Behavior Research Methods
2010, 42 (4), 930-943
doi:10.3758/BRM.42.4.930

Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study of multilevel-modeling approaches

John M. Ferron, Jennie L. Farmer, and Corina M. Owens
University of South Florida, Tampa, Florida

While conducting intervention research, researchers and practitioners are often interested in how the intervention functions not only at the group level, but also at the individual level. One way to examine individual treatment effects is through multiple-baseline studies analyzed with multilevel modeling. This analysis allows for the construction of confidence intervals, which are strongly recommended in the reporting guidelines of the American Psychological Association. The purpose of this study was to examine the accuracy of confidence intervals of individual treatment effects obtained from multilevel modeling of multiple-baseline data. Monte Carlo methods were used to examine performance across conditions varying in the number of participants, the number of observations per participant, and the dependency of errors. The accuracy of the confidence intervals depended on the method used, with the greatest accuracy being obtained when multilevel modeling was coupled with the Kenward–Roger method of estimating degrees of freedom.

Single-case designs allow for the examination of intervention effects for either a single participant or a single case (e.g., one class of students with data collected at the classroom level). During a single-case study, data are collected at multiple points over time, allowing for the inspection of intervention effects over time. This research design typically includes a baseline phase, in which data are collected prior to the implementation of the intervention, and an intervention phase. Additional designs allow for the removal of the intervention, the reintroduction of the intervention, and the maintenance of the intervention. Furthermore, several cases can be examined together in a multiple-baseline design. This design can be used to study multiple participants, multiple settings, or multiple behaviors and is recommended to have a minimum of three (Barlow & Hersen, 1984) or four baselines (Kazdin & Kopel, 1975). A graphical display of a multiple-baseline design is provided in Figure 1. The figure is composed of a separate line graph for each of the 4 participants: Marie, Claire, Cody, and Chloe. For each participant, the outcome (minutes reading) is graphed across time, with a vertical line separating the baseline phase from the intervention phase. Note that the intervention begins at a different point in time for each participant, which produces baselines of different lengths (or multiple baselines).

Single-case designs, such as multiple-baseline designs, offer several advantages over group designs. Single-case designs allow researchers to investigate intervention effects at the individual level rather than strictly at the group level. This provides researchers with more information about the intervention effects, because variation in individual effects is lost or obscured in the average effects reported in group-design studies (Barlow & Hersen, 1984; Morgan & Morgan, 2001). In addition, it allows for examination of the intervention and its effects over time, since there are multiple data points, rather than a single score for each participant. Because of the nature of the design, it is also particularly well suited for populations with low prevalence rates, since large samples are not required. Furthermore, single-case and multiple-baseline studies allow researchers to be responsive to the needs of the participant(s), because data are collected at multiple points; therefore, if patterns emerge such that the intervention is not effective, it can be modified. Finally, these designs reduce the gap between research and practice by allowing practitioners to implement research in their current settings.

Among single-case designs, the multiple-baseline design is often preferred, because the staggering of the intervention (see Figure 1) makes it difficult to attribute changes to maturation or history (i.e., an event that just happened to coincide with the intervention). 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 (Swanson & Sachse-Lee, 2000), counseling (Sharpley, 1981), and reading (Neuman & McCormick, 1995). A search of

J. M. Ferron, ferron@usf.edu

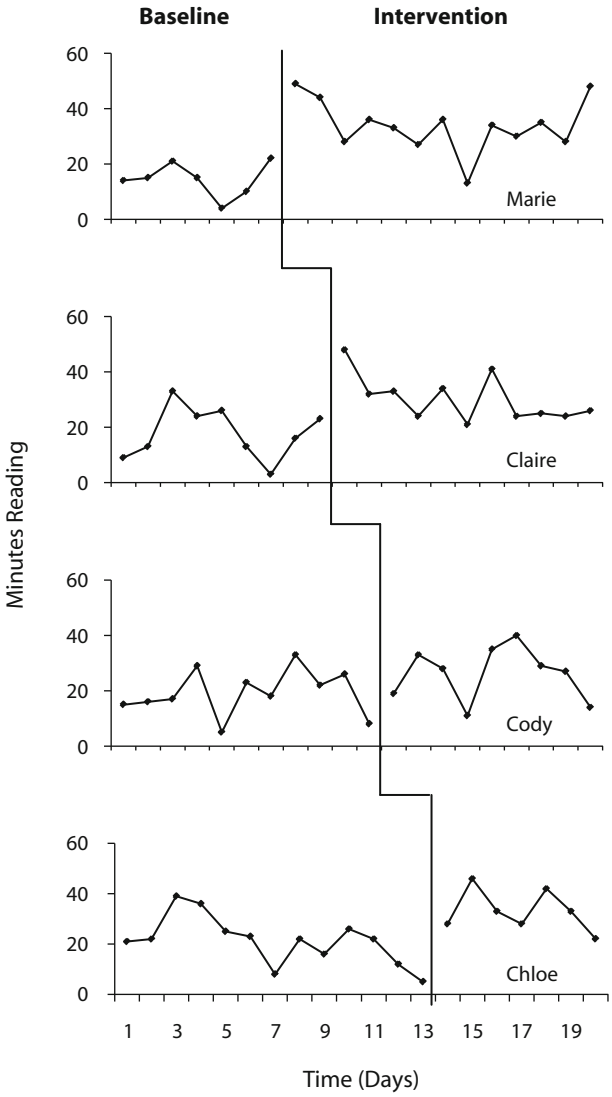

Figure 1. Graphical display of a multiple-baseline design. The outcome (minutes reading) is graphed across time (days) for each of the 4 participants. The staggered vertical line separates the baseline phases from the intervention phases.

the Web of Science database using "multiple baseline" produced 75 studies that were published in 2008. A survey of the first 20 studies showed that the number of baselines ranged from 3 to 10, with a median of 4, and the average number of observations in an individual's time series ranged from 7 to 58, with a median of 24. These studies included interventions to improve academic achievement, social skills performance, communication skills, daily task completion, training procedures, and healthy living with students with disabilities, students who are at risk for school failure, adults in the workplace and college settings, and patients with Alzheimer's disease, thus indicating a broad use of this research design.

Furthermore, the current movement in education and psychology to use interventions within a response to intervention (RTI) framework (Glover & DiPerna, 2007) for students who are struggling and to assist in the identification of students with disabilities for special education services creates an increased need to evaluate the effectiveness of interventions over time and at the individual level. For example, an essential aspect of mathematics RTI is the use of individual students' performance data to make crucial tiered instructional decisions regarding how to best improve mathematical learning outcomes (Allsopp, McHatton, Ray, & Farmer, 2010). Both for practitioners that need to evaluate intervention effectiveness at the individual level and for researchers committed to the belief that interventions and their effects will be more fully understood if they are studied at the individual level, it is critical to understand the quality of inferences made about individual treatment effects.

Need for Confidence Intervals of Individual Treatment Effects

An individual *treatment effect* (T) is generally defined as the difference between what would be observed for the individual under the treatment condition (Y_t) and what would be observed for that individual under the control condition (Y_c); therefore, the treatment effect is the difference between two potential responses: $T = Y_t - Y_c$ (Gadbury & Iyer, 2000; Holland, 1986; Rubin, 1974). The inability to observe the same participant at the same time under both treatment and control conditions is generally referred to as the fundamental problem of causal inference. To overcome this problem, assumptions have to be made, and treatment effect inferences become uncertain (Holland, 1986).

In the context of single-case research, the individual *treatment effect* can be defined as the difference between what is observed during treatment for a participant and what would have been observed for that participant had no intervention taken place. Unfortunately, we do not know exactly what would have been observed if we had not intervened. Instead, we make assumptions about how the outcome changes with time and then use these assumptions along with baseline observations to make projections about what would have happened had we not intervened. Therefore, any estimate of the intervention effect is done with some uncertainty, which, to some extent, is tied to the amount of unexplained variation (or instability) in the baseline observations. To reduce uncertainty, researchers often design their single-case studies in ways that minimize the variation in baseline observations (e.g., making each observation in the same setting), but even with such efforts, some unexplained variation remains. Consequently, when researchers focus on individual treatment effects, they should not only provide point estimates of the individual effects, but should also provide measures of precision to index the uncertainty in these estimates. The American Psychological Association (2010) strongly recommends the reporting of confidence intervals for estimates like treatment effects, because they provide information on both location and precision.

