Behavior Research Methods
2009, 41 (2), 372-384
doi:10.3758/BRM.41.2.372

Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches

John M. Ferron
University of South Florida, Tampa, Florida

Bethany A. Bell
University of South Carolina, Columbia, South Carolina

Melinda R. Hess
University of South Florida, Tampa, Florida

Gianna Rendina-Gobioff
Professional Testing Inc., Orlando, Florida

and

Susan T. Hibbard
Florida Gulf Coast University, Fort Myers, Florida

Multiple-baseline studies are prevalent in behavioral research, but questions remain about how to best analyze the resulting data. Monte Carlo methods were used to examine the utility of multilevel models for multiple-baseline data under conditions that varied in the number of participants, number of repeated observations per participant, variance in baseline levels, variance in treatment effects, and amount of autocorrelation in the Level 1 errors. Interval estimates of the average treatment effect were examined for two specifications of the Level 1 error structure ($\sigma^2\mathbf{I}$ and first-order autoregressive) and for five different methods of estimating the degrees of freedom (containment, residual, between–within, Satterthwaite, and Kenward–Roger). When the Satterthwaite or Kenward–Roger method was used and an autoregressive Level 1 error structure was specified, the interval estimates of the average treatment effect were relatively accurate. Conversely, the interval estimates of the treatment effect variance were inaccurate, and the corresponding point estimates were biased.

The basic interrupted time-series design includes two phases: a baseline phase consisting of a series of observations preceding the introduction of a treatment, and a treatment phase consisting of a series of observations following the introduction of a treatment (see Figure 1). Although inferences about treatment effects are made from studies that utilize basic interrupted time-series designs, 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 (e.g., an event that happened to occur around the time of the intervention; Shadish, Cook, & Campbell, 2002). In an effort to reduce the plausibility of alternative explanations for shifts in time-series data, researchers often turn to more complex interrupted time-series designs, such as the reversal design and the multiple-baseline design (see Figure 1). The reversal design increases the number of phases by withdrawing and later reintroducing a treatment, whereas the multiple-baseline design includes interrupted time-series data from multiple participants (or behaviors, or settings) where an intervention is staggered to occur at different times within the different series.

These interrupted time-series designs—also called *single-participant*, *single-case*, or *single-subject* designs—are valued for several reasons. First, it is important for one to have designs that allow researchers to estimate individual treatment effects. By accumulating individual treatment effects, we can gain a better understanding of person-specific effects and their variation than we can obtain through traditional group designs, which focus on average effects (Barlow & Hersen, 1984; Morgan & Morgan, 2001). Second, researchers need designs that allow them to study treatment effects in low-incidence or highly fragmented populations (Dukes, 1965; Van den Noortgate & Onghena, 2003a). Third, it is valuable for one to have research designs with features that are closely aligned

J. M. Ferron, ferron@tempest.coedu.usf.edu

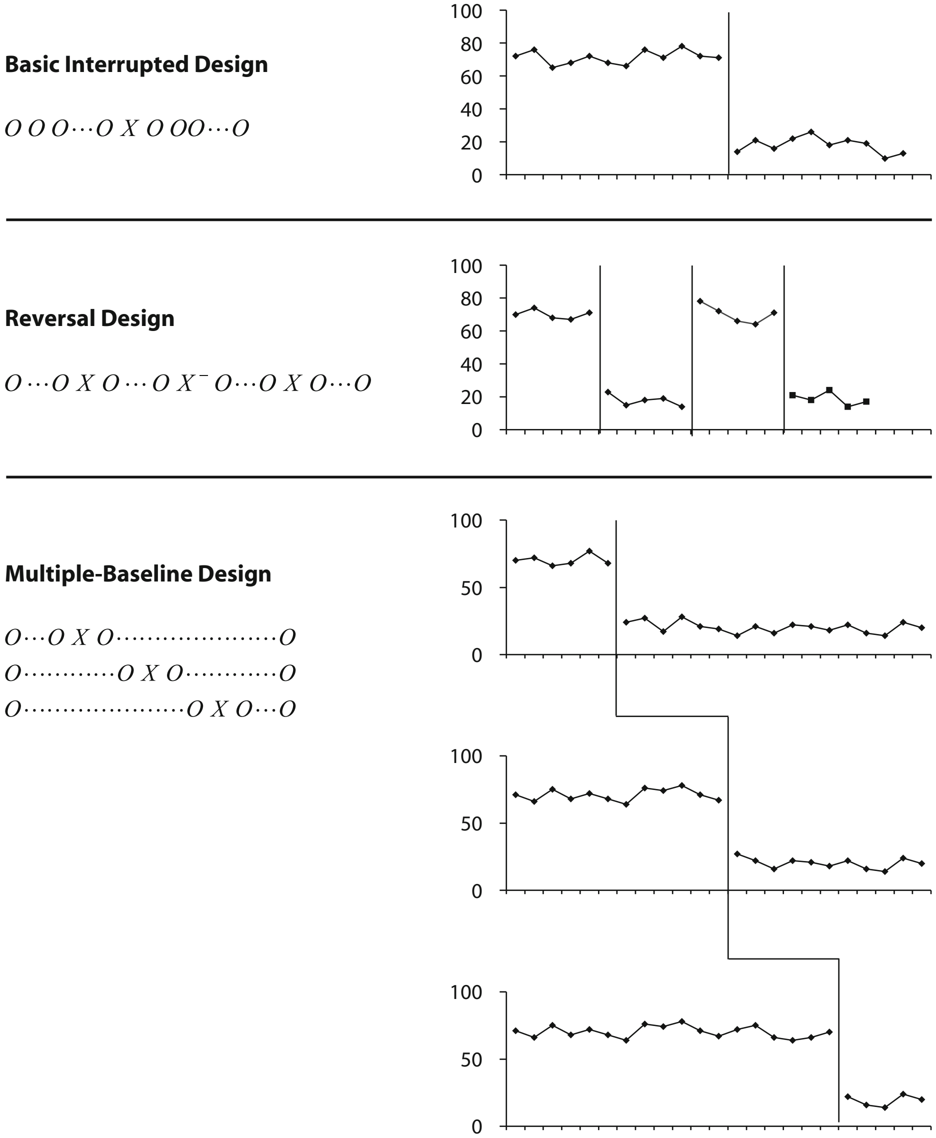

Figure 1. Design diagrams and graphical displays of the basic interrupted time-series design, the reversal design, and the multiple-baseline design. The *O*s represent observations; the *X*s represent the implementation of an intervention; the X^- represents removal of the intervention.

with practice (Kratochwill & Piersel, 1983; Morgan & Morgan, 2001). These designs allow engagement of clinicians and practitioners in research, thereby lessening the gap between research and practice.

Among clinicians and practitioners, the multiple-baseline design is often preferable to the reversal design because it does not require withdrawal of the treatment (Barlow & Hersen, 1984; Ferron & Scott, 2005). Consequently, the multiple-baseline design has become widely used. A search in PsycINFO for "multiple baseline" that was restricted to the year 2007 produced 132 entries; 82 of these were journal articles from a variety of different fields. For example, a multiple-baseline design across settings was used to examine the effects of an intervention on the social interaction skills of a child with Asperger syndrome (Bock, 2007). A multiple-baseline design across participants was used to examine the effect of a treatment on clinical perfectionism in participants with Axis I disorders (Glover, Brown, Fairburn, & Shafran, 2007). Furthermore, a multiple-baseline design across behaviors was used to examine the effects of treatment on verbal productivity in a participant with anomic aphasia (Wambaugh & Ferguson, 2007). Other multiple-baseline applications included a study of the effects of a treatment for depression in a primary care setting (Naylor, Antonuccio, Johnson, Spogen, & O'Donohue, 2007), a study of the effects of an instructional intervention on phoneme-segmentation fluency with at-risk kindergarten children (Musti-Rao & Cartledge, 2007), and a study of the effects of training on the aggressive behaviors of individuals with mild mental retardation (Singh et al., 2007).

