Behav Res (2012) 44:795–805
DOI 10.3758/s13428-011-0180-y

Synthesizing single-case studies: A Monte Carlo examination of a three-level meta-analytic model

Corina M. Owens · John M. Ferron

Published online: 18 December 2011

Abstract Numerous ways to meta-analyze single-case data have been proposed in the literature; however, consensus has not been reached on the most appropriate method. One method that has been proposed involves multilevel modeling. For this study, we used Monte Carlo methods to examine the appropriateness of Van den Noortgate and Onghena's (2008) raw-data multilevel modeling approach for the meta-analysis of single-case data. Specifically, we examined the fixed effects (e.g., the overall average treatment effect) and the variance components (e.g., the between-person within-study variance in the treatment effect) in a three-level multilevel model (repeated observations nested within individuals, nested within studies). More specifically, bias of the point estimates, confidence interval coverage rates, and interval widths were examined as a function of the number of primary studies per meta-analysis, the modal number of participants per primary study, the modal series length per primary study, the level of autocorrelation, and the variances of the error terms. The degree to which the findings of this study are supportive of using Van den Noortgate and Onghena's (2008) raw-data multilevel modeling approach to meta-analyzing single-case data depends on the particular parameter of interest. Estimates of the average treatment effect tended to be unbiased and produced confidence intervals that tended to overcover, but did come close to the nominal level as Level-3 sample size increased. Conversely, estimates of the variance in the treatment effect tended to be biased, and the confidence intervals for those estimates were inaccurate.

Keywords Single-subject · Research synthesis · Multilevel modeling · Hierarchical linear modeling · Simulation

C. M. Owens · J. M. Ferron
Department of Educational Measurement and Research,
University of South Florida,
Tampa, FL, USA

C. M. Owens (✉)
Battelle Centers for Public Health Research and Evaluation,
2111 Wilson Boulevard, Suite 900,
Arlington, VA 22201, USA
e-mail: owensc@battelle.org

Quantitative integration of study results, termed *meta-analysis*, involves the combining of data across multiple studies to evaluate and summarize research findings. The term *meta-analysis* was first coined by Glass (1976) and was defined as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (p. 3). This type of research is an important way to determine the relationships among variables and the effectiveness of interventions across studies. It also allows researchers to integrate study findings with the goal of generalization. Quantitative integration of study findings should cross research domains and include all types of quantitative research, including single-case research. However, meta-analysis of single-case research has resulted in much disagreement in the field.

Although the use of single-case designs has grown over the past decades, the majority of the literature on meta-analysis has focused on group comparison studies and has left out single-case research (Van den Noortgate & Onghena, 2008). This lack of a literature related to single-case designs is often the reason that these designs are excluded from meta-analyses. This exclusion of single-case designs is of concern when one considers the plethora of information that single-case research can add to the literature. Single-case designs not only provide information related to average treatment effects, but also offer

information related to how that treatment effect is related to specific cases. Meta-analyses of single-case designs offer the ability to summarize and evaluate the overall effect without the loss of that specific case information. In addition, the meta-analysis of single-case data would increase the generalizability of research findings.

Researchers have proposed a variety of methods to meta-analyze single-case data (Allison & Gorman, 1993; Busk & Serlin, 1992; Center, Skiba, & Casey, 1985–1986; Gingerich, 1984; Onghena & Edgington, 2005; Scruggs, Mastropieri, & Castro, 1987; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008). Van den Noortgate and Onghena (2003a, 2003b, 2007, 2008) proposed the use of multilevel modeling to aggregate single-case data for the purposes of meta-analysis. The authors suggested aggregating single-case data in three different ways. The first option includes individual-level raw data from each primary study in the meta-analysis and makes the assumption that all of the dependent variables across studies are measured in the same way. Van den Noortgate and Onghena (2008) illustrated this first option in a series of models, provided in Eqs. 1–5 below.

Equation 1 represents within-person variation, which can be modeled with a basic regression equation. Specifically, an outcome (y) is modeled on measurement occasion i for participant j in study k (y_{ijk}) as a linear function of a single predictor, *phase*:

$$y_{ijk} = \pi_{0jk} + \pi_{1jk} phase + e_{ijk}, \tag{1}$$

where *phase* represents a dummy-coded variable indicating whether measurement occasion i took place during the baseline (0) or treatment (1) phase. π_{0jk} is the level of the outcome during baseline for participant j from study k; π_{1jk} is the treatment effect for participant j from study k; and e_{ijk} is the within-phase error variance.

At the second level, the variation across participants is modeled in the following equations:

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \tag{2}$$

and

$$\pi_{1jk} = \beta_{10k} + r_{1jk}, \tag{3}$$

where β_{00k} is the average baseline level for study k and β_{10k} is the average treatment effect for study k, and the error terms r_{0jk} and r_{1jk} allow for variation in both baseline levels and treatment effects among the participants within study k.

At the third level, the variation across studies is modeled in the following equations:

$$\beta_{00k} = \gamma_{000} + u_{00k} \tag{4}$$

and

$$\beta_{10k} = \gamma_{100} + u_{10k}, \tag{5}$$

where γ_{000} is the overall average baseline level and γ_{100} the overall average treatment effect, and the error terms u_{00k} and u_{10k} allow for variation in both the average baseline levels and the average treatment effects among studies. It should be noted that errors on all levels were assumed to be independently normally distributed and to have a mean of zero. However, multilevel models are quite flexible, and the use of a complex covariance structure, such as a first-order autoregressive structure, is possible to account for dependent errors.