Although the creation of confidence intervals for individual treatment effects is desirable, not all of the analyses that have been recommended and used with multiple-baseline data can be used to create confidence intervals.

Graphical displays coupled with visual analyses are widely advocated because of the clarity with which they communicate the observed data (Ferron & Jones, 2006; Parsonson & Baer, 1992), but they do not result in confidence intervals of individual treatment effects. Similarly, randomization tests, which are recommended because of their ability to control the Type I error rate (Edgington, 1980; Ferron & Sentovich, 2002; Koehler & Levin, 1998), also fail to provide a mechanism for creating confidence intervals for individual effects. Regression-based approaches, however, can provide confidence intervals for individual treatment effect estimates. Time series analysis provides one option, but these analyses are generally not considered viable for series with less than 50 observations (Box, Jenkins, & Reinsel, 1994). We will focus on ordinary least squares (OLS) regression and multilevel modeling, both of which have been recommended for use with relatively short series.

OLS Regression

The simplest estimate of an individual treatment effect is obtained by taking the mean of the treatment phase observations for a participant and subtracting the mean of the baseline phase observations for that participant. These means are the most common statistics reported in multiple-baseline studies (Ferron & Jones, 2002), and a confidence interval around the mean difference can be obtained by analyzing the interrupted time series data from the individual using the following regression model:

$$y_i = \beta_0 + \beta_1 phase + e_i, \quad (1)$$

where y_i is the observed value at the ith point in time, *phase* is a dummy coded variable (0, *baseline*; 1, *treatment*), β_0 is the baseline mean, β_1 is the difference in means between baseline and treatment phases (i.e., the individual treatment effect), and e_i is error at the ith point in time, which accounts for within-phase variation around the phase mean.

The regression model can be adapted to accommodate trends in the phases (e.g., Center, Skiba, & Casey, 1985; Huitema & McKean, 2000) and would typically be estimated using OLS methods. The use of OLS regression for analyzing multiple-baseline data has been advocated (Huitema & McKean, 1998), but the use of OLS regression methods has also raised concerns, because the errors in the model are assumed to be independent. Many have argued that errors closer in time may be more similar to each other than independently selected errors, and, therefore, the errors may be positively autocorrelated instead of independent (Kratochwill et al., 1974; Matyas & Greenwood, 1997). Furthermore, positive autocorrelation impacts the statistical inferences such that there is a greater chance of Type I errors (Greenwood & Matyas, 1990; Toothaker, Banz, Noble, Camp, & Davis, 1983), which in turn implies that 95% confidence intervals would contain the actual effect less than 95% of the time. Although the negative effects of autocorrelation on statistical inferences are not contended, there has been considerable debate about the degree to which behavioral time series data are autocorrelated (Busk & Marascuilo, 1988; Huitema, 1985; Huitema & McKean, 1998; Matyas & Greenwood, 1997) and, therefore, the degree to which OLS estimates may be inappropriate.

Multilevel Modeling

Multilevel models have been suggested as an alternative method for analyzing multiple-baseline data (Nugent, 1996; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003a, 2003b, 2007). The first level of the multilevel model can be defined in a manner that mirrors the OLS regression model:

$$y_{ij} = \beta_{0j} + \beta_{1j} phase + e_{ij}, \quad (2)$$

where y_{ij} is the observed value at the ith point in time for the jth participant; *phase* is a dummy coded variable (0, *baseline*; 1, *treatment*); β_{0j} is the baseline mean for the jth participant; β_{1j} is the difference in means between baseline and treatment phases for the jth participant (i.e., the jth participant's treatment effect); and e_{ij} is error at the ith point in time for the jth participant, which accounts for within-phase variation around the phase mean. The errors for the jth participant could be assumed to be independent with a variance of σ^2 or could be assumed to have a more complex covariance structure, such as an autoregressive structure, which would allow the errors that were closer together in time to be more similar. As with OLS, the first level of the multilevel model (Equation 2) could be expanded to accommodate trends in the phases (Van den Noortgate & Onghena, 2003b).

The second level of the multilevel model can be defined to account for variation in participants' baseline levels and variation in the participants' treatment effects:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \quad (3)$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \quad (4)$$

where γ_{00} is the average baseline level and γ_{10} is the average treatment effect. The error u_{0j} indicates how far the jth participant's baseline level is from the average baseline level and is assumed to have a mean of 0 and a variance of τ_{00}. The u_{1j} error is the discrepancy between the jth participant's treatment effect and the average treatment effect and is assumed to have a mean of 0 and a variance of τ_{11}.

Multilevel modeling leads to three different types of parameter estimates. First, there are variance components, such as the between-participants variance in baseline levels (i.e., τ_{00}) and the between-participants variance in treatment effects (i.e., τ_{11}). Second, there are fixed effects, which provide estimates of the average baseline level (i.e., γ_{00}) and the average treatment effect (i.e., γ_{10}). Finally, there are estimates of the random effects, which provide the regression coefficients for each participant. More specifically, β_{0j} is the baseline level for the jth participant, and β_{1j} is the treatment effect for the jth participant.

Multilevel-Modeling Estimates of Individual Effects

The β_{1j} estimates from the multilevel model estimate the same individual treatment effect parameters that are estimated through OLS regression, but it is expected that

the multilevel-modeling estimates will differ from the OLS estimates. OLS estimates of an individual effect are based entirely on the data from the individual. Multilevel-modeling estimates, however, are empirical Bayes estimates, which depend not only on the data from the individual, but also on the data from the other participants. Empirical Bayes estimates are obtained by creating a weighted average of an estimate that is based on information only from the individual and an estimate that is based on the average of all of the participants' data (Raudenbush & Bryk, 2002). When we have a very reliable estimate based on the information from the individual, the average estimate is given very little weight. Under these circumstances, the empirical Bayes estimates would be expected to be very similar to the OLS estimates. As series lengths become shorter and the time series data more variable, the estimates based on the data from one participant become less reliable. As these estimates become less reliable, more weight is given to the average estimate.

Empirical Bayes estimates are expected to be biased (Raudenbush & Bryk, 2002), which means that if we could replicate the study many times for the same individual, the average of the estimates from these studies would not be expected to be the true individual treatment effect. This bias results because the estimate is not based solely on the individual's data; rather, it is a weighted average of the estimate based on the individual's data and the estimate based on averaging all of the participants' data. Although bias is problematic, there is an expectation that the process of using a weighted average will help to stabilize unreliable individual estimates. This should then lead to individual estimates that tend to be closer to the true individual effects than estimates that are based solely on the individual's data (Raudenbush & Bryk, 2002). This expectation was examined in a simulation study of a traditional longitudinal design where the number of measurement occasions was varied from 7 to 13 and the number of participants was varied from 25 to 250 (Candel & Winkens, 2003). Using the mean square error to index the closeness of the estimates to the true parameter values, Candel and Winkens found that empirical Bayes estimates outperformed OLS estimates for all sample sizes studied.

Multilevel-Modeling Confidence Intervals for Estimates of Individual Effects

There are reasons to expect not only that the point estimates will differ, but also that the confidence intervals will differ between OLS and multilevel modeling. The same general approach of computing the confidence interval for an individual effect is used in these two methods:

$$CI = \hat{\beta} \pm t_{\text{crit}} * SE, \tag{5}$$

where $\hat{\beta}$ is the individual effect estimate, t_{crit} is the critical value from the t distribution corresponding to the desired level of confidence, and SE is the standard error. As was previously mentioned, the treatment effect estimates ($\hat{\beta}$) will vary between OLS and multilevel modeling. In addition, OLS assumes that the errors are independent, whereas multilevel modeling can allow for serial dependency (or autocorrelation). The estimation of autocorrelation has consequences for the SE and, thus, the width of the confidence intervals. Finally, how the degrees of freedom are estimated impacts the critical t value and, thus, the width of the confidence interval. Although a variety of degrees of freedom estimates can be utilized in both OLS and multilevel-modeling frameworks, applied researchers typically select among the degrees of freedom methods that are easily accessible in standard statistical software (e.g., SAS, SPSS). The choices readily available differ between the procedures for OLS estimation (e.g., the REG procedure in SAS) and the procedures for multilevel modeling (e.g., the MIXED procedure in SAS).

