Figure 1 provides a graphical display of the fidelity ratings for each of the five mothers as a function of time. Visual inspection of the graph suggests the average fidelity ratings increase after implementation of the intervention. Further statistical analyses will allow us to rule out chance as a viable explanation for these observed increases and to quantify the size of the effect.

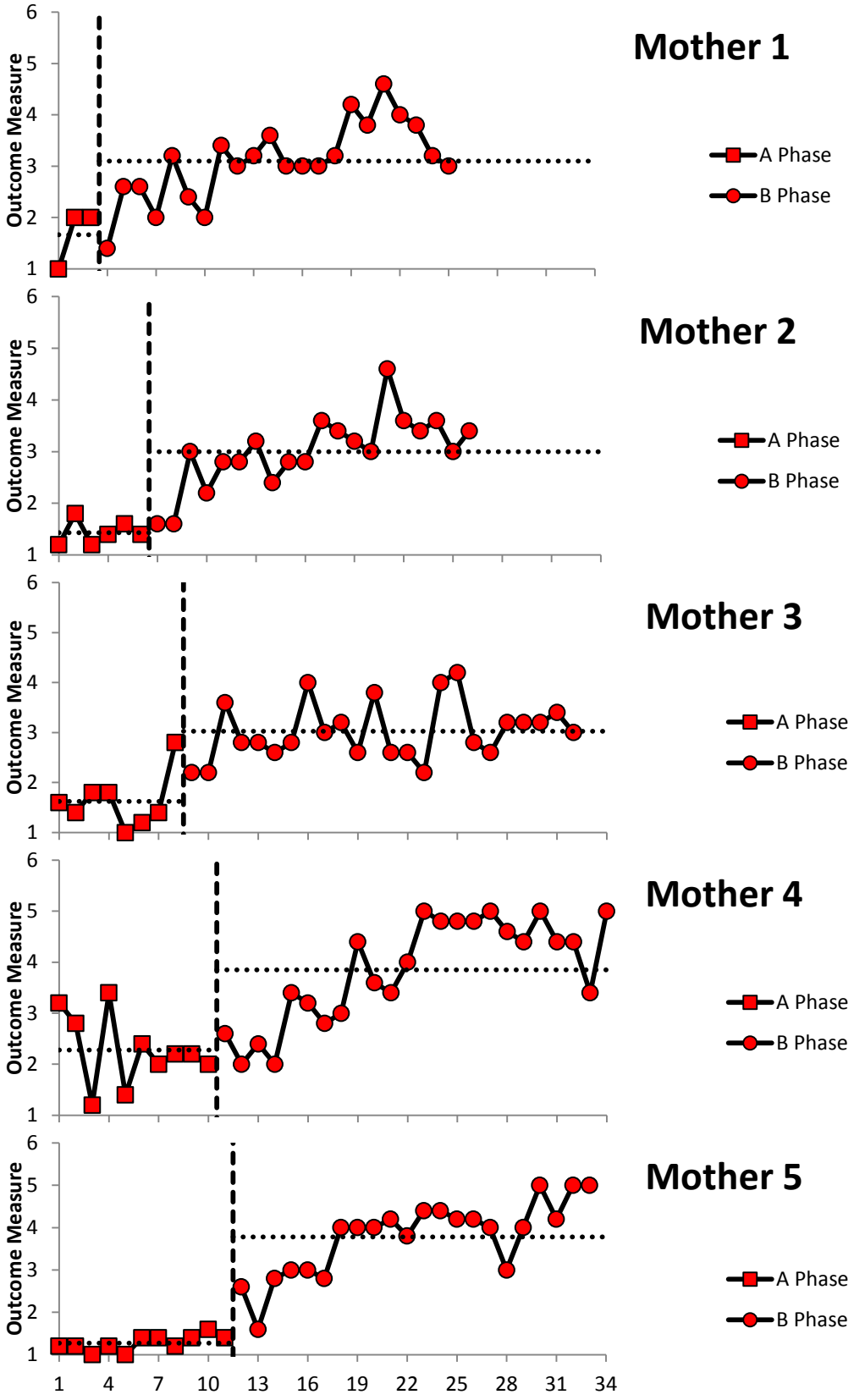

Figure 1: Graphs for average fidelity of all mothers

2.2 Randomization Test

To test the null hypothesis of no treatment effect, we used the Wampold-Worsham randomization test for multiple-baseline designs using the ExPRT program **(**Levin, Evmenova, & Gafurov, 2014).