Van den Noortgate and Onghena (2008) second option assumes that the dependent variable is measured differently across studies, and therefore that scores from individuals need to be standardized before combining them into one analysis. First, the individual-level raw data are standardized by performing an ordinary least squares (OLS) regression for each participant separately, dividing their scores by each resulting root-mean squared error, and then combining the data into the models defined in Eqs. 1–5 (Van den Noortgate & Onghena, 2008).

The third option proposed by Van den Noortgate and Onghena (2008) does not include individual-level data from each study in the meta-analysis. Instead, standardized regression coefficients are calculated for each study and included in the meta-analysis as effect sizes representing a standardized change in level and change in slope. In this option, Eq. 1 needs slight modifications to appropriately meta-analyze single-case data. The first level of the model is adapted to model the effect sizes or standardized regression coefficients from each study rather than the individual-level data:

$$\hat{\pi}_{0jk} = \pi_{0jk} + e_{jk}, \tag{6}$$

with $\hat{\pi}_{0jk}$ representing the observed effect size for participant j in study k, modeled as the true effect size (π_{0jk}) for participant j in study k plus some random variation or error (e_{jk}), where the Level-1 error variance matrix is assumed to be known. The second- and third-level equations (see Eqs. 2–5), describing the variation across participants and between studies, remain the same.

Multilevel modeling provides estimates of the (co) variance at each level, but typically it only estimates fixed-effect parameters at the highest level. Therefore, variance and covariance estimates across all levels, as well as the fixed effects at the third level—the average baseline across studies and the average treatment effect across studies—can be reported. These types of parameter estimates offer the ability not only to provide information on the overall treatment effect, but also information related to the variability of that overall average treatment effect. In addition, predictors can be added to the model to account for that variability.

Van den Noortgate and Onghena (2008) argued that single-case study conclusions are restricted to the participants that were investigated, but multilevel modeling provides the ability to combine results from multiple participants and studies to gain information about not only the average treatment effect, but also whether and how the treatment effect varies across participants and studies. Another advantage of multilevel modeling is that it can be used to aggregate data from single-case studies that include multiple participants. This use of multiple data sources or effect sizes from the same study is typically problematic and has not been addressed by other proposed single-case meta-analytic methods. Multilevel modeling is structured to account for that "nesting" of data within studies by allowing for variation within participants, between participants of the same study, and between studies (Van den Noortgate & Onghena, 2008).

Although several advantages of this approach exist, concerns can be raised when multilevel models are used with small samples. Although the fixed-effect estimates have been shown to be unbiased in small samples, the standard errors for these effects become questionable, as do the estimates of the variance components. More specifically, if the variance parameters are estimated through restricted maximum likelihood (REML) and these variance estimates are then used to solve the mixed-model equations, the resulting fixed-effect estimates are unbiased (Kackar & Harville, 1984; Kenward & Roger, 1997). This analytic result does not depend on appeals to large-sample theory. The standard errors for the fixed effects, however, tend to be underestimated because the sampling error in the variance estimates is not taken into account. As a consequence, variance inflation factors have been proposed (Kackar & Harville, 1984), as well as methods that approximate the degrees of freedom based on the inflated variances (Kenward & Roger, 1997). Although these approaches are recommended for making multilevel inferences with small samples, it should be noted that the exact sampling distribution for the fixed effects has not been derived, and thus, these adjustments are approximations. Furthermore, what we know from statistical theory about the variance estimates from REML estimation of multilevel models is based on large-sample theory, and questions have thus been posed about the accuracy of the variance estimates and the validity of the variance inferences when sample sizes are small (Raudenbush & Bryk, 2002).

Given these concerns with small sample size, several simulation studies have been conducted to examine two-level models of single-case data (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Ferron, Farmer, & Owens, 2010; Ferron, Owens, & Bell, 2010). The results have shown that unbiased estimates of fixed effects (e.g., average treatment effects) are obtained. In addition, accurate confidence intervals for fixed effects and for estimates of individual treatments were obtained under a variety of series length and sample size conditions. This accuracy depended on the method used to estimate the degrees of freedom, with the best performance being associated with the Kenward–Roger approach. Variance components tended to be biased, however, particularly the estimates of between-person variance (e.g., variance in the treatment effect). These findings provide some empirical support for using multilevel models with single-case data and some motivation for this study, which aims to empirically evaluate the utility of inferences made from a three-level model to meta-analyze single-case data.

Purpose

The purpose of this study was to examine the appropriateness of Van den Noortgate and Onghena's (2008) raw-data multilevel modeling approach (their first option) to the meta-analysis of single-case data. Specifically, we examined the fixed effects (i.e., the overall average baseline level and the overall average treatment effect) and the variance components (e.g., the between-person within-study variance in the average baseline level, the between-study variance in the overall average baseline level, and the between-person within-study variance in the average treatment effect) in a three-level multilevel model. More specifically, the bias of point estimates, the confidence interval coverage rates, and the confidence interval widths were examined as a function of specific design and data factors.

