Multilevel models for multiple-baseline data: modeling across-participant variation in autocorrelation and residual variance

Eun Kyeng Baek · John M. Ferron

Published online: 18 July 2012
© Psychonomic Society, Inc. 2012

Abstract Multilevel models (MLM) have been used as a method for analyzing multiple-baseline single-case data. However, some concerns can be raised because the models that have been used assume that the Level-1 error covariance matrix is the same for all participants. The purpose of this study was to extend the application of MLM of single-case data in order to accommodate across-participant variation in the Level-1 residual variance and autocorrelation. This more general model was then used in the analysis of single-case data sets to illustrate the method, to estimate the degree to which the autocorrelation and residual variances differed across participants, and to examine whether inferences about treatment effects were sensitive to whether or not the Level-1 error covariance matrix was allowed to vary across participants. The results from the analyses of five published studies showed that when the Level-1 error covariance matrix was allowed to vary across participants, some relatively large differences in autocorrelation estimates and error variance estimates emerged. The changes in modeling the variance structure did not change the conclusions about which fixed effects were statistically significant in most of the studies, but there was one exception. The fit indices did not consistently support selecting either the more complex covariance structure, which allowed the covariance parameters to vary across participants, or the simpler covariance structure. Given the uncertainty in model specification that may arise when modeling single-case data, researchers should consider conducting sensitivity analyses to examine the degree to which their conclusions are sensitive to modeling choices.

Keywords Multiple baseline · Level-1 error structure · Single-case study · Autocorrelation

Single-case research is the intensive study of a single case (e.g., a single participant or a single group, such as a class of students with data collected at the classroom level) by repeatedly measuring an outcome, allowing for the inspection of intervention effects over time. Basic single-case design includes two phases: a baseline phase, consisting of a series of observations prior to the implementation of a treatment, and a treatment phase, consisting of a series of observations following the implementation of a treatment. Although inferences about treatment effects are made from studies that utilize the basic single-case design, the validity of these inferences can be questioned because a shift in the time series may be the result of something other than the treatment, such as history (i.e., an event that happened to occur around the time of the intervention) or natural maturation (Shadish, Cook, & Campbell, 2002). In an effort to reduce the possible alternative explanations for shifts in time-series data, researchers often turn to more complex single-case designs, such as the reversal design and the multiple-baseline design.

The multiple-baseline design is preferred among single-case designs. The multiple-baseline design includes interrupted time-series data from multiple participants (or behaviors, or settings) in which an intervention occurred at different times within the different series. Therefore, it allows researchers to study several cases together. In addition, the staggering of the intervention makes it more difficult for shifts to be attributed to maturation or history. This strengthens the internal validity without requiring the researcher to remove the treatment (Barlow & Hersen, 1984). The utility of multiple-baseline designs has been well established in a variety of fields within psychology and education, including school psychology (Skinner, 2004; Van den Noortgate & Onghena, 2003a), special education

E. K. Baek (✉) · J. M. Ferron
Department of Measurement and Research in College of
Education, University of South Florida,
EDU 105, 4202 E. Fowler Avenue,
Tampa, FL 33620-7750, USA
e-mail: ebaek@usf.edu

(Swanson & Sachse-Lee, 2000), counseling (Sharpley, 1981), and reading (Neuman & McCormick, 1995).

Autocorrelation

In single-case data, the potential exists for nonmodeled effects to influence multiple consecutive observations in the series, leading to dependencies in the errors, or *autocorrelation*. For an individual, each observation can be deviated from the individual's growth trajectory due to a number of nonmodeled factors in one's life (e.g., being ill, moving to a new school). For example, consider the positive social behavior of a child. The social behavior of the child may increase at a constant rate, which could be illustrated by her growth trajectory, but her observations may deviate from this trajectory. This could be due to the recent death of her pet. She might feel sad and depressed, which could affect observations of the outcome. These disturbances may affect the next two or more sequential observations. As can be seen in this example, in single-case data, it is plausible that the errors that were closer together in time would be more similar, which, in turn, leads to autocorrelation. If the factors affecting the errors are more similar at close points in time, the errors may also be more similar at close points in time. This finding that there may be autocorrelation among the errors has led to the development of a variety of possible error structures, such as unstructured, compound symmetry, banded toeplitz or moving average, first-order autoregressive [AR(1)], AR(1) plus a diagonal, AR(1) plus a common covariance, and an AR(1) generalization for unequally spaced observations (Goldstein, 1995; Goldstein, Healy, & Rasbash, 1994; Heitjan & Sharma, 1997; Jennrich & Schluchter, 1986; Ware, 1985; Wolfinger, 1993; Yang & Goldstein, 1996). Figure 1 illustrates the first-order autoregressive structure, a symmetric two-toeplitz structure (or first-order moving-average structure), and the more general symmetric toeplitz structure (Lütkepohl, 1996).