Although the utility of multiple-baseline designs is well established, there is no consensus on how to analyze and present the resulting data. Among the methods traditionally employed are visual analyses (Parsonson & Baer, 1992), randomization tests (Bulté & Onghena, 2009; Koehler & Levin, 1998; Marascuilo & Busk, 1988), ordinary least squares regression (Huitema & McKean, 1998), first-order autoregressive models (McKnight, McKean, & Huitema, 2000), and more general time-series models (Velicer & Fava, 2003). In recent years, multilevel models (also called hierarchical linear models, or mixed linear models) have also been suggested as a method for combining single-case data within and across studies (Nugent, 1996; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003a, 2003b).

Multilevel Model for Multiple-Baseline Data

As shown in Equations 1–3, in the basic multilevel model for multiple-baseline data across participants, an outcome (y) is modeled for participant j as a linear function of a single predictor, *phase*,

$$y_{ij} = \beta_{0j} + \beta_{1j}\,\text{phase}_{ij} + r_{ij}, \tag{1}$$

where phase is a dichotomous variable indicating whether the observation is from the baseline or treatment phase, β_{0j} is the level of the outcome during baseline for the jth participant, β_{1j} is the treatment effect for the jth participant, and r_{ij} is error leading to within-phase variation. At the second level of the model, the regression coefficients of the first level are allowed to vary randomly across participants:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{2}$$

and

$$\beta_{1j} = \gamma_{10} + u_{1j}, \tag{3}$$

where γ_{00} is the average baseline level, γ_{10} is the average treatment effect, and u_{0j} and u_{1j} are errors that are assumed to be normally distributed. These errors lead to variation in both baseline levels among participants and treatment effects among participants.

Multilevel modeling is an appealing option because (1) it allows one to use data from multiple cases in a single analysis; (2) the models are flexible enough to handle dependent error structures, heterogeneous variances, and moderating effects; (3) software for estimating multilevel models is accessible and familiar to many applied researchers; and (4) interval estimates can be obtained for effects of interest. The size of the average treatment effect can be gauged by the fixed effect for treatment; variation in the treatment effect can be estimated by the variance component for the treatment effect; potential moderators of the treatment effect can be examined through the inclusion of cross-level interaction effects; and individual treatment effects can be provided from the empirical Bayes estimates.

Although multilevel models allow the estimation of parameters that address questions of interest among single-participant researchers, some concerns can also be raised. First, the multilevel model assumes that the time series are independent of each other. In some multiple-baseline applications, however, one would expect the units to be interdependent (Bulté & Onghena, 2009; Marascuilo & Busk, 1988). A multiple-baseline design across behaviors will often produce series that are interdependent, because the treatment of one behavior may affect other behaviors in the individual and because an unmeasured variable that contributes to the error may have an impact on multiple behaviors. A multiple-baseline design across participants, however, may involve either independent or interdependent participants (Marascuilo & Busk, 1988). For example, four students working together in the same class would tend to have interdependent series, whereas four students from different classes may have independent series. It would appear that multilevel modeling would be most appropriate for applications in which the participants are independent.

The second concern focuses on the appropriateness of the sample size. The restricted maximum likelihood methods typically used to estimate multilevel models were developed under large-sample theory, and recommendations for use often suggest sample sizes of at least 30 upper-level units (Hox, 1998). Although researchers may prefer to have large numbers of participants, practical constraints often lead to multiple-baseline studies with 4 to 8 individuals. Consequently, the amount of data available from a single study may not be adequate for a multilevel analysis.

On the basis of previous research (see, e.g., Maas & Hox, 2004; Mok, 1995; Raudenbush & Bryk, 2002), one may anticipate that interval estimates of the average treatment effect would have a better chance of performing well under small-sample-size conditions than would interval estimates of the variance in the treatment effect. More specifically, fixed-effect estimates are unbiased when sample size is small, but variance estimates are not (Raudenbush & Bryk, 2002). Maas and Hox (2004) reported a 25% upward bias in the Level 2 variance components when there were only 10 Level 2 units each of size five. Mok (1995), who studied a variety of Level 1 samples sizes, reported relative bias as high as 34% with 5 Level 2 units, 18% with 10 Level 2 units, and 10% with 20 Level 2 units.

Although the fixed-effect estimates are unbiased, questions can be raised about confidence intervals and significance tests, because they depend not only on the effect estimates, but also on the estimated standard errors and degrees of freedom. Several alternative estimates for the degrees of freedom have been proposed (Fai & Cornelius, 1996; Kenward & Roger, 1997). Although the differences among these methods have a trivial impact when the Level 2 sample size is large, the differences would appear to be material for small-sample contexts, such as multiple-baseline designs.

Methods for Estimating Degrees of Freedom

Consider a multiple-baseline design across 6 participants with 20 observations per participant, in which the researcher uses the model defined in Equations 1–3. The degrees of freedom for an inference about the average treatment effect (γ_{10} in Equation 3) could be estimated using a variety of approaches. The simplest would be to take the number of Level 1 units summed across participants and subtract the number of fixed effects, which for this example yields 118 degrees of freedom (i.e., $120 - 2$). This simple method of estimating degrees of freedom, which is referred to as the *residual* method, and which was the default in the earliest versions of PROC MIXED in SAS, would seem inappropriate for multiple-baseline data because of the anticipated clustering of observations within participants.

An alternative would be to estimate the degrees of freedom as the number of Level 2 units minus the number of fixed effects within the Level 2 equation containing the fixed effect of interest. For this example, we would obtain 5 degrees of freedom (i.e., $6 - 1$) for the inference about the average treatment effect. This is the approach used by the HLM software, as well as the approach used in the *containment* method in SAS as long as the effect is listed on the random statement in PROC MIXED. Another alternative, referred to in SAS as the *between–within* method, is to partition the residual degrees of freedom into between-participants and within-participants degrees of freedom. For our example, this approach would yield 5 between-participants and 113 within-participants degrees of freedom. The treatment effect would be assigned the within-participants degrees of freedom because the treatment varies within participants.

Other approaches to estimating the degrees of freedom rely on estimates of the variance–covariance matrix of the vector of responses. The *Satterthwaite* method used in SAS is a generalization of the procedure described by Fai and Cornelius (1996), which builds on the work of Satterthwaite (1941). The *Kenward–Roger* method is an extension of the Satterthwaite method, and was developed by Kenward and Roger (1997) to adjust for small-sample bias in the variance estimation. These methods, as well as those mentioned previously, are defined more formally in the Appendix. More detailed descriptions are also available elsewhere (SAS Institute Inc., 2004; Schaalje, McBride, & Fellingham, 2001).

Considering the mathematical differences in the approaches used to estimate degrees of freedom, one might anticipate that the confidence interval coverage for the average treatment effect would vary on the basis of the method employed. The containment method would be expected to lead to higher coverage than would the residual or between–within methods, and the Kenward–Roger and Satterthwaite methods would be expected to produce coverage in between the containment and residual methods, with the consequences of using the different methods diminishing as the Level 2 sample size increased. However, it is unknown under what circumstances acceptable interval estimates can be made from multiple-baseline data.

Several researchers have examined the performance of the Kenward–Roger method of estimating degrees of freedom in the context of more traditional repeated measures designs (Fouladi & Shieh, 2004; Gomez, Schaalje, & Fellingham, 2005; Kenward & Roger, 1997; Kowalchuk, Keselman, Algina, & Wolfinger, 2004; Schaalje et al., 2001). Although each of these studies has indicated promising results, with Type I error rates estimated to be close to the nominal level across a variety of design and data conditions, variability in performance was also noted. For example, using a three-group design with 3 participants per group, each measured at three points in time, Gomez et al. (2005) found Type I error control to vary on the basis of the covariance structure. More specifically, when data were generated and analyzed assuming compound symmetry, the estimated Type I error rate for the main effect of treatment was .0525 ($\alpha = .05$). Conversely, when the data were generated and analyzed on the basis of a first-order autoregressive with random effects model, the Type I error rate for the treatment effect was estimated to be .1165 ($\alpha = .05$).