Method

Monte Carlo simulation methods were used to examine the appropriateness of the inferences of multilevel modeling. The use of simulation methods allowed for the control and manipulation of specific design and data factors. The Monte Carlo study included five factors in the design, which were (a) the number of primary studies per meta-analysis (10, 30, and 80); (b) the modal number of participants per primary study (small [mode = 4] and large [mode = 8]); (c) the modal series length per primary study (small [mode = 10], medium [mode = 20], and large [mode = 30]); (d) the level of autocorrelation (0, .2, and .4); and (e) the variances of the error terms (most of the variance at Level 1 [$\sigma^2 = 1$, $\tau_{\pi 00} = \tau_{\pi 10} = 0.2$, and $\tau_{\beta 00} = \tau_{\beta 10} = .05$], as well as most of the variance at Level 2 [$\sigma^2 = 1$, $\tau_{\pi 00} = \tau_{\pi 10} = 2$, and $\tau_{\beta 00} = \tau_{\beta 10} = .5$]). The values for these factors were chosen to be reflective of the range of values seen in single-case meta-analyses and to be consistent with previous simulation research on single-case studies (Ferron et al., 2009; Ferron, Farmer, & Owens, 2010; Ferron, Owens, & Bell, 2010). The data for

this study were generated on the basis of Van den Noortgate and Onghena's (2008) raw-data, three-level, meta-analytic single-case model shown in Eqs. 1–5. Each data set was analyzed using the same model that was used for data generation (see Eqs. 1–5). The three-level model was estimated using REML via the PROC MIXED method with the Kenward–Roger degrees of freedom in SAS version 9.2 (SAS Institute, Inc., 2008). In addition, a first-order autoregressive model for the Level-1 errors was specified. On the basis of the current model, the treatment effect was modeled as a change in level, and estimates were obtained for autocorrelation, variance within participants, variance in baseline levels across participants and studies, and variance in treatment effects across participants and studies. Errors for the within-participant model (e_{ijk}) were generated using the ARMASIM function in SAS version 9.2 with a variance (σ^2) of 1.0 and autocorrelation values of 0, .2, or .4, as previously discussed. Level-2 errors were generated from a normal distribution using the RANNOR random number generator in SAS version 9.2. The variances of the Level-2 errors were defined on the basis of the previously discussed levels of 0.2 and 2, and the covariance between r_{0jk} and r_{1jk} was set to 0. The covariance between these Level-2 errors was set to zero in keeping with past simulation research (Ferron et al., 2009; Ferron, Farmer, & Owens, 2010; Ferron, Owens, & Bell, 2010), as well as with Van den Noortgate and Onghena's (2003a, 2007) application of multilevel modeling to single-case data. Level-3 errors were generated from a normal distribution using the RANNOR random number generator in SAS version 9.2. The fixed effects (γ_{000} and γ_{100}) were set to 1.0. The variances of the Level-3 errors were defined on the basis of the previously discussed levels of .05 and .5, and the covariance between u_{00k} and u_{10k} was set to 0. The covariance between these Level-3 errors was set to zero in keeping with Van den Noortgate and Onghena's (2007) application of multilevel modeling to single-case data.

The estimated models were checked for consistency with the data generation. Several checks were used to verify the accuracy of the simulation program by running the program for a small number of replications. The vectors created during data generation were examined for consistency with the data specifications, and output data sets from the PROC MIXED statements were examined to ensure that the intended models were being analyzed.

The appropriateness of Van den Noortgate and Onghena's (2008) raw-data multilevel modeling approach to the meta-analysis of single-case data was evaluated by examining the bias and/or relative bias of the point estimates, the confidence interval coverage, and the confidence interval width of both the fixed effects and the variance components. This was accomplished by creating box plots, across all conditions, for each dependent variable. Then, the results of the simulation were analyzed using PROC GLM in SAS version 9.2 for both the fixed effects and the variance components, such that the dependent variables were bias, relative bias (where appropriate), confidence interval coverage, and confidence interval width, and the independent variables were the five factors. Models were built with the purpose of finding effects whose eta-squared (η^2) values were .06 or greater. The effect size, η^2, was calculated in order to determine the proportion of variability associated with each effect. Those values were compared to Cohen's (1988) standards for interpreting η^2 values, with a small effect size being $\eta^2 = .01$, a medium effect size being $\eta^2 = .06$, and a large effect size being $\eta^2 = .14$ or greater. Each model was first created as a main-effects-only model. If this model explained more than 94% of the total variability, no further complex models were then investigated, because none of the interaction effects could have an η^2 of .06. However, if less than 94% of the total variability was explained, interactions were included in the model. Two-way interactions were added to the model first, followed by three-way, and then four-way interactions, until at least 94% of the variability was explained.

Results

Fixed effects

Bias The extent to which the fixed effects from a three-level meta-analytic single-case model were biased, as a function of the specific design factors, was evaluated via the average amount that the estimated parameter differed from the true parameter. The results indicated that, regardless of condition, the fixed effects were unbiased, with average bias values of zero. The unbiased fixed-effect estimates revealed in this research are consistent with theoretical expectations (Kackar & Harville, 1984) and previous research regarding the utility of the inferences made from fixed effects in two-level models (Ferron et al., 2009; Raudenbush & Bryk, 2002). Therefore, the use of fixed effects from a three-level meta-analytic single-case model would be expected to provide unbiased estimates of the average baseline level and average treatment effect across studies, if the model was correctly specified.

Confidence interval coverage The proportions of the 95% confidence intervals that contained the parameter value were estimates of the confidence interval coverage of the fixed effects from a three-level meta-analytic single-case model. The confidence interval coverage rates of the fixed effects—both the overall average baseline level and the overall average treatment effect—tended to overcover, with means of .961 and

.960, respectively (see Fig. 1). Further examination of the extent to which the fixed effects varied as a function of the specific design factors illustrated that the 95% confidence interval coverage rates of the fixed effects came close to a .95 coverage rate as the Level-3 sample size increased (see Fig. 2). These findings suggest that whenever possible, researchers should increase the Level-3 sample size or number of primary studies included in the meta-analysis. In addition, these findings validate findings in the previous literature related to two-level models for single-case data, in which larger numbers of upper-level units were reported as leading to greater accuracy and precision (Ferron et al., 2009).

