Behav Res (2012) 44:1244–1254
DOI 10.3758/s13428-012-0213-1

Multilevel meta-analysis of single-subject experimental designs: A simulation study

Maaike Ugille · Mariola Moeyaert · S. Natasha Beretvas · John Ferron · Wim Van den Noortgate

Published online: 31 May 2012

Abstract One way to combine data from single-subject experimental design studies is by performing a multilevel meta-analysis, with unstandardized or standardized regression coefficients as the effect size metrics. This study evaluates the performance of this approach. The results indicate that a multilevel meta-analysis of unstandardized effect sizes results in good estimates of the effect. The multilevel meta-analysis of standardized effect sizes, on the other hand, is suitable only when the number of measurement occasions for each subject is 20 or more. The effect of the treatment on the intercept is estimated with enough power when the studies are homogeneous or when the number of studies is large; the power of the effect on the slope is estimated with enough power only when the number of studies and the number of measurement occasions are large.

Keywords Meta-analysis · Multilevel · Single-subject experimental design · Effect size

Single-case or single-subject experimental designs (SSEDs) are used when one is interested in the effect of a treatment for one specific subject, a person or another entity. In the most basic design, the time series design, the subject is observed several times before the treatment, during the so called baseline phase, and several times during or after the treatment. Because it is difficult to generalize the results from such an experiment to other subjects, the experiment can be replicated within or across studies. Next, the results of several single-case studies can be combined using meta-analytic techniques. There is a plethora of research indicating how to perform a meta-analysis of group comparison studies, in which study subjects are typically measured only once or a few times (e.g., Cooper, 2010; Lipsey & Wilson, 2001). In contrast, procedures necessary to conduct a meta-analysis of SSEDs are not well documented, and it is not straightforward how SSEDS should be meta-analyzed. This is because SSED data differ from group comparison study data. SSEDs entail a far smaller number of subjects for whom many repeated measures are taken. The resulting small-sample size and interrupted time series data may involve cyclical patterns or serial dependencies (West & Hepworth, 1991).

Reviews of meta-analyses of SSEDs indicate that a multitude of methods are used (Beretvas & Chung, 2008; Maggin, O'Keeffe, & Johnson, 2011), each method having its own advantages and weaknesses. Most studies about the methodology of SSED meta-analysis have focused on the question of which effect size should be used to describe the treatment effect in each study being synthesized (e.g., Campbell, 2004; Parker, Vannest, & Davis, 2011; Wolery, Busick, Reichow, & Barton, 2010). There have been many proposals: nonparametric approaches (e.g., the percentage of nonoverlapping data statistic; Scruggs, Mastropieri, & Casto, 1987), parametric approaches (e.g., standardized mean difference; Gingerich, 1984), and regression-based methods (Alison & Gorman, 1993; Center, Skiba, & Casey, 1985–1986; Van den Noortgate & Onghena, 2003a,

M. Ugille (✉) · M. Moeyaert · W. Van den Noortgate
Faculty of Psychology and Educational Sciences,
University of Leuven,
Vesaliusstraat 2,
3000 Leuven, Belgium
e-mail: maaike.ugille@ppw.kuleuven.be

S. N. Beretvas
University of Texas at Austin,
Austin, TX, USA

J. Ferron
University of South Florida,
Tampa, FL, USA

2003b). Maggin et al. (2011b) compared the weaknesses and advantages of several methods (both parametric and nonparametric), and on the basis of this overview, use of the hierarchical linear or multilevel model seemed most promising. Use of a multilevel model is consistent with the logic of visual analysis, can control for threats to interpretation (e.g., autocorrelation), and has attractive statistical properties (e.g., being able to capture differences between subjects and/or studies in the magnitude of the effect).