For the OLS model shown in Equation 1, the degrees of freedom are typically computed as

$$df_{\text{OLS}} = n - 2, \tag{6}$$

where n is the number of observations in the individual's time series. In multilevel modeling, there are a variety of methods to estimate the degrees of freedom that have been programmed as options for those using SAS. The specific options include the containment, residual, between–within, Satterthwaite, and Kenward–Roger methods. We will discuss each of these methods as they apply to estimating the degrees of freedom for the individual treatment effects (i.e., the random effects) from the model defined in Equations 2–4 and further specified in the SAS programming lines provided in the Appendix. For more general and detailed descriptions of these degrees of freedom methods, see SAS Institute (2004); Schaalje, McBride, and Fellingham (2001); and Kenward and Roger (1997).

The default method when using the MIXED procedure in SAS, which is known as the containment method, varies its computation method depending on the model specified. For the model shown in Equations 2–4 and specified in the programming lines in the Appendix, the degrees of freedom for the individual treatment effects (i.e., the random effects) are estimated as

$$df_{\text{containment}} = n_2(n_1 - 2), \tag{7}$$

where n_2 is the number of participants and n_1 is the number of observations in each participant's time series. An alternative is the residual method, which for the model specified in Equations 2–4 and the Appendix simplifies to

$$df_{\text{residual}} = (n_2 * n_1) - 1. \tag{8}$$

Another option is the between–within option, which generally partitions the residual degrees of freedom into between-participants and within-participants portions. For the individual treatment effects of the model specified in Equations 2–4 and the Appendix, all of the residual degrees of freedom are given to the within-participants portion, and therefore, the degrees of freedom using the between–within method turn out to be the same as the residual degrees of freedom for this application.

These relatively simple methods for estimating the degrees of freedom tend to overestimate the degrees of freedom when there is a complex covariance structure, such as autocorrelated errors, so other degrees of freedom es-

timation methods have been developed. The Satterthwaite (1941) method uses the variance–covariance matrix of the observed time series to approximate the degrees of freedom using a generalization of the procedure described by Fai and Cornelius (1996), which builds on the work of Satterthwaite. Unlike the containment, between–within, and residual methods, when the Satterthwaite method is used for the model specified in Equations 2–4 and the Appendix, it typically produces a different degrees of freedom estimate for each individual treatment effect. Note that this degrees of freedom method is the one used by the SPSS MIXED procedure.

The Kenward–Roger method (Kenward & Roger, 1997) is an extension of the Satterthwaite method. More specifically, Satterthwaite-type degrees of freedom are computed, but the computation is made after the estimated variance–covariance matrix of fixed and random effects is inflated to adjust for bias. The method for inflating the variance–covariance matrix of fixed and random effects is described by Harville and Jeske (1992) and by Prasad and Rao (1990). With some data sets, the method leads to no change in the estimated variance–covariance matrix of fixed and random effects, and therefore, the Satterhwaite and Kenward–Roger methods may return the same degrees of freedom estimates for the individual treatment effects. When sample sizes are very large, the differences among the degrees of freedom methods are expected to have negligible effects on the confidence intervals, but as sample size gets smaller, the differences are expected to result in meaningful differences in the confidence intervals.

Research on the Functioning of Multilevel Modeling With Multiple-Baseline Data

Interestingly, research into the statistical functioning of multilevel models with multiple-baseline data has been focused on the variance components and the average effects, and the individual effects have not been examined. Ferron, Bell, Hess, Rendina-Gobioff, and Hibbard (2009) conducted a Monte Carlo simulation study that looked at multiple-baseline studies having 4, 6, or 8 participants in which series lengths of 10, 20, or 30 observations were used. They found that the variance components were substantially biased. More specifically, using restricted maximum-likelihood estimation, they found that as the number of participants went from 4 to 6 to 8, the average relative bias estimates for the variance in the treatment effect went from 0.34 to 0.25 to 0.21, which indicates a substantial bias even in the largest sample size examined.

These results were not too surprising, given other research in which estimation of the variance components in multilevel models for more traditional longitudinal designs was examined. Monte Carlo studies of growth curve models having as few as 30 participants and series lengths of 4 or 8 (Kwok, West, & Green, 2007), and series lengths of 3–12 (Ferron, Dailey, & Yi, 2002) have all shown substantial biases in the variance components when the model was misspecified. Furthermore, in two of these studies (Kwok et al., 2007; Murphy & Pituch, 2009) biases in the variance components under correct model specification were also examined, and when the number of participants was small ($N = 30$), the bias was substantial. For example, Murphy and Pituch found that when the model was correctly specified; the number of participants was 30; the series length was 5; and the autoregressive and moving average parameters were .5 and .3, respectively; the relative bias in the intercept variance was .20 and the relative bias in the slope variance was .17.

Research on the estimation of average effects, however, has revealed that it is possible to obtain accurate results even when sample sizes are small. In a study of multiple-baseline designs, Ferron et al. (2009) found that the fixed effect estimate of the average treatment effect was unbiased for all sample sizes studied. In addition, it was shown that accurate confidence intervals could be obtained for the average treatment effect if an autocorrelated error structure was specified and either the Kenward–Roger or Satterthwaite method was used to estimate the degrees of freedom. Under these circumstances the proportion of times that the confidence interval contained the parameter value for 95% confidence intervals ranged from .935 to .965 (i.e., the coverage estimates were close to the nominal value).

The value of using the Kenward–Roger method to estimate degrees of freedom for fixed effect inferences has also been considered more generally in the methodological literature on the application of multilevel models to repeated measures data. This method is theoretically the most appropriate when sample size is small and there is a complex covariance structure, but exactly how well it will work and the degree to which it will provide superior performance has not been derived mathematically. Consequently, its performance has been examined in Monte Carlo studies (Fouladi & Shieh, 2004; Gomez, Schaalje, & Fellingham, 2005; Kenward & Roger, 1997; Kowalchuk, Keselman, Algina, & Wolfinger, 2004; Schaalje et al., 2001). In these studies, designs in which the participants came from two to five groups and were repeatedly measured from three to five times were considered. The number of participants ranged from 9 to 45 across these simulation studies, and even with these relatively small sample sizes, accurate hypothesis tests (and confidence intervals) were obtained for most conditions. An exception was found by Gomez et al. (2005) when they examined a three-group design with 3 participants per group in which each participant was measured at three points in time. When data were generated and analyzed assuming compound symmetry, the estimated Type I error rate was .0525 ($\alpha = .05$), but when the data were generated and analyzed assuming a first-order autoregressive with random effects model, the Type I error rate for the treatment effect was estimated to be .1165 ($\alpha = .05$).

A general conclusion that can be drawn from the studies in which the functioning of multilevel modeling with small sample sizes was examined is that the degree to which multilevel modeling functions appropriately depends on the type of parameter being estimated. When the focus is on a variance component (e.g., the variance in treatment effects), multilevel modeling has problems

when the sample size is small, but when the focus is on a fixed effect (e.g., the average treatment effect), multilevel modeling often performs well even with small sample sizes, as long as the error structure is correctly specified and the degrees of freedom are appropriately estimated. Given the inability to mathematically derive performance under small sample size conditions, and the variation in performance when focus shifts from variance components to fixed effects, it is unclear how multilevel modeling will perform when the focus is on the individual treatment effect estimates.