These findings are also consistent with general methodological research on more traditional designs of repeated measurements using multilevel models and the Kenward–Roger degrees of freedom (Fouladi & Shieh, 2004; Gomez, Schaalje, & Fellingham, 2005; Kenward & Roger, 1997; Kowalchuk, Keselman, Algina, & Wolfinger, 2004; Schaalje, McBride, & Fellingham, 2001). These previous simulation studies indicated that, across a variety of conditions and sample sizes, Type I error rates have been close to the nominal α level, but variability in performance was noted. For example, Gomez, Schaalje, and Fellingham (2005) examined a three-group design with 3 participants per group and each participant measured at three points in time, and they found that Type I error control varied with the type of covariance structure. In particular, their results indicated that when the data were generated and analyzed assuming compound symmetry, the estimated Type I error rate was .052 (α = .05). However, when the data were generated and analyzed assuming a first-order autoregressive-with-random-effects model, the estimated Type I error rate was .1165 (α = .05).

Confidence interval width The average difference between the upper and lower limits of the 95% confidence intervals defined the confidence interval widths of the fixed effects from a three-level meta-analytic single-case model. The confidence interval widths of the fixed effects—both the overall average baseline level and the overall average treatment—were relatively small, with average confidence interval width estimates of 0.428 and 0.459, respectively (see Fig. 3). To gain a better understanding for widths of this size, it is helpful to recall that the Level-1 variance was set to 1.0 and that both fixed effects were set to 1.0. Therefore, average confidence interval widths of 0.459 for the overall average treatment effect would produce interval estimates that ranged from around 0.770 to 1.230.

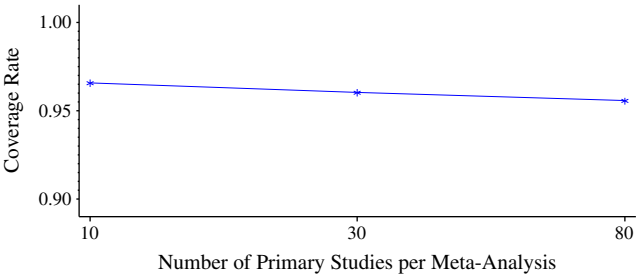

Fig. 2 Line graph showing the estimated confidence interval coverage rates for the overall average treatment effect as a function of the number of primary studies per meta-analysis

Further examination of the extent to which the confidence interval widths of the fixed effects varied as a function of the specific design factors indicated that the interaction between the Level-3 sample size and the variances of the error terms impacted the variability in confidence interval widths of the fixed effects (see Fig. 4). Specifically, confidence interval widths of the fixed effects were smallest when the Level-3 sample sizes were largest (mode = 80) and most of the variance in the error terms was at Level 1, or less variance was at Levels 2 and 3. This finding is similar to previous research examining two-level models for single-case data (Ferron et al., 2009; Ferron, Owens, & Bell, 2010), which found that confidence interval widths of the treatment effect decreased with more participants, more observations per participant, and smaller variance components. The present study's results suggest that a larger number of upper-level units and less variability between persons and studies would produce more precise confidence intervals of the fixed effects.

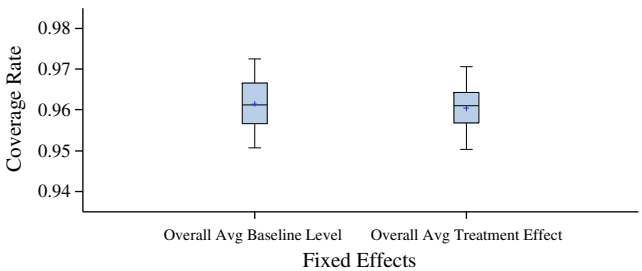

Fig. 1 Box plots showing the distribution of confidence interval coverage rates for each fixed effect in the three-level model

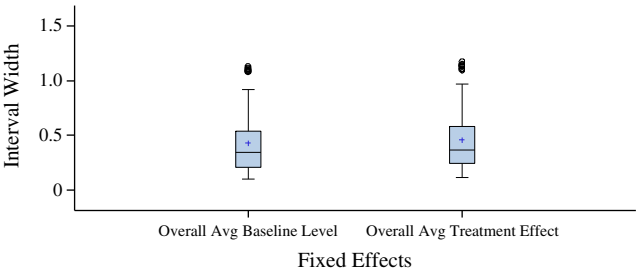

Fig. 3 Box plots showing the distribution of confidence interval width estimates for each fixed effect in the three-level model

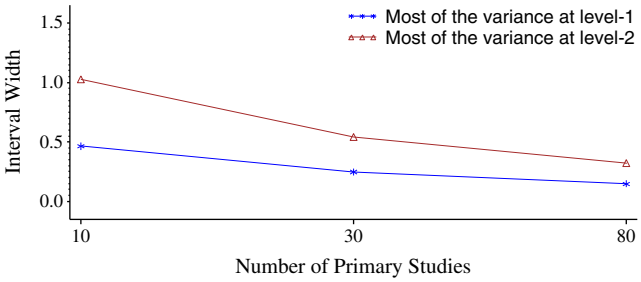

Fig. 4 Line graph showing the estimated confidence interval widths of the overall average treatment effect as a function of the variances of the error terms at each level of the number of primary studies per meta-analysis