In this study, we will examine the effectiveness of using a multilevel meta-analysis to synthesize effect sizes from a set of SSEDs' results. Van den Noortgate and Onghena (2008) illustrated this approach using a reanalysis of the study results combined in the meta-analysis of Shogren, Fagella-Luby, Bae, and Wehmeyer (2004). However, an illustration using real data does not prove that the multilevel approach results in proper parameter and standard error estimates for the effect size and variance components. Simulation research makes it possible to investigate the latter question, because population parameters are specified in advance and, therefore, also known. First, we will discuss the multilevel approach to meta-analysis of SSEDs; next, we will present our simulation study.

Multilevel meta-analysis of SSEDs

Meta-analysis is a set of statistical methods for combining the results of various studies addressing the same research question (Glass, 1976). In order to combine these analysis results, study results are typically first converted to a common standardized effect size metric. The advantage of using effect sizes is that it is not necessary that all raw data are available in all studies. Effect sizes may already be reported in a study, or they can be calculated on the basis of available test statistics. One possible way to calculate effect sizes when using SSEDs is to make use of a regression model. Center et al. (1985–1986) proposed using the following regression model to analyze data from an SSED study:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 D_i + \beta_3 T_i D_i + e_i \quad (1)$$

where Y_i is the score of the subject at the ith point in time, D_i is a dummy variable that equals 0 in the baseline phase and 1 in the treatment phase, and T_i is a time-related variable that equals 0 on the first day of the treatment phase. Therefore, β_0 is the baseline intercept, and β_1 is the linear trend during the baseline. β_2 refers to the treatment effect on the intercept for the trend during the intervention phase, and β_3 refers to the effect of the treatment on the time trend. Van den Noortgate and Onghena (2003a) proposed using the last two regression coefficients as effect size measures: β_2 for the immediate treatment effect and β_3 for the treatment effect on the time trend.

If we can obtain, for each case, estimates for these two effect sizes, either by using Eq. 1 on the raw data or on the basis of reported summary statistics or test statistics, then the resulting effect sizes can be combined in two separate meta-analyses: one for the immediate treatment effect and one for the treatment effect on the time trend. There are several ways to do this. A three-level model presented by Van den Noortgate and Onghena (2003b, 2008) makes it possible to model variability in effect sizes at each of the three levels. At the first level, the effect size estimates of the immediate treatment effect for case j from study k may be modeled to randomly deviate from the unknown population effect size:

$$b_{2jk} = \beta_{2jk} + r_{2jk} \quad \text{with } r_{2jk} \sim N\left(0, \sigma^2_{r_{2jk}}\right) \quad (2)$$

with b_{2jk} the ordinary least squares (OLS) estimate of β_{2jk}. The random deviations of the observed regression coefficients are assumed to be normally distributed with a variance that depends on the coefficient and the subject. Because we have only one estimate per case for a coefficient, this variance cannot be estimated in the three-level analysis. However, the sampling variance of the regression coefficients, $\sigma^2_{r_{2jk}}$, can be estimated in the original OLS regression analysis used to estimate the regression coefficient or can be derived from summary or test statistics reported in the primary SSED study. Because these variances are estimated before the actual meta-analysis is performed, the three-level meta-analysis (as well as other typical meta-analyses) can be regarded as a "variance known problem" (Raudenbush & Bryk, 2002).

The population effect sizes β_{2jk} from study k can be modeled as varying over subjects around a study-specific mean effect θ_{20k} (second level) as follows:

$$\beta_{2jk} = \theta_{20k} + u_{2jk} \quad \text{with } u_{2jk} \sim N\left(0, \sigma^2_{u_{2jk}}\right) \quad (3)$$

and the effects for studies can vary between studies (third level):

$$\theta_{20k} = \gamma_{200} + \upsilon_{20k} \quad \text{with } \upsilon_{20k} \sim N\left(0, \sigma^2_{\upsilon_{20k}}\right) \quad (4)$$

The same fundamental three-level model could also be used to model variability in the treatment's effect on the time trend, using the following level one, two, and three equations,

$$b_{3jk} = \beta_{3jk} + r_{3jk} \quad \text{with } r_{3jk} \sim N\left(0, \sigma^2_{r_{3jk}}\right) \quad (5)$$

$$\beta_{3jk} = \theta_{30k} + u_{3jk} \quad \text{with } u_{3jk} \sim N\left(0, \sigma^2_{u_{3jk}}\right) \tag{6}$$

$$\theta_{30k} = \gamma_{300} + \upsilon_{30k} \quad \text{with } \upsilon_{30k} \sim N\left(0, \sigma^2_{\upsilon_{30k}}\right), \tag{7}$$

respectively.