The first-order autoregressive structure [AR(1)] contains two parameters, with σ^2 to represent the variance of errors and ρ to represent the autocorrelation coefficient. The correlation

for errors that are separated by one point in time is ρ , and the correlations for errors that are separated by two, three, and n points in time are ρ^2 , ρ^3 , and ρ^n , respectively. The symmetric two-toeplitz structure [TOEP(2)] also contains two parameters, with σ^2 on the main diagonal to represent error variance and σ_1 to represent the covariance between errors that are separated by one point in time. Unlike the AR(1) structure, it assumes that the correlation for errors that are separated by more than one point in time is zero. The symmetric toeplitz contains as many parameters as time points, with σ^2 on the main diagonal to represent error variance, and σ_1 , σ_2 , σ_3 , and σ_n to represent the covariances between errors that are separated by one point in time, two points in time, three points in time, and n points in time, respectively. Unlike the AR(1) structure, the correlations between errors are not restricted to be less when they are farther apart in time.

Statistical methods for single-case study

Statistical models for analyzing single-case data have been considered for decades. In the early seventies, several ANOVA-based procedures were suggested (Gentile, Roden, & Klein, 1972; Shine & Bower, 1971), and later ordinary least squares (OLS) regression procedures were illustrated (Center, Skiba, & Casey, 1985–1986; Huitema & McKean, 1998). Concerns with these general linear model (GLM) based approaches have been expressed because they assume that the errors in the statistical model are independent (e.g., Kratochwill et al., 1974; McKnight, McKean, & Huitema, 2000), as opposed to autocorrelated.

The effect of positive autocorrelation on GLM-based methods has been well documented. In general, we would expect the regression coefficients to be unbiased, but the standard errors of the regression coefficients to be underestimated (Neter, Wasserman, & Kutner, 1990). This, in turn, leads to confidence intervals that tend to be too small and significance tests with inflated Type I error rates. The degree to which confidence intervals and significance tests are impacted increases with the level of autocorrelation, and has been documented for various series lengths and patterns of autocorrelation (Beretvas & Chung, 2008; Greenwood & Matyas, 1990; Huitema, McKean, & McKnight, 1999; Scheffé, 1959; Toothaker, Banz, Noble, Camp, & Davis, 1983).

In order to handle this issue, some alternative approaches have been suggested. Under the generalized least squares (GLS) regression method, the autocorrelation can be estimated for an individual, and the estimate can be included in subsequent analyses to account for serial dependence (Maggin et al., 2011). Several methods are available to estimate the autocorrelations under the GLS regression framework (Solanas, Manolov, &

Fig. 1 Example covariance structures for dependent errors

Sierra, 2010), and McKnight, McKean, and Huitema (2000) found that a double-bootstrapping procedure for refining autocorrelation estimates was more accurate than several more traditional GLS methods.

Multilevel models have been suggested as an another alternative method for analyzing single-case data (Nugent, 1996; Shadish & Rindskopf, 2007; Shadish, Rindskopf, & Hedges, 2008; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). Multilevel modeling is flexible enough not only to allow for dependent error structures, but also to allow the covariance parameter values to differ across the participants in a single analysis. For this approach, any of the variety of error structures illustrated in Fig. 1 can be assumed.

There are many advantages to Using multilevel modeling for analyzing multiple-baseline single-case data (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009): (1) It allows one to use data from multiple cases in a single analysis, while OLS or GLS approaches require individual analysis for each case of the multiple-baseline design; (2) it allows one to estimate parameters corresponding to the shift in level and the shift in slope, and thus allows researchers to distinguish different types of treatment effects; (3) the models are flexible enough to handle various dependent error structures, heterogeneous variances, and moderating effects; and (4) software for estimating multilevel models is accessible and familiar to many applied researchers. Due to these benefits of using multilevel models to analyze multiple-baseline designs, the multilevel modeling approach was selected as the focus of this study.