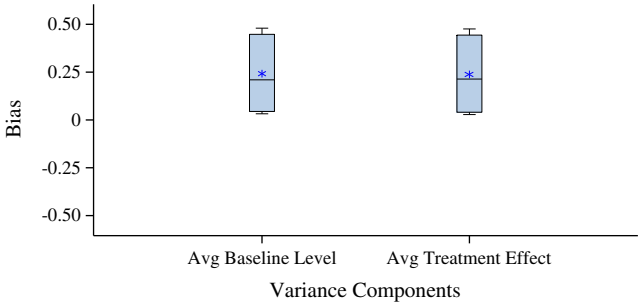

Fig. 6 Box plots showing the distribution of bias estimates for each Level-2 variance component in the three-level model

Variance components

Bias The extent to which the variance components from a three-level single-case meta-analytic model were biased, as a function of the specific design factors, was evaluated via the average amount that the estimated parameter differed from the known parameter. As expected, the Level-3 and Level-2 variance components tended to be biased. Specifically, the Level-3 variance components, of the between-study variances in both the overall average baseline level and the overall average treatment effect, tended to be underestimated, with mean bias values of –0.241 and –0.237, respectively (see Fig. 5). The Level-2 variance components, of the between-person within-study variances in both the average baseline level and the average treatment effect, tended to be overestimated, with mean bias estimates of 0.243 and 0.238, respectively (see Fig. 6). These findings are not too surprising, given other research from a broader methodological perspective. Previous Monte Carlo research on growth curve models, in studies having as few as 30 participants and a series length of 4 or 8 (Kwok, West, & Green, 2007) and a series length of 5 or 8 (Murphy & Pituch, 2009), has reported substantial bias in the variance components when the model was correctly specified and the number of participants was small ($N = 30$).

In the present study, bias in the Level-3 variance components was mainly impacted by one factor, the variances of the error terms. As the variances of the error terms shifted from most of the variance at Level 1 to most of the variance at Level 2, the Level-3 variance components tended to become increasingly underestimated and progressively more biased. Conversely, the Level-2 variance components became increasingly overestimated and progressively more biased as the variances in the error terms shifted from most of the variance at Level 1 to most of the variance at Level 2.

As in previous research on two-level models with single-case data (Ferron et al., 2009), Level-1 variance or within-person residual variance was slightly biased, but unlike in that previous research, the bias in the estimates of within-person residual variance remained constant at around 8%, regardless of the Level-3 or Level-2 sample size (see Fig. 7). However, results from this study did reveal that the within-person residual variance of the three-level model became increasingly biased as the level of autocorrelation increased. As autocorrelation increases, the errors between observations within a person become more similar, and therefore it seems reasonable that it would be more difficult to produce an unbiased estimate of the within-person variability. However, the estimate of the autocorrelation was on average unbiased. Both the within-person residual variance and the autocorrelation bias results were not consistent with the previous literature on two-level models, which found both parameters to be substantially biased (Ferron et al., 2009). However, the present study did focus on a three-level model rather than the previously investigated two-level model, and

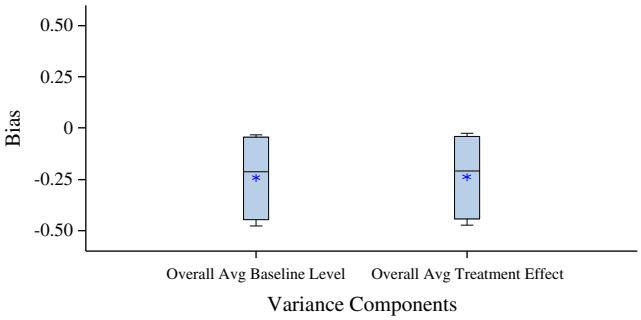

Fig. 5 Box plots showing the distribution of bias estimates for each Level-3 variance component in the three-level model

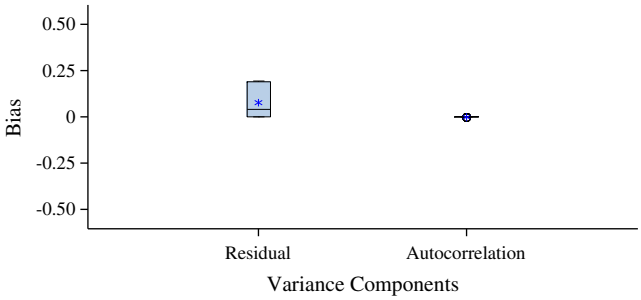

Fig. 7 Box plots showing the distributions of bias estimates for the within-person residual variance and the amount of estimated autocorrelation in the three-level model

therefore more information was ultimately available in the estimation of those parameters.

Confidence interval coverage The extent to which the confidence interval coverage estimates of the variance components from a three-level meta-analytic single-case model produced accurate confidence intervals, as a function of the specific design factors, was estimated by the proportion of the 95% confidence intervals that contained the parameter value. The Level-3 variance components, of the between-study variances in both the overall average baseline level and the overall average treatment effect, tended to overcover, with means of .998 and .995, respectively (see Fig. 8).

Further examination of these effects indicated that the main factors that influenced the variability in confidence interval coverage rates of the Level-3 variance components were the Level-3 sample size, the Level-2 sample size, and the variances of the error terms. Specifically, the confidence interval coverage rates of the Level-3 variance components were closest to a .95 coverage rate when the Level-3 sample size was smallest (10 primary studies), Level-2 sample size was smallest (mode = 10), and most of the variance in the error terms was at Level 1 (see Fig. 9). Recall that the bias of the Level-3 variance components was smallest when the Level-3 sample size was smallest, Level-2 sample size was smallest, and most of the variance in the error terms was at Level 1. Therefore, given the relative bias results, it is not surprising that the confidence interval coverage was problematic for the Level-3 variance components.