The mean difference between treatment and baseline phases was 1.43, 1.57, 1.4, 1.57, and 2.51, respectively, and thus the observed test statistic was 1.70. This value was compared to the randomization distribution of 120 test statistic values that were obtained by finding the average mean difference for each of the possible random assignments. The observed test statistic was the 3rd largest value in the randomization distribution, and thus the one-tailed test resulted in a *p*-value of .0250, and thus we reject the null hypothesis that the treatment had no effect. As mentioned earlier, we do not have a parameter estimate that can be interpreted as an effect size with the randomization test. This motivates our next analysis.

2.3 Multilevel Model

2.3.1 Model 1: No Trend

Model 1 is a hierarchical model of parent average fidelity as a function of treatment where it is assumed that there is no trend in either phase. The model also assumes a first-order autocorrelation and that the baseline levels and the effects are allowed to vary across mothers. The model was estimated using restricted maximum likelihood estimation and the Kenward-Roger approach for fixed effect inferences as implemented in the MIXED Procedure of SAS.

Equations for model 1:
Level -1:

$$Y_{ij} = \pi_{0j} + \pi_{1j} * Phase_{ij} + e_{ij}$$

where Y_{ti} is average fidelity of mother *i* at time point t, π_{0i} is baseline level of fidelity, π_{1i} is shift in average fidelity level, $Phase_{ti}$ equals to 0 if mother *i* is in baseline and equals to 1 when mother *i* is in the treatment condition at time point *t*, e_{ti} is level-1 error term.

Level - 2:

$$\pi_{0j} = \beta_{00} + r_{oj}$$
$$\pi_{1j} = \beta_{10} + r_{1j}$$

where β_{00} is overall average baseline level across all five mothers, β_{10} is overall average treatment effect, r_{oi} and r_{1i} are variations of base level and treatment effect from the average values for person *i*.

The level-2 equations can be substituted into the level-1 equation to create the combined model or mixed linear model below:

$$Y_{ij} = \beta_{00} + \beta_{10} * Phase_{ij} + r_{oj} + r_{1j} * Phase_{ij} + e_{ij}$$

Results of Model 1 are presented in Table 1. The results showed that the estimated average baseline and the average shift in levels (or average treatment effect) were 1.65 and 1.64, respectively. This would place the average of the treatment phase at 3.29. The treatment effect was statistically significant with a *p*-value that equals .0015.

The variance components are also shown in Table 1 for completeness, but the level-2 variances should be viewed cautiously given the small sample size coupled with the use of restricted maximum likelihood estimation. The autocorrelation of .68 may be partially attributed to trends in the treatment phase that were not modeled. Based on Figure 1, it appears that there is an upward trend during the treatment phase for each of the participants.

Table 1: Hierarchical modeling regression results of average mother fidelity for Model 1

Parameter	*Estimate*	*se*	*df*	*t*	*p*
Fixed Effects					
Average Baseline Level	1.65	0.17	4.39	9.58	.0004
Average Treatment Effect	1.64	0.22	4.06	7.60	.0015
Variances					
Between Baseline Levels	0.13	0.10			
Between Treatment Effects	0.07	0.19			
Within Baseline Phase Variance	0.23	0.06			
Baseline Phase Autocorrelation	-.32	.19			
Within Treatment Phase Variance	0.79	0.22			
Treatment Phase Autocorrelation	.68	.09			

2.3.2 Model 2: With Trend

Model 2 is a hierarchical model of parent average fidelity as a function of treatment and time into treatment. In this model, we will allow no trend in baseline, a linear trend in treatment, and variation between baseline levels across mothers as well as variation in treatment effects across mothers.

Equations for Model 2:

Level -1:

$$Y_{ij} = \pi_{0j} + \pi_{1j} * Phase_{ij} + \pi_{2j} * Phase_{ij} * Time_{ij} + e_{ij}$$

Where π_{0i} is the baseline level, π_{1i} represents the shift in level between the treatment phase and the baseline extrapolation approximately half way through the intervention (i.e., Time was centered so that 0 corresponded to the 11^{th} treatment observation), and π_{2i} is the slope in the treatment phase for person *i*.