When the scale used is not the same for all subjects, the effect sizes b_{2jk} and b_{3jk} are not comparable across subjects and, therefore, cannot be combined. For example, if the dependent variable for a first subject can range from 0 to 10 and for a second subject from 0 to 100, the expected unstandardized effect size of the second subject may be 10 times larger than the effect size of the first subject. Unstandardized effect sizes are appropriate only when the dependent variable is measured in the same way for all subjects. In practice, this situation is rare, unless studies are exact replications of each other. In the example, also the expected residual standard deviation σ_e will be 10 times larger for the second subject. Therefore, effect sizes can be made comparable over subjects by standardizing b_{2jk} or b_{3jk}, by dividing them by the estimated residuals' standard deviation $(\widehat{\sigma_e})$ (Van den Noortgate & Onghena, 2003b, 2008). This residual standard deviation can be estimated by estimating the OLS regression model in Eq. 1 separately for each subject and then finding the square root of the mean square error or the RMSE of each separate regression. When there are a lot of measurements for each subject, this standard deviation can be estimated reasonably well. On the other hand, when there are only a few measurements available for each subject, this standard deviation might be poorly estimated, resulting in poor estimates of the standardized effect size. Therefore, in this study, we are primarily interested in the performance of a multilevel meta-analysis when there are not many measurements for each subject.

During a single-subject experiment, subsequent measurements can be influenced by common (random) factors, with the result that measurements close to each other in time may be correlated. This phenomenon of autocorrelation can be included in the previous multilevel model by specifying an autoregressive covariance structure for the first-level errors (i.e., the e_i from Eq. 1). In the present study, however, we will focus on the basic model with no autocorrelation.

The parameters typically estimated in a multilevel analysis are the fixed effects regression coefficients (e.g., γ_{200} referring to the average immediate treatment effect over cases and studies and γ_{300} referring to the average treatment effect on the linear trend in Eqs. 4 and 7, respectively). Although this multilevel approach seems well suited for single-case data, in practice, it is seldom used. One reason might be that little is known about the functioning of this model for SSED meta-analysis. This study assesses the performance of this method in more detail, using a set of realistic situations.

Method

To evaluate the performance of the three-level approach, we used a simulation study consisting of several steps. In a first step, raw data were simulated. In a second step, the unstandardized regression coefficients of Eq. 1 were estimated for each subject, using OLS estimation. Next, these estimates were divided by the residual within-phase standard deviation, to obtain corresponding standardized effect sizes. Finally, the unstandardized and the standardized effect sizes were separately analyzed using the three-level meta-analytic approach (using Eqs. 2–7), and the results were compared with the parameter values used to generate data. Despite the fact that meta-analyses with unstandardized effect sizes will rarely be performed in practice, they are analyzed in this study, because if problems arise, we can find out whether these problems are due to the standardization of the effects or to the multilevel meta-analysis itself.

To simulate the raw data, the following measurement occasion (level one) model was used:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}T_{ijk} + \beta_{2jk}D_{ijk} + \beta_{3jk}T_{ijk}D_{ijk} + e_{ijk} \quad \text{with } e_{ijk} \sim N\left(0, \sigma^2_e\right) \tag{8}$$

with occasions nested within individuals at level two:

$$\begin{cases} \beta_{0jk} = \theta_{00k} + u_{0jk} \\ \beta_{1jk} = \theta_{10k} + u_{1jk} \\ \beta_{2jk} = \theta_{20k} + u_{2jk} \\ \beta_{3jk} = \theta_{30k} + u_{3jk} \end{cases} \text{with} \begin{bmatrix} u_{0jk} \\ u_{1jk} \\ u_{2jk} \\ u_{3jk} \end{bmatrix} \sim N(0, \Sigma_u), \tag{9}$$

and within studies at level three:

$$\begin{cases} \theta_{00k} = \gamma_{000} + \upsilon_{00k} \\ \theta_{10k} = \gamma_{100} + \upsilon_{10k} \\ \theta_{20k} = \gamma_{200} + \upsilon_{20k} \\ \theta_{30k} = \gamma_{300} + \upsilon_{30k} \end{cases} \text{with} \begin{bmatrix} \upsilon_{00k} \\ \upsilon_{10k} \\ \upsilon_{20k} \\ \upsilon_{30k} \end{bmatrix} \sim N(0, \Sigma_\upsilon) \tag{10}$$

Note that we simulated data on the same scale for each subject and study. However, this is not a limitation in this simulation study. If we had simulated data using different scales, this effect would be neutralized by the standardization (e.g., if for a specific subject we had multiplied each score by five, both the estimated regression coefficients and the estimated residual standard deviations would be 5 times larger, so the estimated standardized coefficients would remain unchanged). By simulating data on the same scale, however, it is possible to evaluate the use of the multilevel model for SSED meta-analysis for situations in which standardization is required or is not required at the same time.

The effect sizes were synthesized using the restricted maximum likelihood estimation procedure implemented in SAS PROC MIXED (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). We used the Satterthwaite approach to estimating degrees of freedom. This approach has been shown for two-level analyses of multiple-baseline data to

provide relatively accurate confidence intervals for estimates of the average treatment effect (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009), using a model without linear trends to simulate and analyze the data. On the basis of these results, we expect that use of this estimation procedure for three-level analyses would also result in relatively accurate confidence intervals for the overall treatment effects.

The γ_{200} coefficient represents the shift in level that occurs due to the treatment. Data were generated assuming no effect ($\gamma_{200} = 0$) and 2 times the within-phase standard deviation ($\gamma_{200} = 2$). We chose values of 0 (no effect) and 0.2 times the within-phase standard deviation for the overall effect on the slope, γ_{300}. These values were based on reanalyses of meta-analyses (Alen, Grietens, & Van den Noorgate, 2009; Denis, Van den Noortgate, & Maes, 2011; Kokina & Kern, 2010; Shogren et al., 2004; Wang, Cui, & Parrila, 2011). The effects of the baseline regression coefficients γ_{000} and γ_{100}, were set at zero.

The number of studies (K) in each simulated meta-analysis was manipulated to be 10 or 30. A review of 39 social science single-case meta-analyses (Farmer, Owens, Ferron, & Allsopp, 2010) showed that the number of studies included in a meta-analysis ranged from 3 to 117, with 60 % of the meta-analysis including fewer than 30 studies.

We simulated studies with a multiple baseline design. The number of subjects per study (J) was 3, 4, or 7. These values are based on recommendations of Barlow and Hersen (1984) and Kazdin and Kopel (1975), a survey of multiple-baseline studies (Ferron et al., 2010), a review of Farmer et al. (2010), and a survey of single-case studies of Shadish and Sullivan (2011).

The series lengths (I) consisted of 10, 20, or 40 observations. The survey of Ferron et al. (2010) found average series lengths that ranged from 7 to 58, with a median of 24, and the survey of Shadish and Sullivan (2011) found that the number of data points per case ranged from 2 to 160, with median and mode equal to 20. A meta-analysis of 85 single-case studies (Swanson & Sachse-Lee, 2000) found that 25 studies had fewer than 11 treatment sessions, 37 studies had between 11 and 29 treatment sessions, and 23 studies had more than 29 treatment sessions.