Multilevel-modeling method

Multilevel modeling allows one to conceptualize longitudinal data at two distinct hierarchical levels. The first-level model describes change within an individual using a growth curve. The second-level model is used to determine how much variability exists across the individual growth curves. The basic multilevel model for single-case data across participants is shown in Eqs. 1 and 2. Equation 1 is for the first level of the multilevel model, which is comparable to the OLS regression model, where y_{ij} is the observed value (outcome) at the i th point in time for the j th participant, Phase is a dichotomous variable for which 0 indicates baseline and 1 indicates treatment, β_{0j} is the baseline mean for the j th participant, and β_{1j} is the difference in means between baseline and treatment phases for the j th participant (i.e., the treatment effect for the j th participant), and e_{ij} is error that leads to within-phase variation and could be assumed to follow any one of the variety of Level-1 covariance structures, such as a first-order autoregressive or a symmetric two-toeplitz structure.

$$y_{ij} = \beta_{0j} + \beta_{1j} \text{Phase}_{ij} + e_{ij} \quad (1)$$

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \mu_{0j} \\ \beta_{1j} &= \gamma_{10} + \mu_{1j} \end{aligned} \quad (2)$$

Equation 2 is for the second level of the multilevel model, which allows variation in baseline levels and treatment effects across participants. In this model, γ_{00} is the average baseline level, γ_{10} is the average treatment effect, and μ_{0j} and μ_{1j} are errors for the baseline level and the treatment effect, which are assumed to be multivariate-normally distributed. This general multilevel modeling equation can be expanded by adding more Level-1 predictors and Level-2 predictors.

Need for an extension of the application of multilevel modeling of single-case data

Although multilevel modeling has many appealing features for analyzing single-case data, some concerns can also be raised. It requires critical assumptions. Multilevel modeling typically assumes that (a) the degree of autocorrelation is the same for all participants and (b) the error variance is the same for all participants. Past applications of multilevel modeling to single-case data (e.g., Van den Noortgate & Onghena, 2003a, 2003b), as well as methodological studies into the statistical functioning of multilevel models with single-case data (Ferron et al., 2009; Ferron, Farmer, & Owens, 2010), have assumed that the Level-1 error covariance matrix is the same for all participants.

It is possible that the Level-1 residual variance and autocorrelation may not be the same across participants. In a recent study by Maggin et al. (2011), they used GLS regression analysis to compute effect sizes from single-case data. For their regression analysis, autocorrelation was taken into account by estimating a first-order autoregressive model [AR(1)] for each case. Two empirical single-case studies from published articles were reanalyzed, and the degree of autocorrelation was estimated for each case. In the first study, the autocorrelation ranged from $-.33$ to $.06$, and in the second study, the autocorrelation ranged from $-.37$ to $-.16$. As was seen in their study, autocorrelation could vary across cases. If variation in the Level-1 error covariance matrix were not taken into account, it might have an impact on the results (fixed effects and random effects). Thus, it is critical to examine the assumption that the Level-1 residual variance and autocorrelation are the same across participants, and the consequences of different approaches to modeling for the Level-1 error covariance matrix.

Purpose

Multilevel modeling allows for the possible dependency of errors, so that researchers need not assume that the errors in the statistical model are independent. However, assumptions about the covariance structure still exist for multilevel modeling applications. The purpose of this study was to extend the application of multilevel modeling of single-case data so as to accommodate across-participant variation in the Level-1 residual variance and autocorrelation.

This study illustrates a multilevel modeling method for estimating varying autocorrelation and residual variances across participants using an SAS program, and it makes these codes accessible to applied researchers for use in their own research. In addition, this approach was used in the reanalysis of published single-case data sets to examine the degree to which the autocorrelation and residual variances differ across participants and the degree to which inferences about treatment effects are sensitive to whether the Level-1 error covariance matrix is allowed to vary across participants.

Method

Search criteria

Single-case studies using multiple-baseline designs were found by searching the Web of Science database in October 2010, using the key word “multiple baseline” coupled with the publication year 2010. A total of 33 articles were first pulled by using those search keywords. Further criteria were then applied in order to select appropriate studies from among those 33 articles for analysis in this study. The criteria for selecting studies were (1) at least five data points per phase (Ferron et al., 2009; Swanson & Sachse-Lee, 2000), (2) at least four baselines (Kazdin & Kopel, 1975), (3) multiple-baseline across-participant designs, (4) full text available for the article, (5) data values that could be clearly read from the graph, (6) an outcome that could be modeled as continuous, and (7) trends within phases that appeared linear. Five studies (from four of the articles) were selected for analysis on the basis of the criteria listed above (Dufrene et al., 2010; Ingersoll & Lalonde, 2010; Koegel, Singh, & Koegel, 2010; Oddo et al., 2010).