Kowalchuk et al. (2004) also examined a three-group design, but they considered 30 participants who were unequally split into groups of size 6, 10, and 14, and who were measured at four points in time. Data were generated using a random coefficient, a heterogeneous first-order autoregressive, or an unstructured covariance structure. Using a nominal α of .05, they found that when the covariance structure was selected using Akaike's information criterion, the Type I error rate estimates varied from .028 to .069 for the time main effect, and from .045 to .072 for the interaction effect. When the covariance structure was selected using Schwarz's Bayesian criterion, the estimates of the Type I error rates ranged from .020 to .088 for the time main effect and from .035 to .081 for the interaction effect. In addition to the reported variation in performance across design and data conditions, the inability to mathematically derive performance under small-sample-size conditions adds to the complexity of generalizing the performance of the Kenward–Roger method to multiple-baseline designs.

Purpose

The purpose of the present study was to examine the quality of treatment effect inferences made from multilevel models of multiple-baseline data. Specifically, the interval estimate of the average treatment effect was examined for each of five methods of approximating the degrees of freedom (containment, residual, between–within, Satterthwaite, and Kenward–Roger), and for each of two methods of specifying the Level 1 error structure ($\sigma^2\mathbf{I}$ or first-order autoregressive). In each case, the quality of the inference was considered for conditions varying in the number of participants, series length, level of autocorrelation, variance among participants in initial level, and variance among participants in treatment effect. The point and interval estimates of the variance in the treatment effect were also examined.

METHOD

Monte Carlo simulation methods were used to examine approaches for making multilevel modeling inferences from multiple-baseline data. The number of simulated participants (Level 2 sample size) was 4, 6, or 8. The number of simulated observations in the time series (series length or Level 1 sample size) for each participant was 10, 20, or 30. By crossing the number of participants with the series length, nine conditions were obtained that covered the range of sample sizes and series lengths typically reported in multiple-baseline studies.

Data were generated on the basis of the two-level model defined in Equations 1–3, with the fixed effects (γ_{00} and γ_{10}) set to 1.0. The within-participants model (Equation 1) is based on an immediate shift in level and is consistent with the multilevel modeling application presented by Van den Noortgate and Onghena (2003a). Furthermore, because it represents the most basic interrupted time-series model (e.g., there are no trends, changes in trends, or seasonal effects), it appeared to be the appropriate model for initial study into the multilevel modeling of multiple-baseline data. Errors for the within-participants model (r_{ij}) were generated using the ARMASIM function in SAS (Version 9.1; SAS Institute Inc., 2005), with a variance (σ^2) of 1.0 and an autocorrelation (ρ) of 0, .1, .2, .3, or .4, which appears ample to cover the range of autocorrelation typically found in behavioral data (Busk & Marascuilo, 1988; Huitema, 1985; Matyas & Greenwood, 1997).

The between-participants model (Equations 2 and 3) allowed baseline level and treatment effects to vary randomly across participants. Level 2 errors were generated from a normal distribution using the RANNOR random number generator in SAS. The variance of u_{0j} (τ_{00}) was equal to 0.1 or 0.3; the variance of u_{1j} (τ_{11}) was equal to 0.1 or 0.3, and the covariance between u_{0j} and u_{1j} was 0. These Level 2 variances were chosen so that the majority of the variance would be in the Level 1 errors (recall $\sigma^2 = 1.0$). Substantial Level 1 variation makes treatment effects more difficult to discern visually and thus motivates statistical analyses. A larger variance component at Level 1 was also consistent with the multilevel modeling application presented by Van den Noortgate and Onghena (2003a) and with several reanalyses of recently published multiple-baseline studies that we conducted using multilevel models. For example, when previously published multiple-baseline data were reanalyzed using multilevel modeling, in some cases it was found that σ^2 was greater than τ_{00}, which exceeded τ_{11} (Figure 2 in Mahar et al., 2006, and Figure 1 in O'Callaghan, Allen, Powell, & Salama, 2006). In some cases, σ^2 was greater than τ_{11}, which exceeded τ_{00} (Figure 3 in Tiger, Hanley, & Hernandez, 2006, and Figure 4 in Tsao & Odom, 2006).

Next, crossing the two variance levels of u_{0j} with the two variance levels of u_{1j} and the five levels of autocorrelation, we examined a total of 20 variance conditions for each of the nine combinations of sample size with series length. For each of these 180 data conditions ($20 * 9$), 5,000 data sets were simulated using SAS IML. The use of 5,000 replications leads to an adequate level of precision when estimating the coverage (e.g., when the coverage is .95, the standard error is .0031).

After each data set was generated, it was analyzed using multilevel modeling with REML estimation via PROC MIXED in SAS. Interval estimates of the fixed effects were made using each of the five methods of estimating degrees of freedom: containment method, residual method, between–within method, Satterthwaite method, and Kenward–Roger method (Kenward & Roger, 1997; Schaalje et al., 2001). For variance components, the confidence intervals were based on the Satterthwaite method.

In the initial simulations, the treatment effect was modeled as a shift in level between the baseline and treatment phases; the Level 1 errors were modeled as $\sigma^2\mathbf{I}$, and the intercept and treatment effects were allowed to vary randomly across participants; see Figure 2 for the PROC MIXED specification when the Kenward–Roger method was used to estimate the degrees of freedom. The Level 1 error specification of $\sigma^2\mathbf{I}$ was chosen because it is a relatively simple structure that is commonly used in multilevel models for more traditional growth curve applications, and because simulation work has shown that in these more traditional applications, tests of fixed effects function relatively well with this specification, even when the errors are generated on the basis of an autoregressive model (Ferron, Dailey, & Yi, 2002).

Simulations for all 180 data conditions were then rerun with the specification of the analysis model changed to estimate a first-order autoregressive model for the Level 1 errors. Again, results were obtained for each of the degrees of freedom methods. The PROC MIXED specification for this second set of simulations is shown in Figure 2, and again the syntax shows the particular case in which the Kenward–Roger method was used to estimate the degrees of freedom.

Convergence rates were high (100% convergence for 94% of the conditions, and 99.98% convergence for the other 6% of the conditions). Several checks were used to verify the accuracy of the simulation program. For a small number of replications, the vectors produced at each stage of data generation were examined, as were the output data sets generated by calls to PROC MIXED and the summary data set that collected results. In addition, the fixed-effect parameter values used in data generation were compared with the values obtained from analyses of the generated data. It was found that the fixed-effect estimates were unbiased, which was theoretically expected and provided additional information that the simulation was functioning appropriately. The code for conducting the simulations is available from the first author.

```
proc mixed covtest cl;

model y = phase / s cl alpha = .05 ddfm = kenwardroger;

random int phase / sub = idlevel2;


proc mixed covtest cl;

model y = phase / s cl alpha = .05 ddfm = kenwardroger;

random int phase / sub = idlevel2;

repeated / type = AR(1) sub = idlevel2;
```

Figure 2. Example PROC MIXED code showing multilevel model specification when the Level 1 error structure was assumed to be σ^2I and when the Level 1 error structure was assumed to be first-order autoregressive. In both cases, the Kenward–Roger method was used to estimate the degrees of freedom.

RESULTS

To estimate confidence interval coverage (CI_{95}) for each fixed effect, the results from the analyses were aggregated across the 5,000 replications for each of the 180 data conditions for each of the five degrees of freedom methods and for each of the Level 1 error specifications. Doing this led to 180 × 5 × 2, or 1,800, confidence interval coverage estimates for each fixed effect. Similarly, analyses were aggregated across replications to estimate the widths of the confidence intervals for each fixed effect. However, to conserve space, only the results for the average treatment effect (γ_{10}) are provided. This is the effect of primary interest in multiple-baseline studies. After examining the fixed effects, attention was turned to the variance components. Both the relative bias in the point estimates and confidence interval coverage were examined.