Purpose

An argument has been made for the importance of estimating individual treatment effects and the corresponding confidence intervals from multiple-baseline data. Traditional OLS methods can be used to do this but are known to create inaccurate confidence intervals when errors are autocorrelated. Multilevel modeling provides an alternative approach, which allows the autocorrelation to be modeled. The purpose of this study was to examine the accuracy of multilevel-modeling estimates of individual treatment effects and their confidence intervals. More specifically, the goal was to examine bias and variability (mean square error, MS_e) in the empirical Bayes estimates and to examine the coverage and width of the corresponding confidence intervals, which were constructed using one of three different methods of estimating the degrees of freedom: the Kenward–Roger, the Satterthwaite, or the containment method. These estimates were examined for conditions that varied in the number of participants, series length, the level of autocorrelation, variance among participants in initial level, and variance among participants in the treatment effect. To provide a comparison with these multilevel-modeling approaches, OLS estimates and confidence intervals were also examined by doing a separate analysis for each individual time series.

METHOD

Monte Carlo simulation methods were used to examine the multilevel-modeling estimates of individual effects and their confidence intervals in the context of multiple-baseline studies. The number of simulated participants (or baselines) was 4, 6, or 8. These numbers were selected on the basis that multiple-baseline studies have been recommended to have at least four baselines (Kazdin & Kopel, 1975); consistency with the other simulation study of multilevel modeling of multiple-baseline data (Ferron et al., 2009); and our survey of multiple-baseline studies published in 2008, which showed studies having from 3 to 10 baselines with a median of 4.

The series lengths were simulated to be 10, 20, or 30 observations. These values were also selected on the basis of multiple considerations. In our survey, we found average series lengths that ranged from 7 to 58, with a median of 24, and in a meta-analysis of 85 single-case studies, Swanson and Sachse-Lee (2000) found that 25 studies had <11 treatment sessions, 37 studies had between 11 and 29 treatment sessions, and 23 studies had >29 treatment sessions. In addition, the values of 10, 20, and 30 are consistent with the other simulation study of multilevel modeling of multiple-baseline data (Ferron et al., 2009). Moreover, we wanted to cross series length with the number of participants in the Monte Carlo design, and the minimum possible series length in a multiple-baseline design with 8 participants, where each would have a separate baseline, would be 9. On the basis of these considerations, a series length of 10 was established as the minimum for the study, and although series lengths of over 30 are sometimes used in practice, we expected that statistical theory and the results of our study would allow us to generalize to longer series lengths.

For each simulated study, the implementation of treatment was staggered, such that each successive participant had a baseline that was one observation longer than that of the previous participant when the series length was 10, two observations longer when the series length was 20, and three observations longer when the series length was 30. As a result, the length of the baseline phases varied among participants within studies but also varied across studies that had different series lengths and numbers of participants. The design with 4 participants and series lengths of 20 is illustrated in Figure 1. When the number of participants was 4 and the series length was 30, the baseline lengths were at their longest, with values of 10, 13, 16, and 19 for the 4 participants. When the number of participants was 8 and the series length was 10, the baseline lengths were at their shortest, with values of 1–8 for the 8 participants.

The data were generated on the basis of the multilevel model shown in Equations 2–4. At the first level, an outcome at the *i*th time for the *j*th participant (y_{ij}) was modeled as a linear function of a single predictor, *phase*,

$$y_{ij} = \beta_{0j} + \beta_{1j} phase + e_{ij}, \quad (9)$$

where *phase* was a dichotomous variable indicating whether the observation was from the baseline or treatment phase, β_{0j} was the average level of the outcome during baseline for the *j*th participant, and β_{1j} was the treatment effect for the *j*th participant. This within-participants model was consistent with the model used by Ferron et al. (2009) in the reanalyses of four multiple-baseline studies and was consistent with the multilevel-modeling application presented by Van den Noortgate and Onghena (2003a), which was focused on a reversal design that was replicated across 6 participants. Furthermore, because it was the most basic interrupted time-series model (e.g., there were no trends, changes in trends, or seasonal effects), it appeared to be the most appropriate model for an initial study into the multilevel modeling of multiple-baseline data. If accurate confidence intervals of individual effects could not be obtained in this model, one would not expect them to be obtained from more complex models. Errors for the within-participants model (e_{ij}) were generated using the ARMASIM function in SAS Version 9.1 (SAS Institute, 2005), with a variance (σ^2) of 1.0 and an autocorrelation (ρ) of 0, .1, .2, .3, or .4. These autocorrelation values were selected to cover the range expected in behavioral studies and were based on a review of studies conducted to index the degree to which errors in behavioral data are autocorrelated (Busk & Marascuilo, 1988; Huitema, 1985; Matyas & Greenwood, 1997).

At the second level, baseline levels (i.e., intercepts) and treatment effects (i.e., shifts) of the first-level model were allowed to vary randomly,

$$\beta_{0j} = \gamma_{00} + u_{0j}, \quad (10)$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \quad (11)$$

where γ_{00} was the average baseline level and γ_{10} was the average treatment effect, each set to a value of 1.0. Level 2 errors were generated from a normal distribution using the RANNOR random number generator in SAS Version 9.1 (SAS Institute, 2005). The variance of u_{0j} (τ_{00}) was equal to 0.1 or 0.5, the variance of u_{1j} (τ_{11}) was equal to 0.1 or 0.5, and the covariance between u_{0j} and u_{1j} was 0.

The variances of the Level 2 errors were chosen such 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 four reanalyses of recently published multiple-baseline studies that were conducted using multilevel models (Ferron et al., 2009) and with

the multilevel-modeling application for a replicated ABAB design presented by Van den Noortgate and Onghena (2003a).

Crossing the two variance levels of u_{0j} with the two variance levels of u_{1j} and the five levels of autocorrelation, a total of 20 data conditions were examined for each of the nine combinations of sample size with series length. For each of these 180 conditions (20 * 9), 1,000 data sets were simulated using SAS IML (SAS Institute, 2005). The use of 1,000 replications leads to 4,000, 6,000, or 8,000 individual effect estimates depending on whether the number of participants was 4, 6, or 8. With 4,000 effect estimates, the *SE* would be .003 if the coverage was .95, and the *SE* would be even smaller for conditions with 6 or 8 participants.

Each data set was analyzed using a separate OLS analysis for each time series by using the REG procedure with a BY statement. The regression model, which was consistent with the model in Equation 1, provided an estimate of the baseline level and the individual treatment effect. The individual treatment effect, or shift in level that occurs with intervention, is the difference between the baseline mean and the intervention mean. Each data set was also analyzed using multilevel modeling with RML estimation via the MIXED procedure. The estimated multilevel model was consistent with data generation. For each participant, a baseline level was estimated as well as the treatment effect (i.e., the shift in the level that occurred with intervention). Estimates were also obtained for the autocorrelation, variance within participants, variance in baseline levels, and variance in treatment effects. The empirical Bayes estimates, which are not part of the default output, were obtained following the procedure described by Van den Noortgate and Onghena (2003a). The specific programming lines that were used to estimate the multilevel model and to obtain the empirical Bayes estimates of the individual effects, along with their confidence intervals, are shown in the Appendix.

Confidence intervals for the individual effects from the multilevel models were obtained using three alternative approaches for estimating the degrees of freedom: the Kenward–Roger method, the Satterthwaite method, and the containment method. The residual and between–within methods were not included, because they provide the largest degrees of freedom estimates and are expected to overestimate the degrees of freedom in this context. The containment method is also expected to overestimate the degrees of freedom, although not by as much. It was included because it is the default method in the MIXED procedure and is therefore likely to be used in practice. By including the containment method as a comparison to the more complex Satterthwaite and Kenward–Roger approaches, it is possible to address the practical question of whether the theoretical advantages of these more complex methods materialize to an extent that warrants moving from the more familiar default approach.