Level - 2:

$$\pi_{0j} = \beta_{00} + r_{oj}$$
$$\pi_{1j} = \beta_{10} + r_{1j}$$
$$\pi_{2j} = \beta_{20} + r_{2j}$$

where β_{00} can be interpreted as the average baseline level across mothers, β_{10} the average shift in level across mothers, and β_{20} the average slope during treatment across mothers, r_{0i}, r_{1i}, and r_{2i} represent the deviations of the i^{th} mother's values from these averages.

Combined model:

$$Y_{ij} = \beta_{00} + \beta_{10} * Phase_{ij} + \beta_{20} * Phase_{ij} * Time_{ij} + r_{oi} + r_{1j} * Phase_{ij} + r_{2j} * Phase_{ij} * Time_{ij} + e_{ij}$$

Table 2 shows the hierarchical modeling regression results for Model 2. The estimated average baseline effect and the average shift in levels across participants in this model were statistically significant and identical with those estimates in Model 1. However, results from Model 2 showed that there was a slope in the treatment phase (0.08) and this trend was statistically significant (p = .0065). In addition, the autocorrelation of the treatment phase was smaller in this model than in Model 1 (i.e. .29 in this model in comparison with .68 in Model 1).

Table 2: Hierarchical modeling regression results of average mother fidelity for Model 2

Parameter	*Estimate*	*se*	*df*	*t*	*p*
Fixed Effects					
Average Baseline Level	1.65	0.17	4.39	9.71	.0004
Average Treatment Effect	1.64	0.21	4.22	7.78	.0012
Average Treatment Slope	0.08	0.02	4.49	4.79	.0065
Variances					
Between Baseline Levels	0.12	0.10			
Between Treatment Effects	0.17	0.15			
Between Treatment Slopes	<0.01	<0.01			
Within Baseline Phase Variance	0.23	0.06			
Baseline Phase Autocorrelation	-.32	.20			
Within Treatment Phase Variance	0.34	0.06			
Treatment Phase Autocorrelation	.29	.11			

3. Conclusions and Recommendations

The two analyses we ran complement each other and the visual analysis of the authors. By conducting a randomization test we were able to formally rule out chance as an explanation for the changes observed in the graph, and by estimating a multilevel model we were able to gauge the size of the treatment effect. We would recommend that those planning multiple-baseline studies incorporate randomization into their design so that a valid randomization test can be conducted. We also recommend that those analyzing single-case data consider using multiple analyses, as we have illustrated here. Multilevel models provide a reasonable way to estimate effects, but rely on many assumptions that are difficult to test with a small data set, and thus it is helpful to also conduct a randomization test. The randomization test does not require distributional assumptions, but by itself it is also limited because it does not provide an estimate of an effect size parameter. Together these analyses provide a more complete understanding of the data and the effect of treatment on the outcome.

References

Baek, E., & 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.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *1*, 91-97.

Christ, T. J. (2007). Experimental control and threats to internal validity of concurrent and nonconcurrent multiple baseline designs. *Psychology in the Schools, 44*, 451-459.

Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics, 5*, 235-251.

Edgington, E. S., & Onghena, P. (2007). *Randomization Tests* (4th ed.). Boca Raton, FL: Chapman & Hall.

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., Farmer, J., and Owens, C. (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., & 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., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (in press). Estimating casual effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*.

Ferron, J., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education, 70*, 165-178.

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

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children, 35*, 269-290.

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

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*, 124-144.

Levin, J. R., Evmenova, A. S., & Gafurov, B. S. (2014). The single-case data-analysis *ExPRT* (*Excel*[7] *Package of Randomization Tests*). In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 185-219). Washington, DC: American Psychological Association.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49*, 301-321.

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods, 5*, 87-101.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2014). Three-level analysis of single-case experimental data: Empirical validation. *The Journal of Experimental Education, 82*, 1-21.

Rindskopf, D., & Ferron, J. (2014). Using multilevel models to analyze single-case design data. In T. R. Kratochwill & J. R. Levin (Eds.) *Single-Case Intervention Research: Methodological and Data-Analysis Advances* (pp. 221-246). American Psychological Association.

Shadish, W.R., Kyse, E.N., & Rindskopf, D.M. (2013). Analyzing data from single-case designs using multilevel models: new applications and some agenda items for future research. *Psychological Methods, 18*, 385-405.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation, 113*, 95-109.

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.

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, Instruments, & Computers, 35*, 1-10.

Van den Noortgate, W., Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today*, 8(2), 196-209.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 142-151.