Interval Coverage for the Average Treatment Effect

Boxplots showing the distribution of coverage estimates for each method of estimating the degrees of freedom and each method of modeling the Level 1 error structure are presented in Figure 3. The residual and between–within methods undercovered, regardless of how the Level 1 error structure was specified, with both of these degrees of freedom methods having an average coverage of .917. The containment method tended to overcover, with an average estimate of .972 across the two Level 1 error structures. When the Level 1 error structure was modeled as $\sigma^2\mathbf{I}$, the average coverage estimates for the Kenward–Roger and Satterthwaite methods were .942 and .940, respectively. When a first-order autoregressive model was specified for the Level 1 errors, both the Kenward–Roger and Satterthwaite methods provided average estimates that were very close to the nominal level of .95 (M = .952 and M = .949, respectively). In addition, for these estimation methods, all 180 estimates were reasonably close to the nominal level (min = .935, max = .965, for Kenward–Roger; min = .935, max = .960, for Satterthwaite).

To explore the variation between estimates, a series of graphs was constructed, each showing the coverage estimate as a function of the degrees of freedom method, the method used to model the Level 1 errors, and one of the data factors (i.e., autocorrelation, number of participants, series length, intercept variance, or treatment effect variance). These graphs show the vast majority of the variation in the coverage estimates. The η^2 was .96 when coverage was modeled with the main effects, with the two-way interactions involving either the degrees of freedom method or the Level 1 error model, and with the three-way interactions involving both the degrees of freedom method and the Level 1 error model.

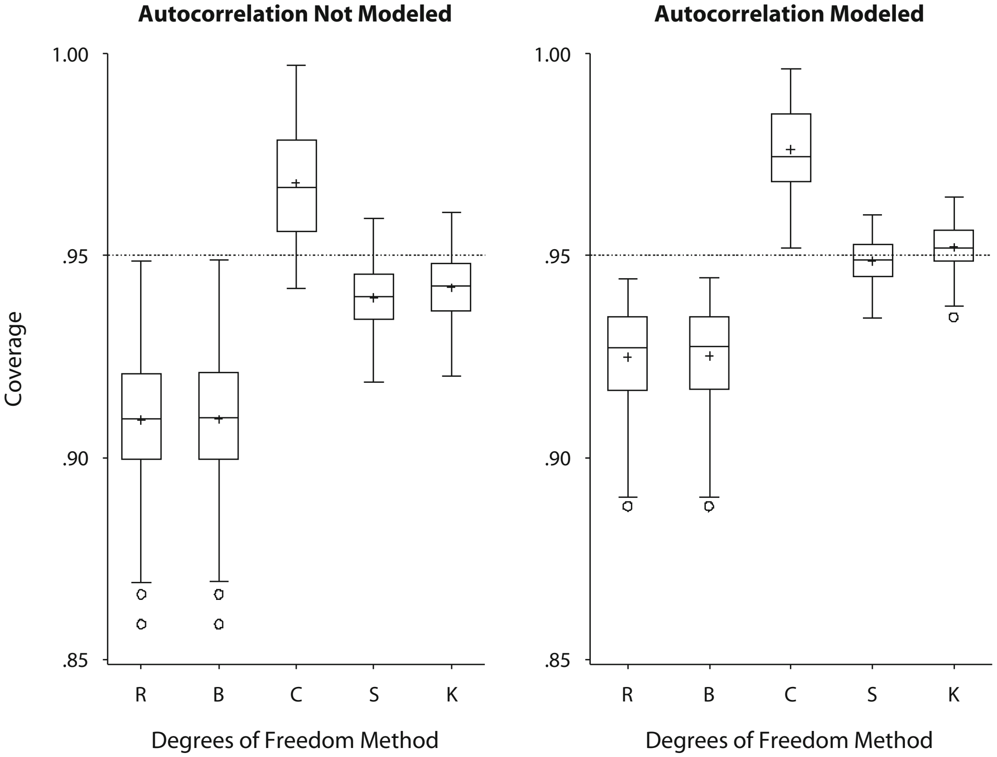

Figure 3. Boxplots showing the distribution of coverage estimates for the 95% confidence interval of the treatment effect, γ_{10}, for each modeled Level 1 error structure and each degrees of freedom method. R, residual method; B, between–within method; C, containment method; S, Satterthwaite method; K, Kenward–Roger method. "Autocorrelation not modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was σ^2I, whereas "autocorrelation modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was first-order autoregressive.

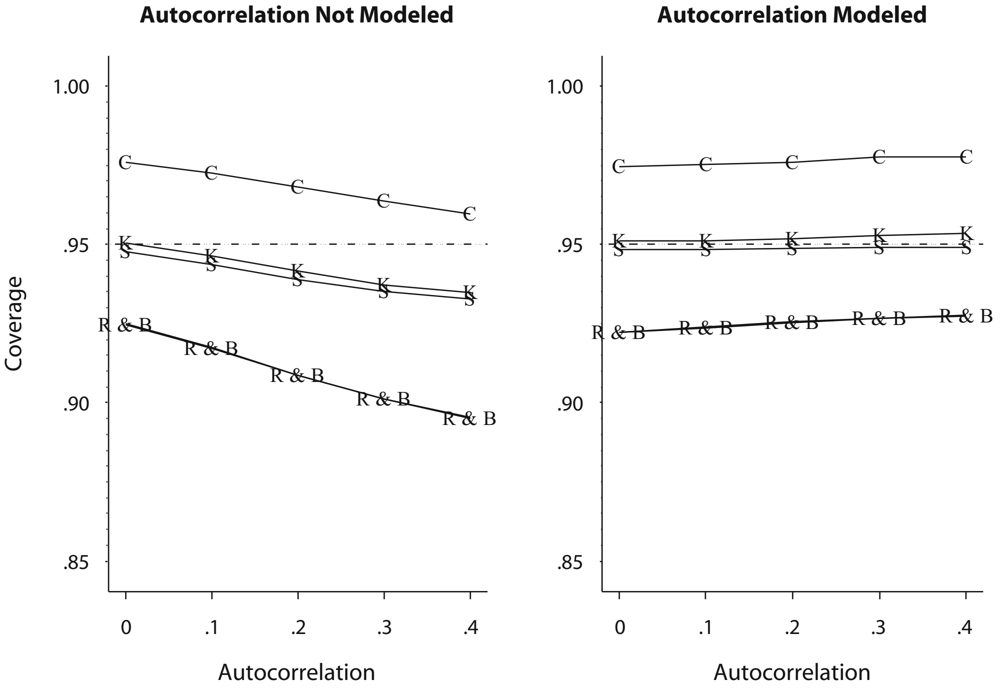

Figure 4. Average coverage for the treatment effect, γ_{10}, as a function of the autocorrelation for each modeled Level 1 error structure and each degrees of freedom method. R, residual method; B, between–within method; C, containment method; S, Satterthwaite method; K, Kenward–Roger method. "Autocorrelation not modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was $\sigma^2\mathbf{I}$, whereas "autocorrelation modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was first-order autoregressive.

The coverage rates as a function of autocorrelation are shown in Figure 4. The interaction between autocorrelation and whether or not autocorrelation was modeled can readily be seen. When the autocorrelation in the Level 1 errors was not modeled (i.e., $\sigma^2\mathbf{I}$ was specified), the average coverage estimates decreased for all degrees of freedom methods as the autocorrelation in the generated errors increased, but when the Level 1 error structure was modeled as first-order autoregressive, these drops in coverage did not occur. Of particular note, when autocorrelation was modeled, the Kenward–Roger and Satterthwaite methods maintained coverage very close to the nominal level, regardless of the autocorrelation. When the autocorrelation was not modeled, both the Kenward–Roger and Satterthwaite methods had coverage estimates that dropped further below the desired level as the autocorrelation in the generated errors increased.