Reliability of data extraction

For the selected five studies, the data were then independently extracted from the graphs in each study by the two authors. The interrater reliability was computed in order to examine the consistency of the data points extracted by the

two authors. The reliability estimates were greater than .97 for each study.

Analysis

The data for each study were then analyzed using two different two-level hierarchical models. More explicitly, for both hierarchical models, the Level-1 model was based on a linear model for each phase,

$$Y = \pi_{0i} + \pi_{1i}\text{Phase}_{it} + \pi_{2i}\text{Time}_{it} + \pi_{3i}\text{Phase}_{it} * \text{Time}_{it} + e_{it},$$

where Phase is a binary variable (0 = *baseline*, 1 = *treatment*) and Time is centered so that 0 corresponds to the first treatment observation. In addition, π_{0i} is the baseline intercept for the i th participant, π_{1i} is the immediate change in level between the baseline and treatment phases for the i th participant, π_{2i} is the baseline slope for the i th participant, and π_{3i} is the change in slopes between baseline and treatment phases. Level-1 error is represented by e_{it} , which indicates within-participant random variation and, in this study, is assumed to follow an AR(1) structure. The Level-2 model allowed the intercepts and slopes for each phase to vary randomly across participants:

$$\pi_{0i} = \beta_{00} + r_{0i},$$

$$\pi_{1i} = \beta_{10} + r_{1i},$$

$$\pi_{2i} = \beta_{20} + r_{2i},$$

$$\pi_{3i} = \beta_{30} + r_{3i},$$

where β_{00} is the average baseline intercept, β_{10} is the average immediate shift in level between phases, β_{20} is the average baseline slope, and β_{30} is the average difference between the baseline and treatment slopes. Level-2 errors are represented by r_{0i} , r_{1i} , r_{2i} , and r_{3i} , which are assumed to be distributed multivariate-normally with a diagonal covariance matrix.

The difference between the first and second hierarchical models is in the Level-1 error structure. For the first hierarchical model, the Level-1 error structure was modeled as first-order autoregressive [AR(1)]¹ where the

¹ When error structures are modeled as being autocorrelated, the first-order autoregressive function, AR(1), seems to be the most commonly used means to account for serial correlation (Murphy & Pituch, 2009), due to its being the simplest dependent error structure and to the technical limitations of some software. Thus, the present study has focused on AR(1) among various types of error structures.

values were held constant across participants (fixed model), while in the second hierarchical model, the Level-1 error structure was modeled as first-order autoregressive where the values were allowed to vary across participants (varying model). More specifically, a single autocorrelation and a single Level-1 error variance were estimated for the fixed model. These values represent the assumed common autocorrelation and common error variance of all participants in the study. Thus, if the autocorrelation is estimated as .3 and the error variance is estimated as 25, these two values are applied to all participants. On the other hand, autocorrelation and the variance of Level-1 error variance are estimated separately for each participant with the varying model; therefore, in this model every participant in the study is allowed to have unique autocorrelation and Level-1 error variance values.

The SAS code showing the specification of each hierarchical model, along with an illustration of how the data should be entered, is provided in the [Appendix](#). The code for the varying model differs from the more traditional fixed model in that it includes in the repeated statement the `GROUP = person` option, which allows for a separate estimate of the Level-1 error covariance matrix to be obtained for each group (or person).

Results

The results from the analyses of the five studies are presented in Table 1. When the Level-1 error covariance matrix is allowed to vary from participant to participant, some relatively large differences in parameter estimates emerged in terms of autocorrelation and residual variances. The changes in modeling of the variance structure did not change the conclusions about which fixed effects were statistically significant in most of the studies, but there was one exception.

Autocorrelation

The variation of autocorrelation across participants within a study was inspected by comparing the varying-model estimates within a study. In the varying model, the autocorrelation estimates varied across participants in all five studies, as is shown in Table 1. For example, in Study 3, the autocorrelation was estimated to be .22 when it was held constant across participants (fixed model), but when it was allowed to vary, it ranged from $-.04$ to $.46$ (varying model). In Study 4-2, the autocorrelation was estimated to be $-.38$ for the fixed model, but it ranged from $-.69$ to $.16$ for the varying model. Similar differences were founded for the rest of

the studies. In conclusion, autocorrelation estimates differed between the fixed model and the varying model for all five studies, in that the autocorrelation estimates varied across participants within the study in the varying model.