For each of the 180 conditions, we obtained the estimates of the bias and MS_e for each method of making point estimates of the individual treatment effects and the coverage and width for each method of making the confidence intervals for the individual treatment effects. More formally,

$$\text{bias} = \frac{\sum_{k=1}^{1,000}\sum_{j=1}^{n_2}\left(\hat{\beta}_{1jk} - \beta_{1jk}\right)}{1,000 * n_2}, \tag{12}$$

where $\hat{\beta}_{1jk}$ is the estimated treatment effect for the jth participant from the kth simulated study, β_{1jk} is the simulated treatment effect for the jth participant from the kth simulated study, n_2 is the number of participants per simulated study, and

$$MS_e = \frac{\sum_{k=1}^{1,000}\sum_{j=1}^{n_2}\left[\left(\hat{\beta}_{1jk} - \beta_{1jk}\right)^2\right]}{1,000 * n_2}. \tag{13}$$

where the symbols are defined as before. Coverage was computed as the proportion of 95% confidence intervals that contained β_{1jk}, and width was computed as the average difference between the upper and lower limits of the 95% confidence intervals.

To verify that the simulation program was generating data consistent with specifications, running the intended models, keeping track of the right values from the results, and correctly summarizing these values, the program was run for a small number of replications. The vectors produced at each stage of data generation were examined for consistency with the specifications, the output data sets generated by calls to the MIXED and REG procedures were examined to ensure that the intended models were being analyzed, and the summary data set in which results were accumulated was examined for accuracy by comparing it back to the output data sets.

RESULTS

Bias

The distribution of bias values for each method of estimating individual treatment effects is illustrated in a boxplot in Figure 2. As was expected, the bias values were close to 0 for the OLS method of estimating individual treatment effects, with an average bias value of 0.01 and a range of values from −0.01 to 0.09. Also as was expected, the empirical Bayes method of estimation led to biased estimates of the individual treatment effects with a mean of −0.24 and values ranging from −0.45 to −0.09. Recall that the value of the average treatment effect was 1.0, thus an average bias estimate of −0.24 represents 24% of the average parameter value, which is substantial.

Variation in bias of the individual treatment effects was explored by modeling bias with the main effects (series length, number of participants, variance in baseline levels, variance in treatment effects, autocorrelation, and

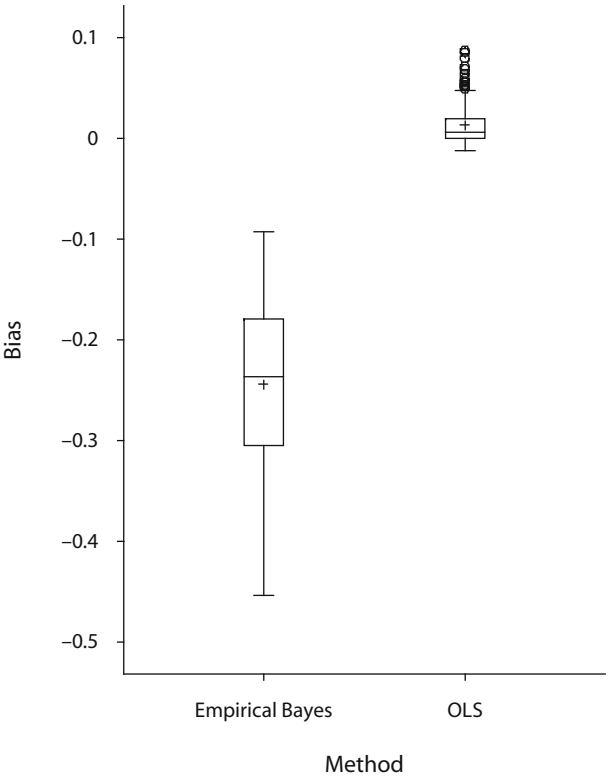

Figure 2. Boxplots showing the distribution of bias estimates for each method of making point estimates of the individual treatment effects.

Table 1
Eta-Squared (η^2) Values for Association of Design Factors With Outcomes

Design Factor	Bias	MS_e	Coverage	Width
Series length	.037	.603	.012	.389
Number of participants	.009	.016	.003	.004
Variance in baseline level	.000	.002	.007	.006
Variance in treatment effect	.006	.000	.009	.003
Autocorrelation	.010	.220	.065	.064
Method	.820	.077	.577	.267
Series length ∗ method	.073	.034	.110	.132
Number of participants ∗ method	.002	.009	.009	.032
Variance in baseline level ∗ method	.000	.002	.003	.002
Variance in treatment effect ∗ method	.013	.000	.022	.009
Autocorrelation ∗ method	.021	.019	.159	.048
Total	.990	.983	.976	.956

Note—For bias and MS_e, the method is empirical Bayes or ordinary least squares (OLS), for coverage and width the method is containment, Satterthwaite, Kenward–Roger, or OLS.

method of estimating the individual treatment effect) and the two-way interactions involving the method used to estimate the individual treatment effects. The proportion of variability associated with each effect is shown in the first column of Table 1. Most of the variability in bias was associated with the method used for estimating the individual treatment effects, followed by the interaction between series length and method and the main effect for series length.

In order to explore these effects further, a line graph was created that modeled the bias values as a function of method, series length, and their interaction and that therefore shows 93% of the variance in bias estimates (Figure 3). Although it is evident that the method of estimation accounts for the majority of the variation in the bias values, as series length gets larger, the bias values for the empirical Bayes estimates become closer to 0. These results are consistent with statistical theory. As series length increases, the empirical Bayes estimates draw more heavily on the data from the individual (and less heavily on the average estimate), thus making the empirical Bayes estimates closer to the OLS estimates.

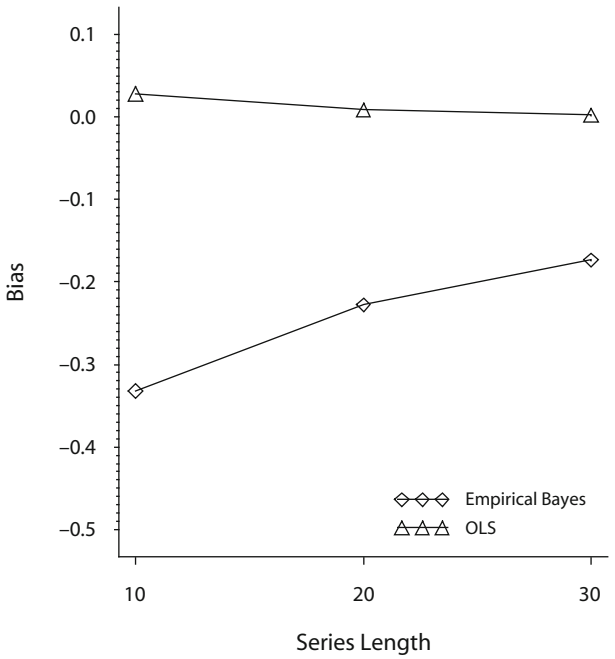

Figure 3. Line graph showing the estimated bias as a function of series length for each method of making point estimates of the individual treatment effects. OLS, ordinary least squares.

MS_e

The distribution of MS_e estimates is shown in Figure 4 for each method of creating the point estimates of the individual effects. The variation in the MS_e estimates that was associated with each design effect is shown in the second column of Table 1. The MS_e decreased with series length, increased with autocorrelation, and—as was theoretically expected—tended to be lower in empirical Bayes estimates than in OLS estimates. The interaction of series length and method can be seen in Figure 5. The difference in MS_e between the empirical Bayes and OLS estimates is greatest when the series length is short and diminishes as the series length increases, which is expected because the empirical Bayes estimates get closer to the OLS estimates as the series length increases.

Although it was expected that the empirical Bayes estimates would be biased and would have a lower MS_e, it was not clear prior to this study how large the bias or difference in MS_e would be for multiple-baseline studies. With the observed bias in the estimate of the individual effect comes the concern that confidence intervals will not be accurate—because these intervals would be created around a biased estimate. The extent to which the confidence intervals are accurate, the primary focus of this study—is examined next.