The coverage for each method is shown as a function of the number of participants in Figure 5. As was anticipated, there was an interaction between the number of participants and degrees of freedom methods, with the difference in coverage rates among the degrees of freedom methods becoming smaller as the number of participants increased. The average coverage rates for both the Kenward–Roger and Satterthwaite methods were close to .95 for all Level 2 sample-size conditions when the autocorrelation was modeled. For the containment method, the overcoverage became less pronounced as the number of participants increased, and for both the residual and between–within methods, the undercoverage became less pronounced as the number of participants increased. Note, however, that with 8 participants (the largest number examined), there were still notable differences among the degrees of freedom methods. When autocorrelation was modeled and there were 8 participants, the average coverage estimates were .954 for the Kenward–Roger method, .950 for the Satterthwaite method, .966 for the containment method, .931 for the between–within method, and .930 for the residual method.

The coverage for each method as a function of series length is shown in Figure 6. When autocorrelation was modeled, all degrees of freedom methods showed small decreases in coverage as series length increased. For the Kenward–Roger method, the average coverage went from .955 when the series length was 10, to .950 when the series length was 30. For the Satterthwaite method, the average coverage went from .949 when the series length was 10, to .948 when the series length was 30. The most pronounced effects for series length were seen for the residual and between–within methods.

Variance in the intercepts (τ_{00}) had almost no impact on treatment effect coverage rates (η^2 of .0028 for the combination of the main effect and interactions); thus, this graph is not provided. However, coverage rates were influenced by variance in the treatment effects (τ_{11}). As shown in Figure 7, the coverage for each of the degrees of freedom methods decreased as the variance in the treatment effects increased. For conditions in which the autocorrelation was

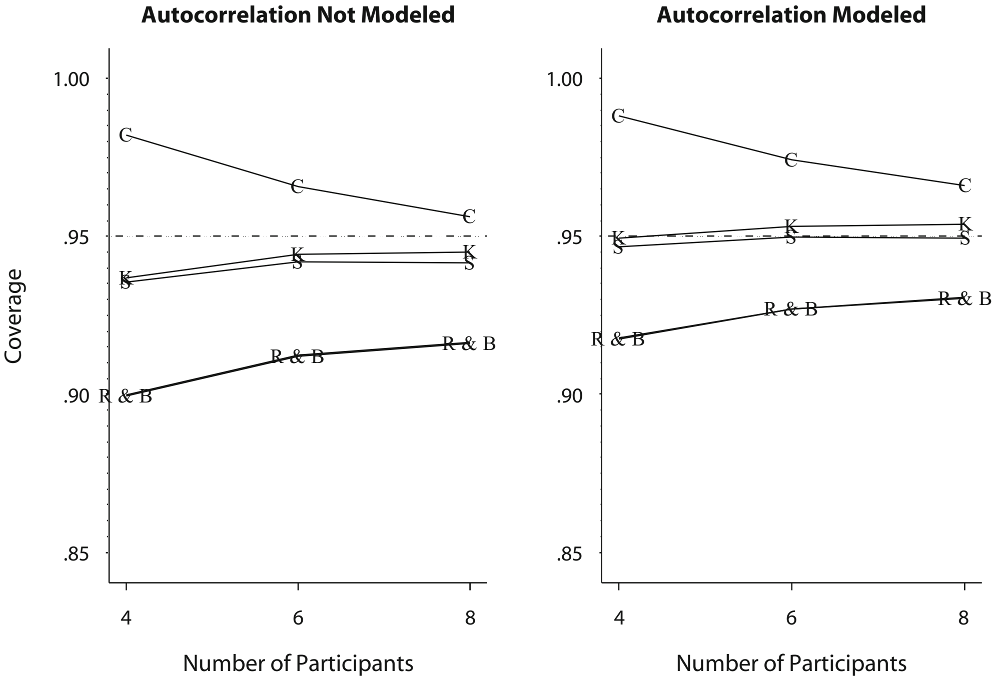

Figure 5. Average coverage for the treatment effect, γ_{10}, as a function of the number of participants for each modeled Level 1 error structure and each degrees of freedom method. R, residual method; B, between–within method; C, containment method; S, Satterthwaite method; K, Kenward–Roger method. "Autocorrelation not modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was σ^2I, whereas "autocorrelation modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was first-order autoregressive.

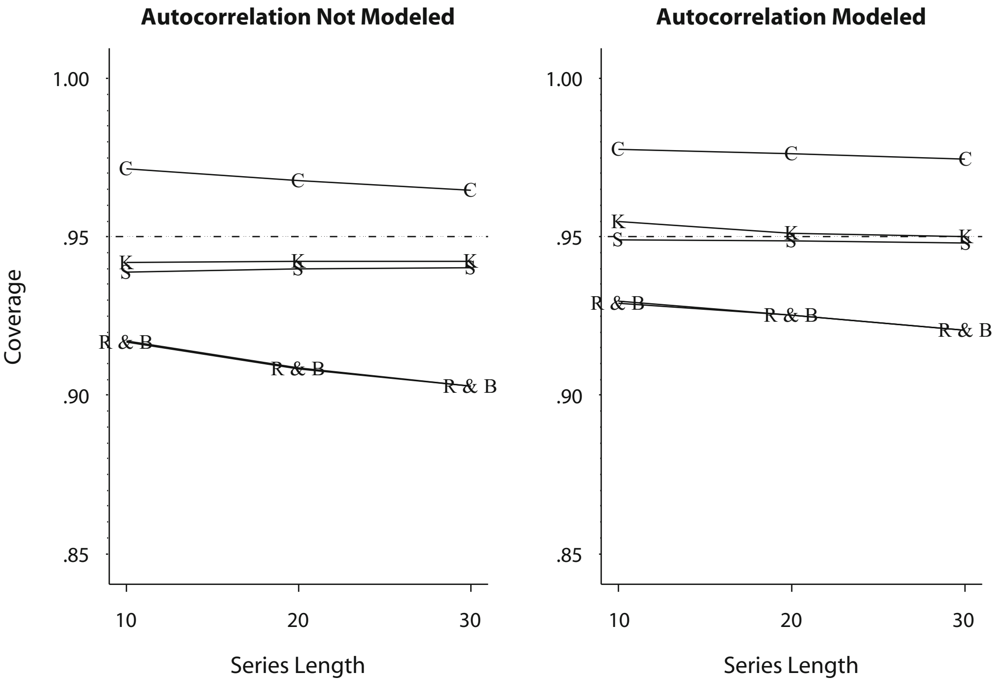

Figure 6. Average coverage for the treatment effect, γ_{10}, as a function of the series length for each modeled Level 1 error structure and each degrees of freedom method. R, residual method; B, between–within method; C, containment method; S, Satterthwaite method; K, Kenward–Roger method. "Autocorrelation not modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was σ^2I, whereas "autocorrelation modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was first-order autoregressive.