The intervention introductions were staggered across subjects within studies. For each combination of the number of subjects and number of data points, the moment at which the treatment started is given in Table 1. For example, when there were 3 subjects and the number of measurement occasions for each subject equaled ten, then for the first subject, the treatment started on the fourth measurement occasion, for the second subject on the sixth measurement occasion, and for the third subject on the eighth measurement occasion, and for all 3 subjects the treatment lasted until the tenth measurement occasion.

Elements of the within-study variance matrix, Σ_u, were manipulated to have conditions with relatively small and relatively large amounts of within-study variance. For simplicity, covariances between regression coefficients were set to zero at the subject and study levels. Therefore, Σ_u is a diagonal matrix, $\sum_u = \mathrm{diag}\left(\sigma^2_{u_0}, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_{u_3}\right)$. A review of several reanalyses indicated that the variance between subjects is sometimes less than the within-person variance (Ferron et al., 2009; Van den Noortgate & Onghena, 2003a) and sometimes greater than the within-person variance (Van den Noortgate & Onghena, 2008). If the within-person (level one) variance is set to 1.0, setting the four diagonal elements of Σ_u to values of 2, 0.2, 2, 0.2 (for the variances in the baseline intercept, baseline slope, treatment effect on the intercept, and treatment effect on the slope residuals, respectively) represents a relatively large amount of within-study variability, while setting the four diagonal elements of Σ_u to values of 0.5, 0.05, 0.5, 0.05, respectively, represents a relatively small amount of within-study variability. Reanalyses of real data sets (Denis et al., 2011; Kokina & Kern, 2010; Shogren et al., 2004) indicated that the variance of the effect of γ_{200} is sometimes much larger than the variance of the effect of γ_{300}. Therefore, we also set the four diagonal elements of Σ_u to values of 8, 0.08, 8, 0.08.

Elements of the between-study variance matrix, Σ_v, were also manipulated. On the basis of reanalyses of meta-analyses (Alen et al., 2009; Denis et al., 2011; Heyvaert, Maes, Van den Noorgate, Kuppens, & Onghena, in press; Kokina & Kern, 2010; Shogren et al., 2004; Wang et al., 2011), we set $\sum_u = \mathrm{diag}\left(\sigma^2_{v_0}, \sigma^2_{v_1}, \sigma^2_{v_2}, \sigma^2_{v_3}\right)$ equal to diag(2, 0.2, 2, 0.2), diag(0.5, 0.05, 0.5, 0.05), and diag(8, 0.08, 8, 0.08).

Crossing the levels of the seven factors leads to a 3 × 3 × 2 × 2 × 2 × 3 × 3 factorial design yielding 648 experimental

Table 1 The number of the measurement occasion at which the treatment started

		Number of subjects		
		3	4	7
Total number of data points	10	4 – 6 – 8	4 – 5 – 7 – 8	4 – 5 – 5 – 6 – 7 – 7 – 8
	20	7 – 11 – 15	7 – 10 – 12 – 15	7 – 9 – 9 – 11 – 13 – 13 – 15
	40	11 – 21 – 31	11 – 18 – 24 – 31	11 – 15 – 15 – 21 – 27 – 27 – 31

conditions. For each condition, 2,000 data sets were simulated and analyzed, with a total of 1,296,000 data sets.

Results

We will successively discuss the estimates of both fixed effects used to describe an intervention's effect (on the intercept and slope), the mean squared error, the estimation of the corresponding standard errors, the confidence interval coverage, the power, and the estimates of the variances. Because it is impossible to discuss the 648 conditions separately, we explored variation between conditions using an ANOVA, modeling main effects and two-way interaction effects, and discuss below only the effect for which the ANOVA showed clear evidence ($p < .001$). This procedure was primarily used to distinguish the most important patterns in the results.

Overall effect size estimates

In the simulation study, the mean population effect on the intercept was 0 or 2, and the effect on the time trend was 0 or 0.2. In each meta-analysis, we estimated these mean effects. These estimates will likely deviate from this mean population effect, because of random variation at each of the three levels.