Level-1 error variance

The results for Level-1 error variance were inspected using the same approach that was used when examining autocorrelation. The variation of the error variance across participants within a study was inspected by comparing the varying-model estimates within each study. Across-participant differences in error variance estimates also emerged throughout all five studies. For example, in Study 2 the error variance was estimated to be 269.54 when it was constant across participants (fixed model), but it ranged from 164.41 to 795.62 when it was allowed to vary across participants (varying model). Similarly, in Study 4-1, the error variance was 1.74 for the fixed model, but it ranged from 0.68 to 3.29 for the varying model. In conclusion, similar to the autocorrelation results, the error variance estimates varied across participants within the study, and this finding was consistent across all five studies.

Fixed effects

The changes in variance structure did not change the conclusions about which fixed effects were statistically significant in most of the studies, but there were some exceptions. For example, in Study 1, the baseline intercept, the immediate change in level, and the baseline slope parameter values changed only slightly from the fixed model to the varying model, and the significance of the parameters did not change. The immediate changes in level were estimated as 0.06 for the fixed model and -0.18 for the varying model. With both models, the change was not statistically significant, and so from both models the researcher would conclude that there was not sufficient evidence to suggest that the treatment had an immediate effect on level. However, the interaction effect, which references the change in slope between phases, and thus part of the treatment effect, changed from being statistically significant ($\beta_{30} = 0.10$, $p = .01$) to not being statistically significant ($\beta_{30} = 0.07$, $p = .19$) when the model was changed from a common Level-1 error covariance matrix to a varying Level-1 error covariance matrix. Thus, the researcher using the fixed model would conclude that the treatment increased the slope, while a researcher using the varying model would not reach this conclusion.

Table 1 Parameter estimates (with standard errors) from the multilevel models with Level-1 error covariance structures that either were fixed or varied across participants

Model	Study 1		Study 2		Study 3		Study 4-1		Study 4-2	
	Fixed	Varying	Fixed	Varying	Fixed	Varying	Fixed	Varying	Fixed	Varying
Random Effects										
Intercept	0.30 (0.33)	0.36 (0.35)	81.59 (172.25)	0	454.72 (478.55)	448.91 (445.15)	0	0	0	0
Phase	0.02 (0.18)	0	70.10 (217.79)	0	779.68 (758.80)	645.01 (640.76)	0	0	0.12 (0.20)	0
Time	0.003 (0.003)	0	0	0	0	0	0.002 (0.01)	0	0.01 (0.01)	0.01 (0.01)
Interaction	0	0.01 (<0.01)	0	0	0	0	0	0	0.01 (0.03)	0.06 (0.06)
Variance	0.32 (0.05) ***	0.50 (0.16)**	269.54 (77.46)***	795.62 (716.66)	352.50 (76.32)***	145.77 (48.09)**	1.74 (0.54)***	3.29 (1.80)*	0.58 (0.14)***	1.12 (0.69)
AR(1)	.12 (.13)	.22 (.24)	.43 (.18)*	.89 (.10)***	.22 (.20)	-.04 (.41)	.52 (.15)***	.52 (.24)*	-.38 (.15)**	-.69 (.20)***
Variance	–	0.16 (0.05)***	–	164.41 (90.80)*	–	238.21 (98.19)**	–	0.68 (0.46)	–	0.25 (0.12)*
AR(1)	–	.23 (.23)	–	.41 (.37)	–	-.02 (.45)	–	.63 (.23)**	–	-.30 (.35)
Variance	–	0.25 (0.07)***	–	339.12 (135.51)**	–	475.60 (214.10)*	–	1.97 (1.29)	–	0.18 (0.09)*
AR(1)	–	.13 (.24)	–	.21 (.35)	–	.08 (.52)	–	.64 (.23)**	–	-.25 (.41)
Variance	–	0.37 (0.10)***	–	322.28 (169.79)*	–	679.11 (353.09)*	–	1.08 (0.45)**	–	0.67 (0.29)**
AR(1)	–	-.05 (.22)	–	.57 (.23)*	–	.46 (.31)	–	.07 (.43)	–	.16 (.32)
Fixed Effects										
Intercept	0.66 (0.33)	0.82 (0.35)	69.38 (6.12)*	73.69 (5.67)***	77.08 (15.37)**	78.71 (17.54)**	7.08 (0.70)***	5.94 (0.69)***	0.10 (0.27)	0.33 (0.28)
Phase	0.06 (0.23)	-0.18 (0.24)	28.80 (14.08)	32.14 (17.47)	32.98 (18.58)	33.44 (19.52)	-6.47 (0.73)***	-5.45 (0.71)***	3.26 (0.39)***	2.73 (0.37)***
Time	-0.05 (0.04)	-0.002 (0.02)	-0.25 (1.15)	-0.37 (1.48)	-1.12 (3.52)	0.12 (4.67)	0.74 (0.17)**	0.55 (0.13)**	-0.21 (0.09)	-0.16 (0.09)
Interaction	0.10 (0.04)**	0.07 (0.05)	2.45 (1.84)	2.35 (2.43)	1.76 (3.56)	0.27 (4.69)	-0.79 (0.24)**	-0.60 (0.21)*	0.32 (0.12)	0.39 (0.16)