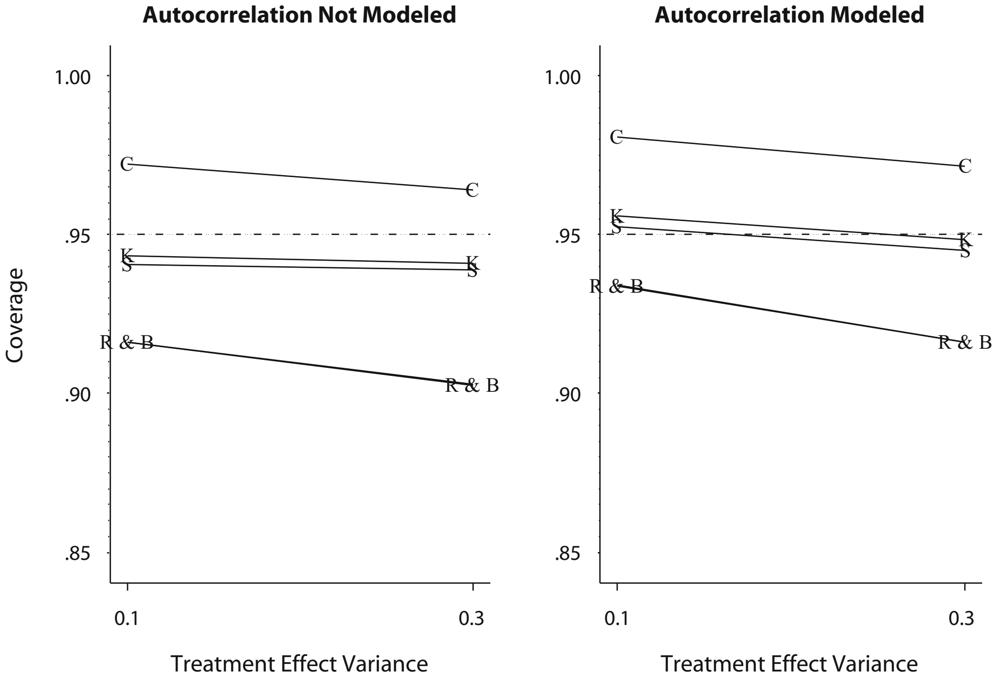

Figure 7. Average coverage for the treatment effect, γ_{10}, as a function of the treatment effect variance for each modeled Level 1 error structure and each degrees of freedom method. R, residual method; B, between–within method; C, containment method; S, Satterthwaite method; K, Kenward–Roger method. "Autocorrelation not modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was σ^2I, whereas "autocorrelation modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was first-order autoregressive.

modeled, as the variance in the treatment effects increased from 0.1 to 0.3, the average coverage for the Kenward–Roger method decreased from .956 to .948, and the average coverage for the Satterthwaite method decreased from .952 to .945.

Interval Width for the Average Treatment Effect

After examining the coverage rates, attention was turned to the widths of the confidence intervals. As was expected, the interval widths were smallest for the residual and between–within methods and largest for the containment method. Also as was expected, the confidence interval widths decreased with more participants, more observations per participant, and smaller variance components. The most notable effect was the number of participants (η^2 = .42), followed by the method for estimating the degrees of freedom (η^2 = .18), and the series length (η^2 = .17). Figure 8 shows the effect of the number of participants on interval width for each degrees of freedom method and each Level 1 error specification.

When the Kenward–Roger method was used to estimate degrees of freedom and the autocorrelation was modeled, the results indicated that as the sample size increased from 4 to 6 to 8, the average width decreased from 1.95 to 1.43 to 1.21, respectively. To get a better feel for widths of this size, it is helpful for one to recall that the Level 1 variance was set to 1.0 in the simulations. It is also helpful for one to consider the magnitude of anticipated treatment effects. In a review that synthesized the results of 150 single-participant studies of school-based interventions, the average mean shift from baseline to treatment was 4.7 times the baseline standard deviation (Gresham et al., 2004). With effects of this magnitude, an interval width of 1.5, for example, could result in an interval estimate around 4.00 to 5.50.

In making decisions about whether to model autocorrelation in the Level 1 errors, one may question how much would be lost in precision if the more complex Level 1 error structure was used, but not needed. Under conditions in which the autocorrelation was 0, results indicated that for the Kenward–Roger method, the average interval width increased from 1.384 to 1.390 when moving from the simple Level 1 error specification of $\sigma^2\mathbf{I}$ to the more complex first-order autoregressive specification. Similarly, under the same conditions, for Satterthwaite method, the loss was also quite small: from 1.371 to 1.374. Interestingly, with both the Kenward–Roger and Satterthwaite methods, when there was a high level of autocorrelation (ρ = .4), using the more complex Level 1 error structure led to both higher coverage rates and smaller interval widths.

One could also question how much precision was lost by being conservative and by using the containment method to estimate the degrees of freedom. When the autocorrelation was modeled, the average width for the containment method was 1.71, whereas the average width for the Kenward–Roger method was 1.53 and the average width for the Satterthwaite method was 1.50. In general, the differences in precision between the containment and other

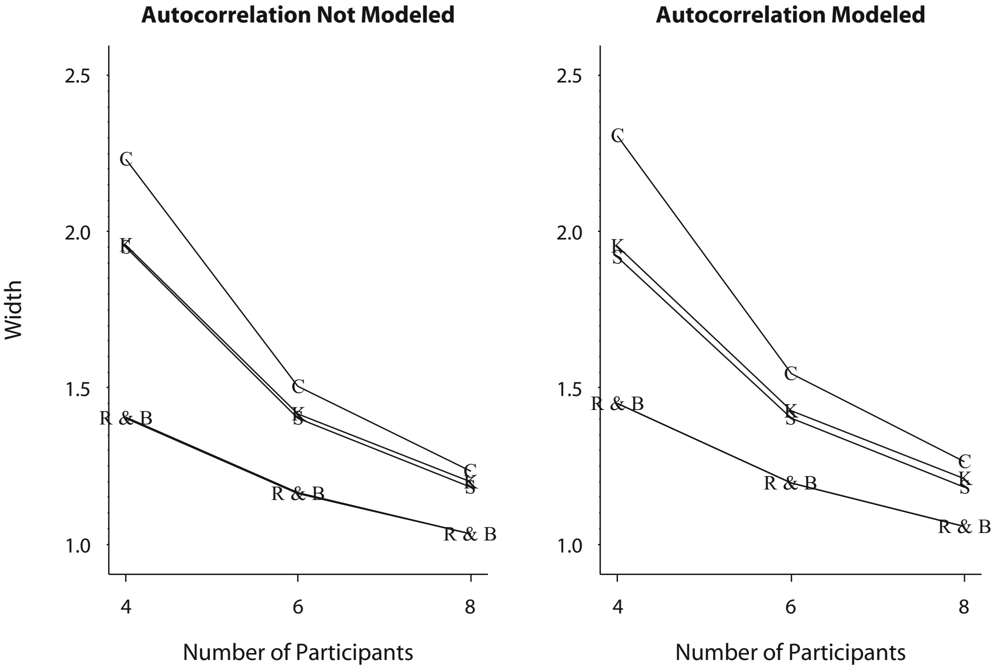

Figure 8. Average interval width for the treatment effect, γ_{10}, as a function of the number of participants for each modeled Level 1 error structure and each degrees of freedom method. R, residual method; B, between–within method; C, containment method; S, Satterthwaite method; K, Kenward–Roger method. "Autocorrelation not modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was σ^2I, whereas "autocorrelation modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was first-order autoregressive.

degrees of freedom methods were much more substantial than the differences in precision between the methods of specifying the Level 1 error structure.

Variance Components

In addition to examining the fixed effects generated from the multilevel models, an examination was made of the variance components. As was expected, these estimates tended to be biased. The distributions of relative bias estimates for the variance components (τ_{00}, τ_{11}, σ^2, and ρ) are provided in Figure 9. Variance in the intercept, or baseline level (τ_{00}), variance in the treatment effect (τ_{11}), and residual variance (σ^2) all tended to be overestimated, whereas the amount of autocorrelation (ρ) tended to be underestimated. The relative bias in τ_{00}, τ_{11}, and ρ tended to decrease as the number of participants increased. Consider, for example, the variance of the treatment effect, which is the component that typically would be of the most interest to single-participant researchers. When autocorrelation was modeled, the average estimate of relative bias decreased from .34 to .25 to .21 as the number of participants increased from 4 to 6 to 8, respectively.

Given the bias in the point estimates, it was not surprising to find that coverage tended to be a problem for the interval estimates of the variance components. Coverage was particularly problematic for the treatment effect variance. When autocorrelation was modeled, the average coverage estimate was .83 with 4 participants, and .85 with 8 participants. In addition to the many cases in which the intervals for treatment effect variance did not cover, there were also cases in which the intervals were so large that they provided no information (e.g., average interval width of 3.3×10^{285} when autocorrelation was modeled).