In Fig. 1, the distribution of the deviations is given for γ_{200} equal to 0 or 2 and the number of measurements equal to 10, 20, or 40 when the unstandardized and the standardized regression coefficients are analyzed. For the unstandardized effect size estimates, close to no bias was identified.

The results differ for the standardized effect sizes. When γ_{200} equaled 0, the estimated bias was close to zero. When γ_{200} equaled 2, the estimated bias of the standardized effect sizes was 0.310 when there are only 10 measurement occasions, 0.110 when there are 20 measurements, and 0.044 when there are 40 measurements. We also sorted all conditions by their estimated bias, and the 100 conditions with the largest bias all had $\gamma_{200} = 2$ and $I = 10$. The condition with the most bias was $\gamma_{200} = 2$, $\gamma_{300} = 0$, $K = 10$, $J = 4$; $I = 10$, $\sigma^2_{v_2} = 8$, and $\sigma^2_{u_2} = 0.5$; in this condition, the bias equaled 0.350. The number of studies, the number of subjects, and the true values of the between- and within-study variances were each found to have no substantial effect on the estimated bias.

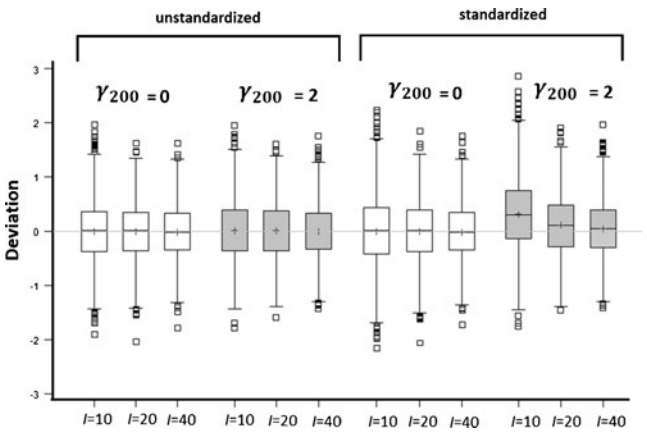

Fig. 1 Distribution of the deviations of the estimated mean effect on the intercept of the treatment from its populations value (γ_{200}) for both the unstandardized and standardized effect sizes, for $\gamma_{300} = 0.2$, $K = 10$, $J = 4$, $\sigma^2_{v_2} = 2$, and $\sigma^2_{u_2} = 2$ conditions

The results for the estimates of γ_{300} were similar to what was found for γ_{200}. There was positive bias when standardized effect sizes were used if $\gamma_{300} = 0.2$ and $I = 10$. If $\gamma_{200} = 2$, $\gamma_{300} = 0.2$, $K = 10$, $J = 4$; $I = 10$, $\sigma^2_{v_2} = 2$, and $\sigma^2_{u_2} = 2$, the bias was 0.030, with a minimum and maximum deviation of −1.071 and 0.883, respectively, and with lower and upper quartiles of −0.115 and 0.178. The relative biases of the two estimates when a meta-analysis on standardized effect sizes was performed were more or less the same: The relative bias of the estimate of both γ_{200} and γ_{300} was 0.155. The standard deviation of the relative deviations, on the other hand, differed: 0.324 for the estimates of γ_{200} and 1.090 for the estimates of γ_{300}.

These results indicate that the positive bias when γ_{200} and γ_{300} are larger than zero and the number of measurements equals 10 results from the standardization of the effects. In this simulation study, however, it is not clear whether this positive bias is caused by calculating a weighted average of the individual regression coefficients or whether these individual regression coefficients themselves are already biased. That is why we also checked the distribution of the deviations from the true value of the individual regression coefficients β_{2jk} for $\gamma_{200} = 2$, $\gamma_{300} = 0.2$, $K = 10$, $J = 4$, $I = 10$, $\sigma^2_{v_2} = 2$, and $\sigma^2_{u_2} = 2$. In this condition, there were 80,000 estimates of β_{2jk}. Table 2 shows that there is almost no bias when the effect sizes are unstandardized and that there is a positive bias of 0.302 when the effect sizes are standardized. This indicates that the bias of the estimated overall effects is mainly due to the standardization of the individual regression coefficients and not a result of calculating a weighted average.