Similar results were found for the Level-2 variance components and the within-person residual variance. The Level-2 variance components, of the between-person within-study variances in both the average baseline level and the average treatment effect, tended to undercover, with means of .612 and .675, respectively (see Fig. 10). Several design factors were found to have impacted the variability in the confidence interval coverage rates of the Level-2 variance components. The confidence interval coverage rates of the between-person

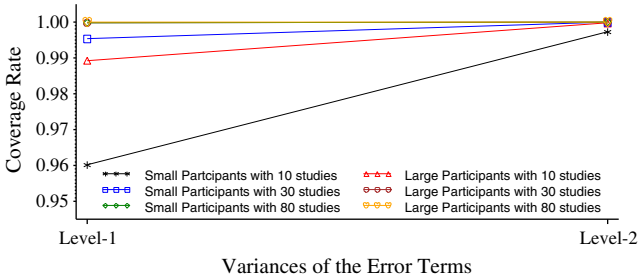

Fig. 9 Line graph showing the estimated confidence interval coverage rates for the between-study variance in the overall average treatment effect as a function of the three-way interaction between number of primary studies per meta-analysis, modal number of participants per primary study, and the variances of the error terms

within-study variance in the average baseline level tended to decrease and move farther away from a .95 coverage rate when the Level-3 sample size increased, the Level-2 sample size increased, and the variances of the error terms shifted from most of the variance at Level 1 to most of the variance at Level 2. In addition, the confidence interval coverage rates of the other Level-2 variance component, the between-person within-study variance in the average treatment effect, tended to decrease and move farther away from a .95 coverage rate as the Level-3 sample size increased and the variances of the error terms shifted from most of the variance at Level 1 to most of the variance at Level 2 (see Fig. 11). Recall that the relative bias results of the Level-2 variance components indicated that estimates of the Level-2 variance components became more biased as the Level-3 sample size increased, the Level-2 sample size increased, and the variances of the error terms shifted from most of the variance at Level 1 to most of the variance at Level 2. Therefore, it was not surprising that the confidence interval coverage of the Level-2 variance components was troublesome. Additionally, these results are consistent with previous findings (Maas & Hox, 2004) from a broader methodological perspective on two-level organizational models, which found that the coverage rates of the Level-2 variance components tended to undercover with small sample sizes ($N = 30$).

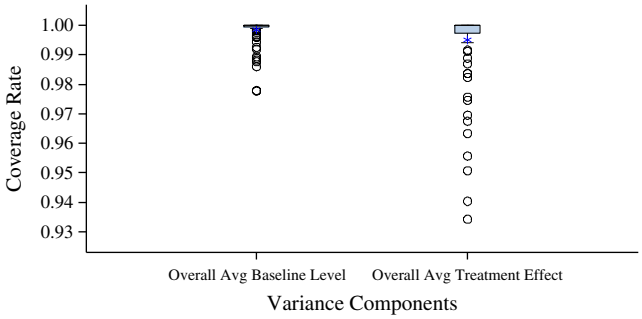

Fig. 8 Box plots showing the distributions of confidence interval coverage rates for the Level-3 variance components in the three-level model

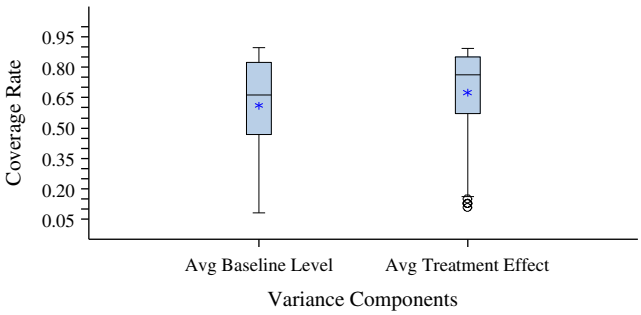

Fig. 10 Box plots showing the distributions of confidence interval coverage rates for the Level-2 variance components in the three-level model

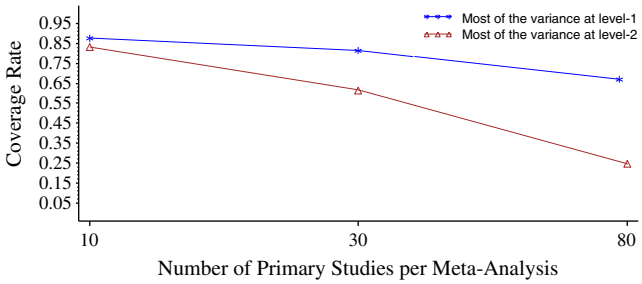

Fig. 11 Line graph showing the estimated coverage rates for the between-person within-study variances in the average treatment effect as a function of the variance of the error terms at each level of the number of primary studies per meta-analysis

Confidence interval coverage rates were the most problematic for the within-person residual variance, with average confidence interval coverage rates well below the nominal level (M = .550; see Fig. 12). However, confidence interval coverage rates of the within-person residual variance were close to a .95 coverage rate when autocorrelation was zero. This finding was consistent with the earlier results, given the bias results for the within-person residual variance. Conversely, confidence interval coverage rates for the amount of estimated autocorrelation were close to a .95 coverage rate (M = .950) regardless of condition, which is not surprising given the bias results for the amount of estimated autocorrelation.