Confidence Interval Coverage

Figure 6 presents boxplots showing the distribution of coverage estimates (i.e., the proportion of times that the confidence interval contains the true individual treatment effect). The distribution of coverage estimates is shown

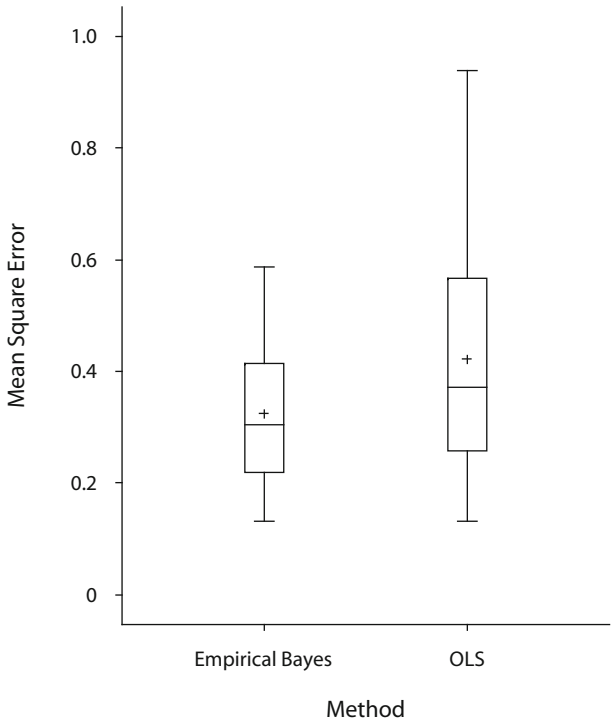

Figure 4. Boxplots showing the distribution of mean square error estimates for each method of making point estimates of the individual treatment effects. OLS, ordinary least squares.

separately for each method of constructing the confidence intervals for the individual treatment effects. With the OLS approach, the coverage estimates range from a high of .959 to a low of .809, with a mean of .899. Variation across conditions was expected for the OLS method, because the intervals can be shown to be accurate when there is no autocorrelation and, as was noted previously, are known to undercover (i.e., coverage less than .95) with positive autocorrelation (Greenwood & Matyas, 1990; Toothaker et al., 1983).

Coverage for the multilevel-modeling methods depended on how the degrees of freedom were computed. Although it was not clear prior to the study how much the method of estimating the degrees of freedom would impact coverage, it was anticipated that if substantial differences emerged, the Kenward–Roger and Satterthwaite methods would have higher coverage estimates than the containment (or default) method. This pattern did emerge, as can be seen in Figure 6. Of particular interest is the accuracy that is obtained when the Kenward–Roger method was used for estimating the degrees of freedom. With this method, the average coverage estimate was very close to the nominal level of .95 (M = .958), and all 180 estimates were relatively close to the nominal level (min = .940, max = .977). Only multilevel modeling with the Kenward–Roger approach to estimating degrees of freedom provided accurate confidence intervals across the range of conditions studied. The Satterthwaite method tended to undercover (M = .920), and the containment method tended to markedly undercover (M = .862).

As is shown in the third column of Table 1, most of the variation in coverage was associated with the method used to construct the confidence interval and its two-way interactions with autocorrelation and series length. To further examine these effects, two line graphs were constructed. One shows the coverage rate as a function of the method used to construct the confidence interval and autocorrelation, whereas the other shows the coverage rate as a function of the method used to construct the confidence interval and series length. These graphs show the vast majority of the variation in the coverage estimates (the combined η^2 was .92 for these factors).

The coverage rates as a function of autocorrelation are shown in Figure 7. The interaction between the method used to construct the confidence interval and autocorrelation can readily be seen. The Kenward–Roger method maintained coverage very close to the nominal level, regardless of the autocorrelation level. However, both the containment and OLS methods had coverage estimates that dropped further below the desired level as the autocorrelation in the generated errors increased. The OLS method had the greatest decrease as autocorrelation increased, moving from .953 to .834, which was expected, since OLS assumes independent errors.

The coverage for each method as a function of series length is shown in Figure 8. As series length increased, the Kenward–Roger method's average coverage went from .963 when the series length was 10 to .953 when the series length was 30. For the Satterthwaite method, the average coverage stayed relatively constant from .920 when the series length was 10 to .921 when the series length was

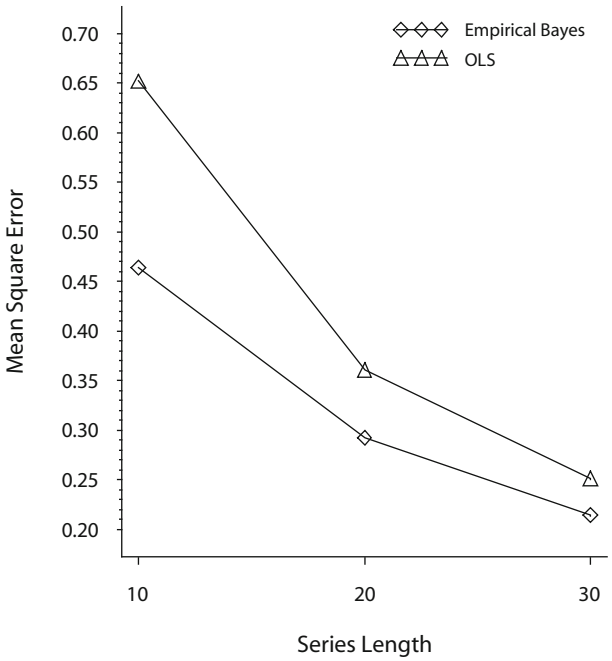

Figure 5. Line graph showing the estimated mean square error as a function of series length for each method of making point estimates of the individual treatment effects. OLS, ordinary least squares.

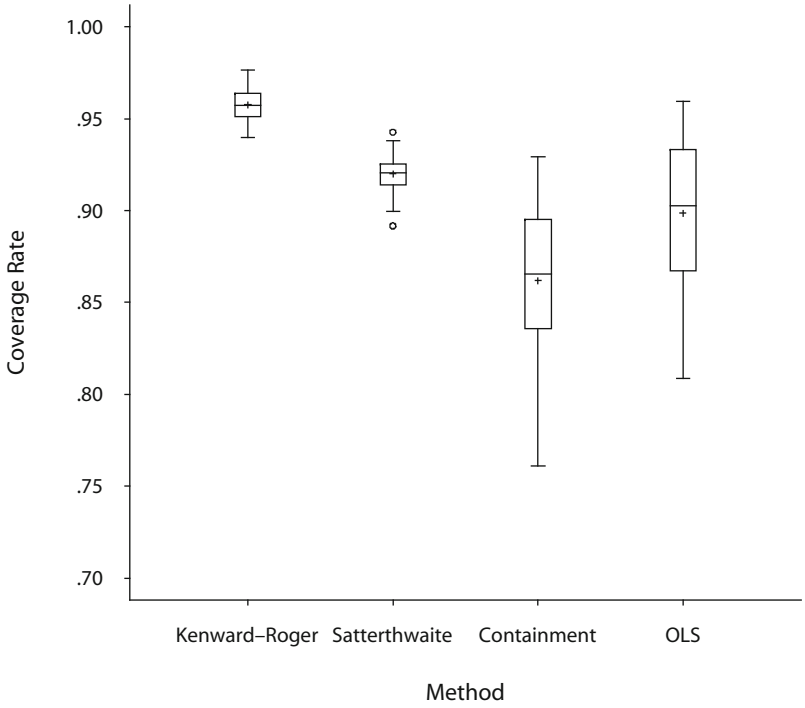

Figure 6. Boxplots showing the distribution of coverage estimates for each method of computing the confidence intervals. OLS, ordinary least squares.

30. The most pronounced effects for series length were seen for the containment method, the coverage of which increased from .820 when the series length was 10 to .894 when the series length was 30. In general, it can be seen that as series length increases, there is less impact of the degrees of freedom method on the coverage. This can be attributed to the increase in degrees of freedom that occurs for all methods as the series length increases. As the degrees of freedom increase, the differences in degrees of freedom will have less impact on coverage.