* $p < .05$. ** $p < .01$. *** $p < .001$

In Study 4-1, the interaction effect was significant at an α level of .01 ($\beta_{30} = -0.79$, $p = .006$) for the fixed model, but was significant only at the .05 level ($\beta_{30} = -0.60$, $p = .03$) for the varying model. Using different models for the error structure did not alter conclusions about the rest of the parameters for fixed effects in Study 4-1. These differences in inferences about treatment effects were only observed within studies that showed significant treatment effects for the fixed model. For the rest of the studies, treatment effects were not significant in both the fixed and varying models, which indicates that the conclusions about the presence of a treatment effect were not sensitive to the method used to model the Level-1 covariance matrix.

Fit indices

The results of fit indices are provided in Table 2. There were inconsistencies across the fit indices within a study and across studies as to which model should be retained. For example, in Study 1, the small-sample-size-corrected version of the Akaike information criterion (AICC) would suggest retaining the fixed model, but the uncorrected Akaike information criterion (AIC) and the Bayesian information criterion (BIC) would suggest retaining the varying model. In Study 2, the AIC and AICC would suggest retaining the fixed model, but the BIC would suggest retaining the varying model. In Study 4-1, the AIC, AICC, and BIC would all suggest retaining the fixed model.

Table 2 Fit indices from the multilevel models with Level-1 error covariance structures that either were fixed or varied across participants

Model	Study 1		Study 2		Study 3		Study 4-1		Study 4-2	
	Fixed	Varying	Fixed	Varying	Fixed	Varying	Fixed	Varying	Fixed	Varying
–2 Log likelihood	233.8	223.1	541.6	534.1	551.5	542.6	161.5	154.7	131.0	120.0
AIC	243.8	243.1	549.6	550.1	559.5	562.6	167.5	170.7	141.0	140.0
AICC	244.3	245.2	550.3	552.8	560.2	567.1	168.1	174.4	142.4	145.9
BIC	240.7	237.0	547.2	545.2	557.0	556.5	165.7	165.8	137.9	133.8

AIC, Akaike information criterion; AICC, small-sample-size-corrected version of AIC; BIC, Bayesian information criterion

Moreover, the results of the fit indices across studies show that the BIC suggests retaining the varying model in Studies 1, 2, 3, and 4-2, but not in Study 4-1. The AICC suggests retaining the fixed model in all five studies. The AIC suggests retaining the varying model in Studies 1 and 4-2, but retaining the fixed model in Studies 2, 3, and 4-1. These inconsistencies make it difficult to be confident that either the fixed model or the varying model is the one that should be selected.

Discussion

This article suggests that there may be differences in autocorrelation and error variances across participants within a multiple-baseline study. It provides a way to model these differences within a multilevel framework and demonstrates some consequences of modeling or not modeling the possible differences between participants in Level-1 error covariance matrices. In our examination of five multiple-baseline studies, we found fairly large differences between the two models (fixed vs. varying) in terms of autocorrelation and error variances. These differences were found consistently across studies. The results also showed that these different models of the Level-1 error covariance structure can lead to different conclusions about the treatment effects.

The conclusions drawn from these findings, however, are limited in several ways. First, we only examined five multiple-baseline studies, which limits generalizability. In addition, we focused on a relatively simple multilevel model, which was appropriate for applications involving multiple baselines across participants when there was a continuous outcome and linear trends within phases. Multilevel modeling has been extended for applications with multiple baselines across settings and behaviors (Beretvas & Chung, 2010; Colvin, Sugai, Good, & Lee, 1997), as well as for noncontinuous outcomes, like count data (Beretvas & Wang, 2011; Shadish & Rindskopf, 2007; Shadish et al., 2008), and for nonlinear trends within phases (Beretvas, 2011; Shadish &

Rindskopf, 2007; Shadish et al., 2008). The theoretical rationale proposed here still holds for these more complex applications. More specifically, autocorrelation and Level-1 error variance may be expected to vary across settings or behaviors, or across participants with nonlinear trends or noncontinuous outcomes. In each of these more complex applications, a researcher may wonder whether across-participant variation should be modeled in autocorrelation and error variance, and whether doing so would alter conclusions about the treatment. Future work will need to extend the modeling of across-participant differences in autocorrelation and error variance to these more complex applications.