Confidence interval width The average difference between the upper and lower limits of the 95% confidence intervals defined the confidence interval widths of the variance components from a three-level meta-analytic single-case model. Confidence interval widths for the Level-3 and Level-2 variance components were so large that they provided no information. These findings are not surprising, given previous research on two-level models for single-case data (Ferron et al., 2009), where the results indicated that the confidence interval widths for the Level-2 variance components were so large that they provided no information.

However, the confidence interval width estimates for the within-person residual variance produced relatively small interval widths (M = 0.146; see Fig. 13), which tended to become even smaller as the Level-3 and Level-1 sample size and the level of autocorrelation increased. For example, consider the fact that the within-person residual variance was set to 1.0; therefore, a small series length, with a mode of 10, would yield a confidence interval from about .914 to 1.086, but a medium series length, with a mode of 20, would produce a confidence interval from .929 to 1.072, and a large series length, with a mode of 30, would provide an even tighter confidence interval, from .939 to 1.061. These results are not too surprising, considering the previously reported tendency of these confidence intervals to undercover.

Likewise, the confidence interval width estimates for the amount of estimated autocorrelation were small (M = 0.090; see Fig. 13) and tended to decrease as the Level-3 and Level-1 sample size increased. Therefore, from the results of the confidence interval widths of the amount of estimated autocorrelation, when the level of autocorrelation was set to zero, a Level-3 sample size of 10 primary studies would produce a confidence interval width from about –.070 to .070, but a Level-3 sample size of 30 primary studies would lead to a confidence interval from around –.040 to .040, and a Level-3 sample size of 80 would yield even greater precision, with a confidence interval from about –.025 to .025. These findings suggest that it is tenable to assume that as the Level-3 and Level-1 sample sizes increase, the estimates of the amount of estimated autocorrelation will become more precise if the model is correctly specified.

Limitations of the study

Because of the design of this study, some generalizability limitations should be considered with regard to this research. The Monte Carlo method used in the study provided control of specific factors to investigate the appropriateness of inferences made from a three-level meta-analytic single-case model in

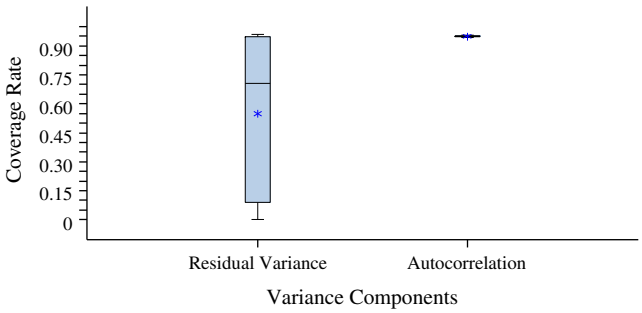

Fig. 12 Box plots showing the distributions of confidence interval coverage rates for the within-person residual variance and the amount of estimated autocorrelation in the three-level model

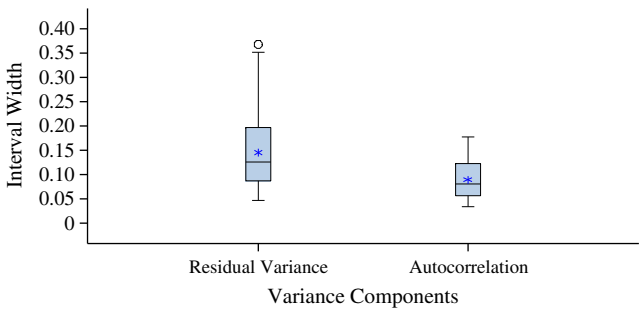

Fig. 13 Box plots showing the distributions of confidence interval width estimates for the within-person residual variance and the amount of estimated autocorrelation in the three-level model

specific situations. While this is a benefit of simulation studies, it also limits the generalizability of the study findings. Therefore, the five design factors (number of primary studies per meta-analysis, modal number of participants per primary study, modal series length per primary study, variances of the error terms, and level of autocorrelation) determine the types of single-case meta-analyses to which the study's findings can be generalized. In addition, another generalizability limitation of this study is the levels of the specific design factors. These levels were chosen to represent a range of possible values seen in single-case meta-analyses as well as in previous simulation work. However, they are not exhaustive of all possible values for each design factor.

Another limitation to consider relates to the model under investigation. The specific model (see Eqs. 1–5) chosen for investigation in this research study makes several assumptions. First, Van den Noortgate and Onghena's (2008) raw-data three-level meta-analytic single-case model assumes that all of the dependent variables were measured in the same way across the primary studies included in the meta-analysis. Second, the model chosen for this analysis was the most basic interrupted time-series model (e.g., no trends or changes in trends). The benefit of choosing this model was that it is the most basic model, and therefore the most logical for an initial study into the three-level meta-analytic modeling of single-case data. In addition, the model and data generation assumed normality of the Level-1 errors, multivariate normality of the Level-2 errors, multivariate normality of the Level-3 errors, and homoscedasticity of the errors at all levels. If the within-person variance varied across the participants either within or across studies, if the autocorrelation varied, or if a more complex time series model (e.g., a second-order or higher autoregressive model) was needed, then the model would be misspecified. The results do not allow for generalizations to performance when there is some degree of misspecification or there is use of more complex model specifications.

Implications

Researchers have suggested that the use of multilevel modeling in meta-analyzing single-case data provides many advantages (Van den Noortgate & Onghena, 2003a, 2007, 2008). Specifically, multilevel modeling provides the ability to combine the results from multiple participants and studies in order to gain information not only about the overall treatment effect, but also about whether and how the treatment effect varies across participants and studies (Van den Noortgate & Onghena, 2008). Another advantage of multilevel modeling is that it can be used to aggregate data from single-case studies that include multiple participants. This use of multiple data sources or effect sizes from the same study is typically problematic and has not been addressed by other proposed single-case meta-analytic methods. Multilevel modeling is structured to account for that "nesting" of data within studies by allowing for variation within participants, between participants in the same study, and between studies (Van den Noortgate & Onghena, 2008).