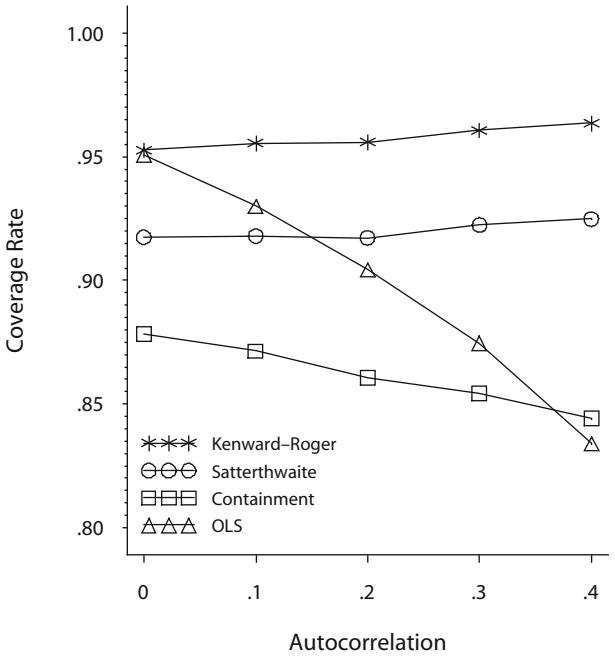

Figure 7. Line graph showing the estimated coverage rate as a function of autocorrelation for each method of computing the confidence intervals. OLS, ordinary least squares.

Confidence Interval Width

The boxplots illustrating the distribution of the interval widths for each method of constructing the confidence intervals for the individual treatment effects are presented in Figure 9. The interval widths were the smallest for the containment method of estimating the degrees of freedom ($M = 1.75$). The Satterthwaite method of estimating the degrees of freedom and the OLS approach were relatively similar ($M = 2.18$ and $M = 2.22$, respectively), and the Kenward–Roger method yielded the largest interval widths ($M = 3.15$). Note that these observations are explainable in light of the coverage results. The methods that provided the smallest confidence intervals were the methods that were least accurate, because the intervals were not wide enough to contain the true treatment effect 95% of the time.

The η^2 for the factors effecting interval width are shown in the last column of Table 1. Most of the variation was related with the series length, followed by the method for constructing the confidence intervals and the interaction between series length and method. In order to explore these effects further, a line graph was created that modeled the interval widths as a function of series length (Figure 10). The graph indicates that although the containment method for estimating the degrees of freedom remains relatively stable regardless of the series length,

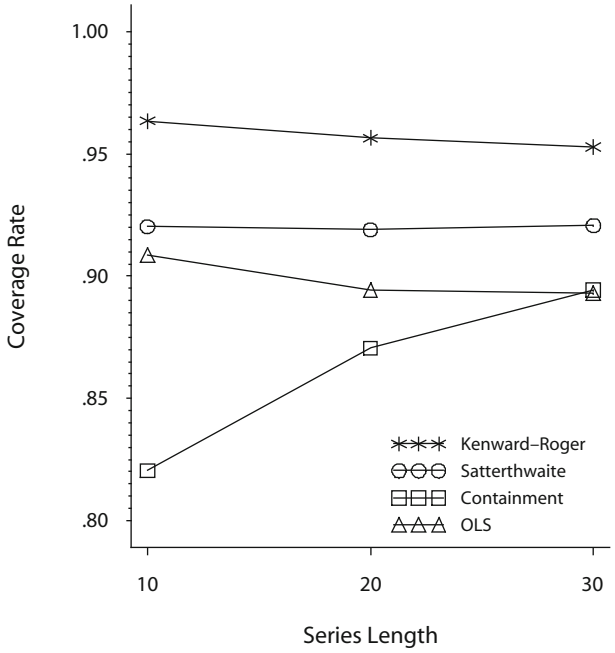

Figure 8. Line graph showing the estimated coverage rate as a function of series length for each method of computing the confidence intervals. OLS, ordinary least squares.

the interval widths decrease slightly for the Satterthwaite method and OLS approach and greatly for the Kenward–Roger method as series length increases. Therefore, the interval widths become more similar as series length increases.

DISCUSSION

This study was focused on a multilevel model in which the treatment effect was conceptualized as the difference in the level of response during intervention and the level of response during baseline. This model was used in the analysis presented by Van den Noortgate and Onghena (2003a) and in several reanalyses of multiple-baseline studies (Ferron et al., 2009). This model is also implied when researchers report the means of each phase and when meta-analyses are done in which the effect size is computed as the standardized difference between the treatment and baseline means (e.g., Busk & Serlin, 1992). Furthermore, rather than looking at the average treatment effects (Ferron et al., 2009), in this study, we expanded the research base by examining effects at the individual level.

Inferences about individual effects are important to practitioners who need to evaluate whether an individual responds to intervention and to researchers who focus on individuals, such as those who study populations with low prevalence rates. In contexts in which it is suspected that the errors may be autocorrelated (such as a multiple-baseline design), multilevel modeling, which can accommodate complex error structures, is preferable to OLS. In this study, we found that when multilevel modeling is used, care must be taken in selecting a method for estimating degrees of freedom. The Kenward–Roger method of estimating the degrees of freedom is preferable to either the Satterthwaite or the containment method when making individual treatment effect inferences from multiple-baseline data, because it provided accurate confidence intervals, regardless of the number of participants, series

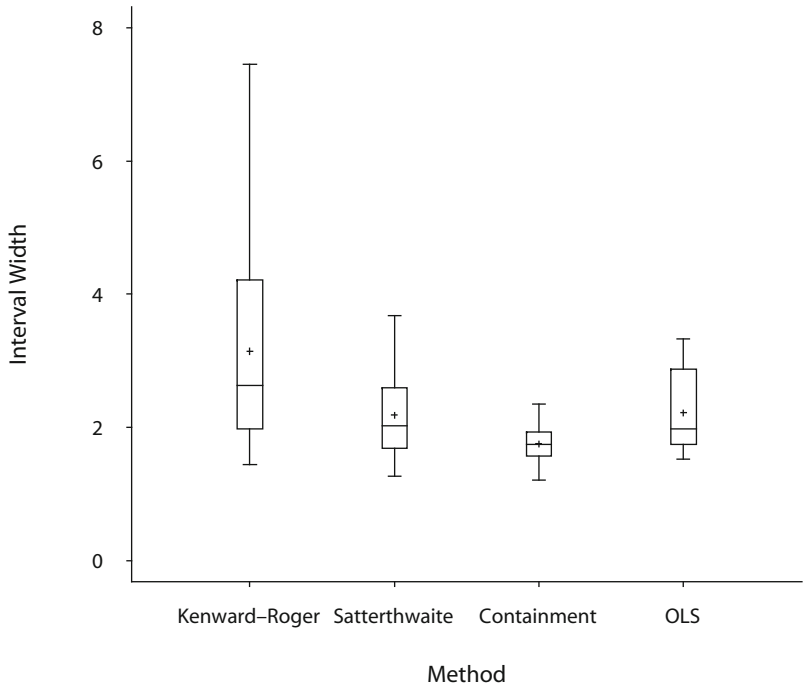

Figure 9. Boxplots showing the distribution of interval width estimates for each method of computing the confidence intervals. OLS, ordinary least squares.

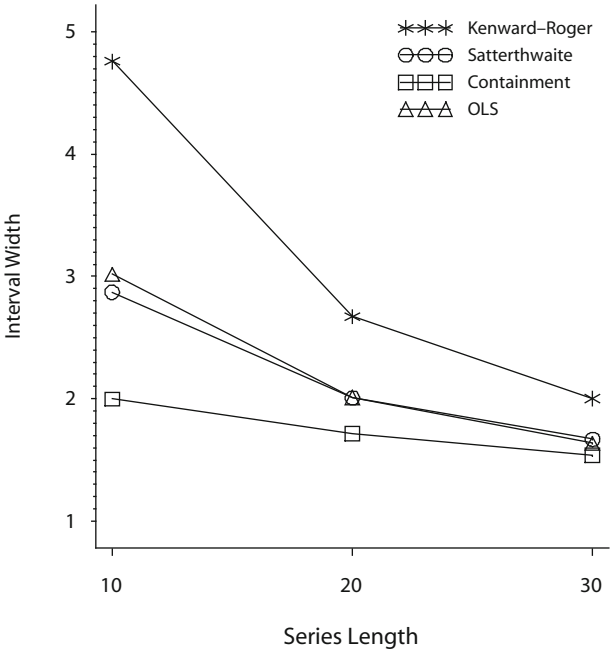

Figure 10. Line graph showing the estimated interval width as a function of series length for each method of computing the confidence intervals. OLS, ordinary least squares.

length, degree of autocorrelation, variance in baseline levels, and variance in treatment effects.