DISCUSSION

Overall, the degree to which the findings are supportive of using multilevel modeling to make inferences from multiple-baseline data depends on the particular inference examined. Estimates of the variance components tended to be biased, and the confidence intervals for these estimates tended to undercover. Given the magnitude of the bias and undercoverage, the results are not encouraging for researchers wishing to make inferences about the variance in the treatment effect from multiple-baseline data. However, it is important to note that some have suggested the use of multilevel models for the meta-analysis of single-participant studies (Van den Noortgate & Onghena, 2003a, 2003b). For a meta-analytic application, the number of participants would exceed the numbers examined in this study, and interval estimates of variance components might perform better. The bias and undercoverage lessened as the number of participants increased, but with 8 participants, the bias and undercoverage were still substantial. Consequently, researchers interested in studying the variation in treatment effects should consider increasing the number of participants utilized in multiple-baseline studies. Future research could be aimed at determining the

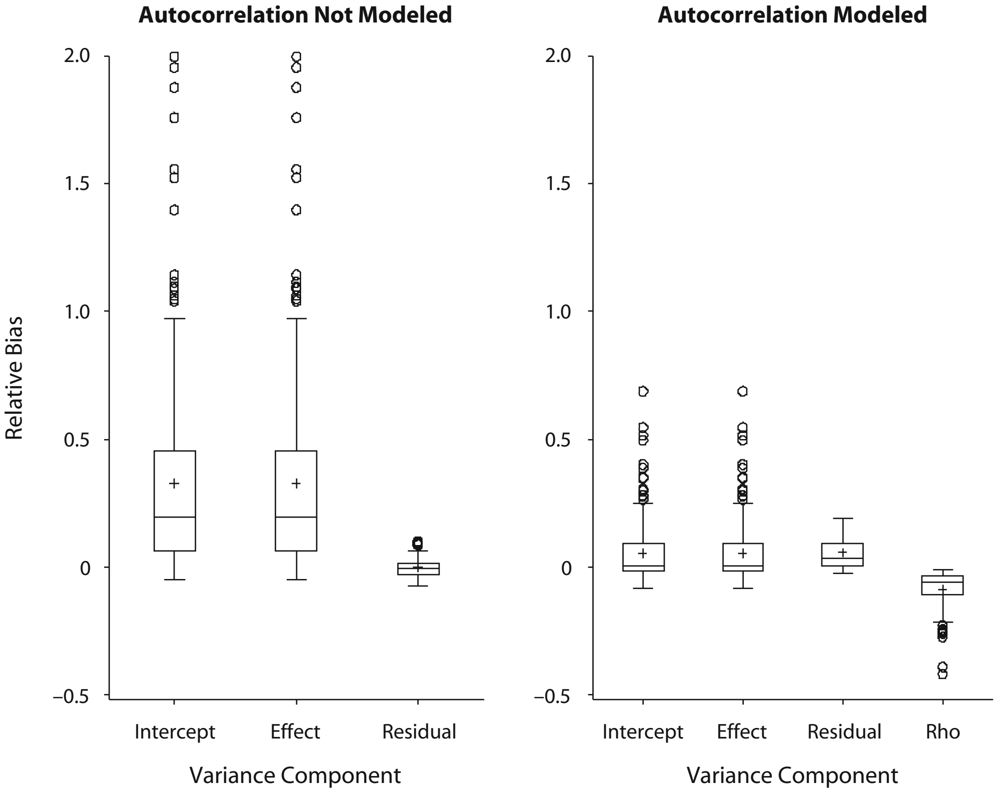

Figure 9. Boxplots showing the distribution of relative bias estimates of the variance components for each modeled Level 1 error structure. "Autocorrelation not modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was σ^2I, whereas "autocorrelation modeled" indicates that the analysis was conducted assuming that the Level 1 error structure was first-order autoregressive. Note that with autocorrelation, relative bias was computed only for conditions in which the parameter value was not 0 (i.e., ρ = .1, .2, or .4).

number of participants needed to make accurate inferences about the variance components. In doing so, alternative methods of estimation, such as the Bayesian approach, should also be considered.

For researchers interested in the average treatment effect, the results of the present study are far more encouraging. Still, researchers should be advised, when possible, to increase the number of participants. With larger Level 2 sample sizes, greater precision could be gained in estimating the average treatment effect. There are contexts, however, in which small sample sizes—like those in the present study—are the only feasible way to conduct the study. Under these conditions, researchers are cautioned to carefully consider how the degrees of freedom are estimated and how the Level 1 errors are modeled. The Satterthwaite and Kenward–Roger methods both provided coverage estimates for the average treatment effect that were close to the nominal .95 level across the conditions studied when autocorrelation was modeled (the lowest estimates were .935). The containment method for estimating degrees of freedom was conservative, providing coverage rates that exceeded the nominal level. Although this method appears relatively safe, it leads to less precise estimates (wider intervals) than does either the Satterthwaite or the Kenward–Roger method. Neither the residual nor the between–within method can be recommended, since each of these methods produced interval estimates that tended to undercover.

The conclusions from the present study should be tempered by recognition of the conditions examined. There are many cases in which a relatively simple multilevel model appears appropriate (see, e.g., Van den Noortgate & Onghena, 2003a, as well as the studies we referenced previously as part of our reanalyses to determine reasonable levels for the variance components). It is for situations like these that the results provide guidance about making treatment effect inferences. It is recognized, however, that some applications may involve more complex treatment effects (e.g., delayed changes in level, transitory effects, effects that change linearly with time in treatment, effects that change nonlinearly with time in treatment, effects that depend on the effects of other participants). Furthermore, some applications may involve more complex error structures (e.g., higher order autoregressive or moving average models, heterogeneous error structures, nonnormally distributed errors at Level 1 or Level 2, multivariate error structures). Finally, some applications could involve variance parameters outside the range of those studied. It is hoped that the present results help to motivate additional methodological research aimed at examining the generalizability of the findings. It is also hoped that future research will address the inferences not addressed in the present study—inferences about moderators of the treatment effect and inferences about treatment effects of individual participants.

AUTHOR NOTE

This work was supported in part by the University of South Florida, College of Education, through a minigrant awarded to J.M.F. We thank C. Mitchell Dayton and Michael Harwell for comments on portions of this research that were presented at the 2007 and 2008 annual meetings of the American Educational Research Association. Correspondence concerning this article should be sent to J. M. Ferron, University of South Florida, 4202 E. Fowler Ave., EDU 162, Tampa, FL 33620-7750 (e-mail: ferron@tempest.coedu.usf.edu).

REFERENCES

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon.

Bock, M. A. (2007). A social–behavioral learning strategy intervention for a child with Asperger syndrome: Brief report. *Remedial & Special Education*, **28**, 258-265. doi:10.1177/07419325070280050101

Bulté, I., & Onghena, P. (2009). Randomization tests for multiple-baseline designs: An extension of the SCRT-R package. *Behavior Research Methods*, **41**, 477-485.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment*, **10**, 229-242.

Dukes, W. F. (1965). *N* = 1. *Psychological Bulletin*, **64**, 74-79. doi:10.1037/h0021964

Fai, A. H.-T., & Cornelius, P. L. (1996). Approximate *F*-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation & Simulation*, **54**, 363-378.

Ferron, J. [M.], Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*, **37**, 379-403. doi:10.1207/S15327906MBR3703_4

Ferron, J. [M.], & Scott, H. (2005). Multiple baseline design. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 2, pp. 941-945). Hoboken, NJ: Wiley.