Mean squared error

The mean squared error (MSE) is equal to the average squared deviation of the estimates from the true value, and therefore, it is preferred that the MSE is as small as possible. The MSE is, as was expected, smaller when the number of studies, the number of cases, or the number of measurement occasions increases. On the basis of the ANOVA, the number of studies and the between-study variance seem to have the most important influence on the MSE. This influence is illustrated in Table 3 for both the unstandardized and the standardized effect sizes.

Table 2 Distribution of the deviations of the individual regression coefficients estimates of β_{2jk}, for $\gamma_{200} = 2$, $\gamma_{300} = 0.2$, $K = 10$, $J = 4$, $I = 10$, $\sigma^2_{v_2} = 2$, and $\sigma^2_{u_2} = 2$ conditions

	Mean	SD	Min	Max
Unstandardized	−0.003	2.452	−10.201	10.444
Standardized	0.302	3.123	−22.297	55.865

Standard error estimates

In addition to assessing parameter estimation for each of the two fixed effects (on the intercept and slope), we also evaluated estimation of the standard errors of each of these effects. These standard errors can be used to construct confidence intervals and to perform tests of the statistical significance of the effects. By definition, the standard error equals the standard deviation of the sampling distribution of the estimated effects. In this simulation study, we performed for each condition 2,000 meta-analyses, which results in 2,000 estimates of the effect. Because of the reasonably large number of estimates, we can regard the standard deviations of the estimates as a good estimate of the standard deviations of the estimator's sampling distribution and can, therefore, use this standard deviation to evaluate the quality of the standard error estimates.

The median of the standard error estimates was found to be almost equal to the standard deviation of the estimates of the effect for both the unstandardized and the standardized effects. As was expected, the standard error decreased when the number of studies, the number of measurement occasions, or the number of subjects increased or in conditions with lower between-study or within-study variance values.

The results were similar for both the standardized and the unstandardized effect sizes. The standard error of the unstandardized estimate of γ_{200} was slightly underestimated in 83.3 % of the conditions and the relative bias over all conditions was −0.022. The difference between the median of the standard error estimates and the standard deviation of the effect size estimates was greatest for $\gamma_{200} = 0$, $\gamma_{300} = 0$, $K = 10$, $J = 3$, $I = 10$, $\sigma^2_{v_2} = 8$, and $\sigma^2_{u_2} = 8$. In this situation, the median of the standard error equaled 1.020, and the standard deviations of the estimates equaled 1.103. For the standardized effects, the standard error was slightly underestimated in 86.4 % of the conditions and the relative bias over all conditions was −0.025. The difference between the median of the standard error and the standard deviation of the estimates was greatest in the same conditions as for the unstandardized data, with the median of the standard error equal to 1.198 and the standard deviations of the estimates equal to 1.311. On the basis of the ANOVA, there seemed to be only a small effect of the number of studies on the difference between the median of the standard error and the standard deviation of the estimates. The results were also similar for the estimates of the standard error of γ_{300}.

Confidence interval coverage

Another way to evaluate estimation of the effect sizes and of their corresponding standard errors is by calculating the proportion of replications in which the confidence interval contained the population effect size. The lower and upper limits of the confidence intervals around the estimated effect are constructed by multiplying the estimated standard error with the right critical value of the *z*-distribution and subtracting from and adding this product to the point estimate of the effect. For a 95 % confidence interval, we expected that the coverage proportion would be around 95 %. Because we simulated 2,000 data sets for each condition, the coverage proportions could be estimated accurately—more specifically, with a standard error of 0.005 $\left(=\sqrt{(0.95*0.05)/2000}\right)$—and therefore, we expected the coverage proportion to range from 94.04 % to 95.96 % (with $\alpha = .05$).