Although this study was focused on a relatively simple multilevel model, there are other applications that may involve more complex treatment effects, such as those that change linearly or nonlinearly with time, those that are delayed, and those that are transitory. These more complex models can also be accommodated within a multilevel-modeling framework, and the theoretical expectations demonstrated here still hold for these more complex models. More specifically, OLS confidence intervals would still be expected to undercover with positive autocorrelation and, of the multilevel-modeling options, the Kenward–Roger method would be expected to perform best when there was a complex covariance structure, even though the degree to which it would perform better has not examined.

In addition, multiple-baseline applications may involve more complex error structures than those that were studied, such as higher order autoregressive or moving average models, heterogeneous error structures, non-normally distributed errors at Level 1 or Level 2, or multivariate error structures for the multiple time series. Similar to more complex treatment effects, more complex error structures can be modeled using multilevel techniques. As the covariance structure becomes more complex, the Kenward–Roger approach maintains a theoretical advantage. The degree to which accurate confidence intervals can also be obtained for these more complex error structures should be explored in future research. Kenward and Roger (2009) proposed a new approximation that has theoretical advantages for models in which the parameterization of the covariance matrix is nonlinear; therefore, future researchers should also consider the functioning of this new approximation for multiple-baseline studies.

Finally, some applications will involve more participants and longer series lengths. In these cases, both statistical theory and the pattern of results obtained from this study lead us to conclude that multilevel modeling with the Kenward–Roger method of estimating degrees of freedom will continue to provide accurate confidence intervals for individual treatment effects. In addition, as series length increases, the width of the confidence intervals will get smaller, thereby giving more precise estimates of the treatment effect. Consider an effect of 4.7 times the baseline standard deviation, which was the average effect size in a review of 150 single-case studies of school-based interventions (Gresham et al., 2004). On the basis of the Kenward–Roger results shown in Figure 9, a series length of 10 would yield a confidence interval from about 2.3–7.1, but a series length of 30 would lead to a confidence interval of about 3.7–5.7, and a series length of greater than 30 would provide an even tighter interval. In summation, results from this study and statistical theory suggest that researchers conducting multiple-baseline studies with multilevel modeling should use the Kenward–Roger method for estimating degrees of freedom.

AUTHOR NOTE

We thank Laura Stapleton for comments on portions of this research that were presented at the 2009 meeting of the American Educational Research Association, and also thank the editor and anonymous reviewers. Correspondence concerning this article should be sent to J. M. Ferron, University of South Florida, 4202 E. Fowler Ave. EDU 105, Tampa, FL 33620-7750 (e-mail: ferron@usf.edu).

REFERENCES

Allsopp, D. H., McHatton, P. A., Ray, S. N. E., & Farmer, J. L. (2010). *Mathematics RTI: A problem-solving approach to creating an effective model*. Horsham, PA: LRP.

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). Needam Heights, MA: Allyn & Bacon.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time series analysis, forecasting, and control*. San Francisco: Holden-Day.

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.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis* (pp. 187-212). Hillsdale, NJ: Erlbaum.

Candel, M. J. J. M., & Winkens, B. (2003). Performance of empirical Bayes estimators of Level-2 random parameters in multilevel analysis: A Monte Carlo study for longitudinal designs. *Journal of Educational & Behavioral Statistics*, **28**, 169-194. doi:10.3102/10769986028002169

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

Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics*, **5**, 235-251. doi:10.2307/1164966

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., BELL, B. A., HESS, M. R., RENDINA-GOBIOFF, G., & HIBBARD, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, **41**, 372-384.

FERRON, J. [M.], 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.], & JONES, P. K. (2002, April). *Visual tests for the analysis of multiple-baseline data*. Paper presented at the annual conference of the American Educational Research Association, New Orleans, LA.

FERRON, J. [M.], & JONES, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *Journal of Experimental Education*, **75**, 66-81. doi:10.3200/JEXE.75.1.66-81

FERRON, J. [M.], & SENTOVICH, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education*, **70**, 165-178.

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

GADBURY, G. L., & IYER, H. K. (2000). Unit-treatment interaction and its practical consequences. *Biometrics*, **56**, 882-885.

GLOVER, T. A., & DIPERNA, J. C. (2007). Service delivery for response to intervention: Core components and directions for future research. *School Psychology Review*, **36**, 526-540.

GOMEZ, E. V., SCHAALJE, G. B., & FELLINGHAM, G. W. (2005). Performance of the Kenward–Roger method when the covariance matrix is selected using AIC and BIC. *Communications in Statistics: Simulation & Computation*, **34**, 377-392.

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.

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

HARVILLE, D. A., & JESKE, D. R. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association*, **87**, 724-731.

HOLLAND, P. W. (1986). Statistical and causal inference. *Journal of the American Statistical Association*, **81**, 945-960.

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

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

KAZDIN, A. E., & KOPEL, S. A. (1975). On resolving ambiguities of the multiple-baseline design: Problems and recommendations. *Behavior Therapy*, **6**, 601-608.

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

KENWARD, M. G., & ROGER, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, **53**, 2583-2595.

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., ALDEN, K., DEMUTH, D., DAWSON, D., PANICUCCI, C., ARNTSON, P., ET AL. (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.

KWOK, O., WEST, S. G., & GREEN, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research*, **42**, 557-592.

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.

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

MURPHY, D. L., & PITUCH, K. A. (2009). The performance of multilevel growth curve models under an autoregressive moving average process. *Journal of Experimental Education*, **77**, 255-282.

NEUMAN, S. B., & MCCORMICK, S. (EDS.) (1995). *Single-subject experimental research: Applications for literacy*. Newark, DE: International Reading Association.

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

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.

PRASAD, N. G. N., & RAO, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163-171.

RAUDENBUSH, S. W., & BRYK, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688-701. doi:10.1037/h0037350

SAS INSTITUTE (2004). *SAS/STAT 9.1 user's guide*. Carey, NC: SAS Institute.

SAS INSTITUTE (2005). *SAS, release 9.12* [Computer program]. Carey, NC: SAS Institute.

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 http://support.sas.com/events/sasglobalforum/previous/index.html.

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.

SHARPLEY, C. (1981). Time series analysis of counseling research. *Measurement & Evaluation in Guidance*, **14**, 149-157.

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.

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*, **8**, 289-309.

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.

VAN DEN NOORTGATE, W., & ONGHENA, P. (2007). The aggregation of single-case results using hierarchical linear models. *Behavior Analyst Today*, **8**, 52-75.

APPENDIX
SAS Programming Lines for Estimating Individual Effects and Their Confidence Intervals

The following MIXED procedure programming lines were used to obtain empirical Bayes estimates of the individual treatment effects and their confidence intervals. Following Van den Noortgate and Onghena (2003a), the dummy coded treatment variable (*phase*) is included in the random statement but not in the model statement. With this specification, the random effects become the empirical Bayes estimates of the treatment effects as opposed to errors that need to be added to the average treatment effect. The Kenward–Roger degrees of freedom are requested by the option `ddfm=kr`. Other degrees of freedom methods were obtained by altering this option.

```
proc mixed data=j2 covtest;
 class person;
 model y = / solution ddfm=kr;
 repeated / type = ar(1) sub=person;
 random intercept phase / sub=person solution cl;
 ods output solutionR = eb1 (keep = estimate lower upper effect);
```

(Manuscript received July 13, 2009;
revision accepted for publication April 3, 2010.)