Fouladi, R. T., & Shieh, Y.-Y. (2004). A comparison of two general approaches to mixed model longitudinal analyses under small sample size conditions. *Communications in Statistics: Simulation & Computation*, **33**, 807-824. doi:10.1081/SAC-200033260

Glover, D. S., Brown, G. P., Fairburn, C. G., & Shafran, R. (2007). A preliminary evaluation of cognitive-behaviour therapy for clinical perfectionism: A case series. *British Journal of Clinical Psychology*, **46**, 85-94. doi:10.1348/014466506X117388

Gomez, E. V., Schaalje, G. B., & Fellingham, G. W. (2005). Performance of the Kenward–Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics: Simulation & Computation*, **34**, 377-392. doi:10.1081/SAC-200055719

Gresham, F. M., McIntyre, L. L., Olson-Tinker, H., Dolstra, L., McLaughlin, V., & Van, M. (2004). Relevance of functional behavioral assessment research for school-based interventions and positive behavioral support. *Research in Developmental Disabilities*, **25**, 19-37. doi:10.1016/j.ridd.2003.04.003

Hox, J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147-154). New York: Springer.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment*, **7**, 107-118.

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

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

Koehler, M. J., & Levin, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods*, **3**, 206-217. doi:10.1037/1082-989X.3.2.206

Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted *F* tests. *Educational & Psychological Measurement*, **64**, 224-242. doi:10.1177/0013164403260196

Kratochwill, T. R., & Piersel, W. C. (1983). Time-series research: Contributions to empirical clinical practice. *Behavioral Assessment*, **5**, 165-176.

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, **58**, 127-137. doi:10.1046/j.0039-0402.2003.00252.x

Mahar, M. T., Murphy, S. K., Rowe, D. A., Golden, J., Shields, A. T., & Raedeke, T. D. (2006). Effects of a classroom-based program on physical activity and on-task behavior. *Medicine & Science in Sports & Exercise*, **38**, 2086-2094.

Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment*, **10**, 1-28.

Matyas, T. A., & Greenwood, K. M. (1997). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215-243). Mahwah, NJ: Erlbaum.

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

Mok, M. (1995). *Sample size requirements for 2-level designs in educational research*. Unpublished manuscript, Macquarie University, Sydney, Australia.

Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design: Bringing science to managed care. *American Psychologist*, **56**, 119-127. doi:10.1037/0003-066X.56.2.119

Musti-Rao, S., & Cartledge, G. (2007). Effects of a supplemental early reading intervention with at-risk urban learners. *Topics in Early Childhood Special Education*, **27**, 70-85. doi:10.1177/02711214070270020301

Naylor, E. V., Antonuccio, D. O., Johnson, G., Spogen, D., & O'Donohue, W. (2007). A pilot study investigating behavioral prescriptions for depression. *Journal of Clinical Psychology in Medical Settings*, **14**, 152-159. doi:10.1007/s10880-007-9064-9

Nugent, W. (1996). Integrating single-case and group-comparison designs for evaluation research. *Journal of Applied Behavioral Science*, **32**, 209-226. doi:10.1177/0021886396322007

O'Callaghan, P. M., Allen, K. D., Powell, S., & Salama, F. (2006). The efficacy of noncontingent escape for decreasing children's disruptive behavior during restorative dental treatment. *Journal of Applied Behavior Analysis*, **39**, 161-171. doi:10.1901/jaba.2006.79-05

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15-40). Hillsdale, NJ: Erlbaum.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, **6**, 309-316. doi:10.1007/BF02288586

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2001). Approximations to distributions of test statistics in complex mixed linear models using SAS Proc MIXED. In *Proceedings of the SAS Users Group International 26th Annual Conference* (Paper 262–26). Available at support.sas.com/events/sasglobalforum/previous/index.html.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasiexperimental designs for generalized causal inference*. Boston: 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.

Singh, N. N., Lancioni, G. E., Winton, A. S. W., Adkins, A. D., Singh, J., & Singh, A. N. (2007). Mindfulness training assists individuals with moderate mental retardation to maintain their community placements. *Behavior Modification*, **31**, 800-814. doi:10.1177/0145445507300925

TIGER, J. H., HANLEY, G. P., & HERNANDEZ, E. (2006). An evaluation of the value of choice with preschool children. *Journal of Applied Behavior Analysis*, **39**, 1-16. doi:10.1901/jaba.2006.158-04

TSAO, L.-L., & ODOM, S. L. (2006). Sibling-mediated social interaction intervention for young children with autism. *Topics in Early Childhood Special Education*, **26**, 106-123. doi:10.1177/02711214060260020101

VAN DEN NOORTGATE, W., & ONGHENA, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, **18**, 325-346. doi:10.1521/scpq.18.3.325.22577

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.

VELICER, W. F., & FAVA, J. L. (2003). Time series analysis. In I. B. Weiner (Series Ed.) and J. Schinka & W. F. Velicer (Vol. Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 581-606). Hoboken, NJ: Wiley.

WAMBAUGH, J. L., & FERGUSON, M. (2007). Application of semantic feature analysis to retrieval of action names in aphasia. *Journal of Rehabilitation Research & Development*, **44**, 381-394.

APPENDIX

Methods for Computing Degrees of Freedom for Tests of Fixed Effects

The residual method defines the degrees of freedom as $n - \text{rank}\ (\mathbf{X})$, where n is the total number of Level 1 units, **X** is the design matrix for the fixed effects, and rank (**X**) will correspond to the number of fixed effects (gamma coefficients). This approach was the default in the earliest versions of PROC MIXED and only gives the correct degrees of freedom when the Level 1 errors are independent and identically distributed, and when there are no Level 2 errors (i.e., a situation in which multilevel modeling is not needed).

The between–within method partitions the residual degrees of freedom into between-participants and within-participants portions. Fixed effects of design variables that change within an individual are assigned the within-participants degrees of freedom. Effects of design variables that do not change within an individual are assigned the between-participants degrees of freedom.

The containment method searches the random statement for effects that match the fixed effect being tested, and then considers the rank contribution of these random effects to the (**X Z**) matrix, where **X** is the design matrix of the fixed effects and **Z** is the design matrix for the random effects. The degrees of freedom for these effects are defined to be equal to the smallest of these rank contributions, which for the type of model considered equals the number of Level 2 units minus the number of fixed effects within the Level 2 equation containing the fixed effect of interest. For effects that are not listed in the random statement, the degrees of freedom are computed as $n - \text{rank}\ (\mathbf{X\,Z})$. The containment method is the default method in PROC MIXED when a random statement is used, and can lead to exact degrees of freedom when the design is balanced, and when the Level 1 errors are independent and identically distributed. This method becomes more questionable as the design becomes less balanced or when a more complex error structure is needed at Level 1.

The Satterthwaite method approximates the degrees of freedom, and is designed for use with unbalanced designs and more complex covariance structures. The method used in SAS is a generalization of the procedure described by Fai and Cornelius (1996), which builds on the work of Satterthwaite (1941). The degrees of freedom are estimated as

$$df = \frac{2\left(c'\hat{\Sigma}_{\hat{\beta}}c\right)^2}{\left[\text{var}\left(c'\hat{\Sigma}_{\hat{\beta}}c\right)\right]},$$

where c is the vector of constants defining the contrast of interest (H_0:$c'\beta = 0$) and $\hat{\Sigma}_{\hat{\beta}}$ is the approximate covariance matrix of $\hat{\beta}$:

$$\hat{\Sigma}_{\hat{\beta}} = (X'\hat{V}^{-1}X)^{-1},$$

where $\hat{V}^{-1}$ is the inverse variance–covariance matrix of y, the vector of responses, and X is the design matrix of the fixed effects.

The Kenward–Roger method also approximates the degrees of freedom, and again is designed for use with unbalanced designs and complex covariance structures. The method, which was developed by Kenward and Roger (1997), inflates $\hat{\Sigma}_{\hat{\beta}}$ to adjust for small-sample bias. The new variance–covariance estimate, $\hat{\Sigma}^*_{\hat{\beta}}$, is then used with the Satterthwaite method to compute the degrees of freedom.

(Manuscript received December 19, 2008;
accepted for publication December 24, 2008.)