The degree to which the findings of this study are supportive of using Van den Noortgate and Onghena's (2008) raw-data multilevel modeling approach to meta-analyzing single-case data depends on the particular effect of interest. This in turn leads to specific implications for those conducting meta-analyses of single-case studies, for single-case researchers, and for methodologists.

Implications for researchers conducting single-case meta-analyses

For researchers interested in the overall average baseline level and overall average treatment effect across studies, the results of this research study are encouraging. If researchers conducting single-case meta-analyses have data that conform to the assumptions of the model examined, they should feel comfortable interpreting the overall average baseline levels and overall average treatment effects across studies. Still, researchers should be advised to increase the Level-3 sample size or the number of primary studies per meta-analysis whenever possible. With larger Level-3 sample sizes, greater accuracy and precision could be gained in estimating the overall average baseline levels and treatment effects across studies. While single-case meta-analysts are constrained by the availability of primary studies, they could adjust their methods for searching (e.g., expanding their search terms) whenever possible, but they are limited by what the field has generated.

On the other hand, statements about the variation in treatment effects across studies, which are also valued by meta-analysts and single-case researchers, should be viewed cautiously. Even assuming that the model was correctly specified, the variance components at Levels 2 and 3 were biased, and confidence intervals for those estimates were inaccurate. Specifically, the Level-3 (between-study) variance components tended to overcover, and the Level-2 (between-person, within-study) variance components tended to undercover and did not show signs of improvement with larger Level-3 sample sizes.

Implications for researchers conducting single-case studies

For researchers conducting single-case studies, the results of this study provide a few recommendations. The results of this study indicated that fixed effects were more precise any time the amount of variability in the model was smaller. Specifically, our study examined shifts in variability at

Levels 2 and 3, but one may anticipate that paying close attention to ways of reducing variability overall would produce greater precision when estimating the overall average baseline levels and treatment effects across studies. For example, single-case researchers should pay attention to baseline variability or stability in an effort to decrease variability at Level 1. Specifically, single-case researchers should consider increasing the number of data points in the baseline to increase the chances of correctly specifying the model, which in turn will decrease the amount of error variability at Level 1. Single-case researchers should also pay attention to the extent to which the intervention is delivered as intended, often termed *treatment fidelity* or *integrity* (Kazdin, 2011). For example, the use of a standardized protocol across participants would ensure similar treatment effects, whereas studies not using a standardized protocol could lack treatment integrity. This lack of treatment integrity could cause increases in between-person variability, and ultimately decrease precision in the estimate of the overall average treatment effect.

Measurement error can also impact variability, and finding ways to decrease that measurement error could ultimately decrease variability overall. For example, single-case researchers should be consistent in their methods of measurement, in an effort to decrease between-person within-study and between-study variability. Therefore, single-case researchers should make every effort to measure outcomes at the same time of day and for the same amount of time across participants, and even across studies assessing similar types of interventions.

A final recommendation to single-case researchers is to consider previous single-case research that has focused in their particular area of interest when determining the most appropriate outcome measure. Specifically, if single-case researchers from similar areas of interest (e.g., reading or math) measured their outcome variables in the same ways across studies, single-case meta-analysts would have a larger number of primary studies to include in this specific meta-analytic model and could feel more confident in their interpretation of the overall average baseline levels and treatment effects across studies.

Implications for methodologists

For methodologists studying the use of multilevel modeling to meta-analyze single-case data, more research needs to be conducted on more complex treatment effects, such as delayed changes in level, trends in the data that change linearly or nonlinearly with time, and transitory effects. Furthermore, violations of assumptions (e.g., nonnormality of the Level-1, Level-2, or Level-3 errors, or heteroscedasticity of the errors at all levels) and various Level-1 error models (e.g., high-order autoregressive or moving-average models) need to be investigated as well. Investigation of these more complex models would allow for a better understanding of the applicability of the models to a variety of conditions.

References

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single-case. *Behavior, Research, and Therapy, 31,* 621–631.

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: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Erlbaum.

Center, B. A., Skiba, R. J., & Casey, A. (1985–1986). A methodology for the quantitative synthesis of intrasubject design research. *Journal of Special Education*, *19*, 387–400.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Erlbaum.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41,* 372–384.

Ferron, J. M., Farmer, J., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study. *Behavior Research Methods, 42,* 930–943.

Ferron, J., Owens, C. M., & Bell, B. A. (2010). *Multilevel models for combining single-case data: A Monte Carlo examination of treatment effect estimates and inferences*. Denver, CO: Paper presented at the Annual Meeting of the American Educational Research Association.

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

Gingerich, W. J. (1984). Meta-analysis of applied time-series data. *The Journal of Applied Behavioral Science, 20,* 71–79.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

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

Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association, 79,* 853–862.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

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

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

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.

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58,* 127–137.

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

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *The Clinical Journal of Pain, 21,* 56–68.

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

SAS Institute, Inc. (2008). *SAS, release 9.2 [Computer program]*. Carey, NC: SAS Institute.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2001). Approximations of 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.

Scruggs, T. E., Mastropieri, M. A., & Castro, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8,* 24–33.

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

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effects 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,* 196–209.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 3,* 142–151.