Future research should also help us to further address the handling of uncertainties that arise when specifying the Level-1 covariance structure. In our analyses of five multiple-baseline studies, the fit indices were not consistent in suggesting whether the fixed or the varying model was most appropriate. Given the uncertainty in model specification that may arise when modeling single-case data, more work in this area is needed. Monte Carlo simulation studies could be designed to index the accuracy of model selection using fit indices. Does one of the indices allow us to reliably distinguish situations in which the estimated autocorrelation and error variance differences between participants are due to sampling error from situations in which it is not? Monte Carlo studies could also be used to explore the consequences of misspecifying the Level-1 error covariance structure as varying when it does not vary across participants or as constant when it does vary across participants. The results from these studies would help inform applied researchers about best practices when modeling single-case data.

In the meantime, applied researchers should consider conducting sensitivity analyses to examine the degree to which their conclusions are sensitive to the approach used to model the Level-1 error variances. More specifically, researchers should consider estimating the treatment effects under both a model that assumes that the Level-1 error covariance structure varies across participants and a model

that assumes that it is constant across participants. The treatment effect estimates and inferences could then be compared across the models. When treatment effect inferences were supported under a variety of plausible Level-1

error variance assumptions, the researcher would have a stronger treatment effect claim than when the conclusion depended on how the Level-1 error covariance matrix was modeled.

Appendix

```

data j1;
input person phase time y;
datalines;
1 0 -5 23
1 0 -4 24
1 0 -3 27
1 0 -2 25
1 0 -1 22
1 1 0 30
1 1 1 35
1 1 2 33
1 1 3 37
1 1 4 35
2 0 -6 20
2 0 -5 18
2 0 -4 20
2 0 -3 19
2 0 -2 20
2 0 -1 20
2 1 0 25
    •
;

*Level-1 error covariance matrix constant across participants;

proc mixed data = j1 covtest;
class person;
model y = phase time phase*time / solution dfm = kenward;
random intercept phase time phase*time / sub = person;
repeated / sub = person type = ar(1);

*Level-1 error covariance matrix varies across participants;

proc mixed data = j1 covtest;
class person;

model y = phase time phase*time / solution dfm = kenward;
random intercept phase time phase*time / sub = person;
repeated / sub = person type = ar(1) group = person;

```

In the repeated statement for the varying model, the “GROUP = person” option allows for a separate estimate

of the Level-1 error covariance matrix to be obtained for each group (or person).

References

- (The superscripts ¹²³⁴ identify the articles from which Studies 1–3, 4-1, and 4-2 [see Tables 1 and 2] were drawn for our analysis.)
- Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). Needham Heights, MA: Allyn & Bacon.
- Beretvas, S. N. (2011). *Estimation of nonlinear trajectories for multiple baseline design data*. New Orleans, LA: Paper presented at the annual meeting of the American Educational Research Association.
- Beretvas, S. N., & Chung, H. (2008). An evaluation of modified R²-change effect size indices for single-subject experimental designs. *Evidence-based Communication Assessment and Intervention*, 2, 120–128.
- Beretvas, S. N., & Chung, H. (2010). *Modeling within-study dependence in multiple-baseline designs meta-analysis*. Denver, CO: Paper presented at the annual meeting of the American Educational Research Association.
- Beretvas, S. N., & Wang, D. (2011). *Synthesis of count data outcome trajectories in single-subjects' experimental design studies*. New Orleans, LA: Paper presented at the annual meeting of the American Educational Research Association.
- 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.
- Colvin, G., Sugai, G., Good, R. H., & Lee, Y. (1997). Using active supervision and precorrection to improve transition behaviors in an elementary school. *School Psychology Quarterly*, 12, 344–363.
- Dufrene, B. A., Reisener, C. D., Olmi, D. J., Zoder-Martell, K., McNutt, M. R., & Horn, D. R. (2010). Peer tutoring for reading fluency as a feasible and effective alternative in response to intervention systems. *Journal of Behavioral Education*, 19, 239–256.³
- 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
- Gentile, J. R., Roden, A. H., & Klein, R. D. (1972). An analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis*, 5, 193–198.
- Goldstein, H. (1995). *Multilevel statistical models*. London, U.K.: Edward Arnold.
- Goldstein, H., Healy, M., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, 13, 1643–1655.
- Greenwood, K. M., & Matyas, T. A. (1990). Problems with the application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment*, 12, 355–370.
- Heitjan, D. F., & Sharma, D. (1997). Modeling repeated-series longitudinal data. *Statistics in Medicine*, 16, 347–355.
- Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods*, 3, 104–116. doi:10.1037/1082-989X.3.1.104
- Huitema, B. E., McKean, J. W., & McKnight, S. D. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement*, 59, 767–786.
- Ingersoll, B., & Lalonde, K. (2010). The impact of object and gesture limitation training on language used in children with autism spectrum disorder. *Journal of Speech Language, and Hearing Research*, 53, 1040–1051.¹
- Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42, 805–820.
- Kazdin, A. E., & Kopel, S. A. (1975). On resolving ambiguities of the multiple-baseline design: Problems and recommendations. *Behavior Therapy*, 6, 601–608.
- Koegel, L. K., Singh, A. K., & Koegel, R. L. (2010). Improving motivation for academics in children with autism. *Journal of Autism Developmental Disorders*, 40, 1057–1066. doi:10.1007/s10803-010-0962-6⁴
- Kratochwill, T., Alden, K., Demuth, D., Dawson, D., Panicucci, C., Arntson, P., & Levin, J. (1974). A further consideration in the application of an analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis*, 7, 629–633.
- Lütkepohl, H. (1996). *Handbook of matrices*. Chichester, U.K.: Wiley.
- 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. doi:10.1037/1082-989X.5.1.87
- Murphy, D. L., & Pituch, K. A. (2009). The performance of multilevel growth curve models under an autoregressive moving average process. *The Journal of Experimental Education*, 77, 255–282.
- Neter, J., Wasserman, W., & Kutner, M. H. (1990). *Applied linear statistical models: Regression, analysis of variance, and experimental designs* (3rd ed.). Homewood, IL: Irwin.
- Neuman, S. B., & McCormick, S. (Eds.). (1995). *Single-subject experimental research: Applications for literacy*. Newark, DE: International Reading Association.
- Nugent, W. (1996). Integrating single-case and group-comparison designs for evaluation research. *The Journal of Applied Behavioral Science*, 32, 209–226.
- Oddo, M., Barnett, D. W., Hawkins, R. O., & Musti-Rao, S. (2010). Reciprocal peer tutoring and repeated reading: Increasing practicality using student groups. *Psychology in the Schools*, 47, 842–858.²
- Scheffé, H. (1959). *The analysis of variance*. New York, NY: Wiley.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasiexperimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- 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., Rindskopf, D. M., & Hedges, L. V. (2008). The state of science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, 3, 188–196.
- Sharpley, C. (1981). Time series analysis of counseling research. *Measurement and Evaluation in Guidance*, 14, 149–157.
- Shine, L. C., & Bower, S. M. (1971). A one-way analysis of variance for single-subject designs. *Educational and Psychological Measurement*, 31, 105–113.
- Skinner, C. H. (2004). Single-subject designs: Procedures that allow school psychologists to contribute to the intervention evaluation and validation process. *Journal of Applied School Psychology*, 20, 1–10.

- Solanas, A., Manolov, R., & Sierra, V. (2010). Lag-one autocorrelation in short series: Estimation and hypothesis testing. *Psicologica*, 31, 357–381.
- Swanson, H. L., & Sachse-Lee, C. (2000). A meta-analysis of single-subject-design intervention research for students with LD. *Journal of Learning Disabilities*, 33, 114–136.
- Toothaker, L. E., Banz, M., Noble, C., Camp, J., & Davis, D. (1983). $N = 1$ designs: The failure of ANOVA-based tests. *Journal of Educational Statistics*, 4, 289–309.
- 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. doi:[10.3758/BF03195492](https://doi.org/10.3758/BF03195492)
- Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *Behavior Analyst Today*, 8, 52–75.
- Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-based Communication Assessment and Intervention*, 2, 142–151.
- Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician*, 39, 95–101.
- Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computation*, 22, 1079–1106.
- Yang, M., & Goldstein, H. (1996). Multilevel models for longitudinal data. In U. Engel & J. Reinecke (Eds.), *Analysis of change: Advanced techniques in panel data analysis* (pp. 191–220). New York, NY: Walter de Gruyter.