The coverage proportions for the unstandardized estimate of γ_{200} ranged from 93.45 % to 97.10 %, and in 91.36 % of the conditions, the coverage proportion lay between 94.04 % and 95.96 %. The coverage proportions for the unstandardized estimate of γ_{300} ranged from 93.60 % to 96.85 %. In 92.90 % of the conditions, the coverage proportion lay between 94.04 % and 95.96 %.

The results for the standardized estimates of γ_{200} differed. In 77.47 % of the conditions, the coverage proportions lay between 94.04 % and 95.95 %; however, for the

Table 3 Mean squared error of (*MSE*) γ_{200} and γ_{300}, *for* $\gamma_{200} = 2$, $\gamma_{300} = 0.2$, $J = 4$, $I = 20$, and $\sigma^2_{u_2} = 2$ conditions

K	$\sigma^2_{v_2}$	*MSE* of γ_{200}		$\sigma^2_{v_3}$	*MSE* of γ_{300}	
		Unstandardized	Standardized		Unstandardized	Standardized
10	0.5	0.129	0.161	0.05	0.011	0.012
	2	0.261	0.307	0.08	0.015	0.017
	8	0.877	0.996	0.20	0.025	0.028
30	0.5	0.041	0.059	0.05	0.004	0.004
	2	0.083	0.104	0.08	0.005	0.006
	8	0.302	0.347	0.20	0.008	0.010

other conditions, the coverage proportions were often too low, with a minimum of 70.65 % when $\gamma_{200} = 2$, $\gamma_{300} = 0.2$, $K = 30$, $J = 7$, $I = 10$, $\sigma^2_{v_2} = 0.5$, and $\sigma^2_{u_2} = 0.5$. There appears to be a problem with coverage when the effect is larger than zero and when the number of measurements occasions is small ($I = 10$); and the problem gets worse when the number of studies is large ($K = 30$) and when the between-study $\left(\sigma^2_{v_2}\right)$ and within-study $\left(\sigma^2_{u_2}\right)$ variances are rather small (Table 4). Figure 1 already showed that there was positive bias when the number of measurement occasions is small and when the effect is larger than zero. When the number of studies increases, the confidence interval becomes narrower, and the lower coverage proportions make it clear that this confidence interval varies around a biased estimator.

The 95 % coverage proportion for the standardized estimates of γ_{300} ranged from 92.8 % to 97.1 %. In 88.27 % of the conditions, the mean coverage proportion lay between 94.04 % and 95.96 %. We already mentioned that the relative bias for the estimates of γ_{200} and γ_{300} was more or less the same but that the standard deviation of the relative deviations of the estimates of γ_{300} from the true value was larger. This will result (again in relative terms) in larger estimated standard errors and wider confidence intervals, and this results in a better coverage proportion, as compared with the coverage proportion of the standardized estimates of γ_{200}.

Power

In our study, we estimated the actual Type I error rate for conditions where the null hypothesis was true (i.e., $\gamma_{200} = 0$ or $\gamma_{300} = 0$), by calculating the proportion of data sets for which the null hypothesis was rejected with α equal to .05. For γ_{200}, this proportion is given in Table 5, in the fourth through seventh columns, for the unstandardized estimates. The results for the standardized estimates were very similar.

When the null hypothesis is false, we want the proportion of correct rejections of the null hypothesis to be as high as possible and, preferably, above .80. The estimated power for $\gamma_{200} = 2$ and $\alpha = .05$ is given in the last four columns of Table 5. On the basis of the ANOVA, the number of studies, the within-study variance, the between-study variance, and all the interactions between these factors seemed to influence

Table 4 Mean coverage proportion of the 95 % confidence interval of γ_{200} for both the unstandardized (U) and standardized (S) effect sizes for $\gamma_{300} = 0.2$ and $J = 4$ (upper section of the table) and of γ_{300} for both the unstandardized (U) and standardized (S) effect sizes for $\gamma_{200} = 2$ and $J = 4$ (lower section of the table)

	K	I	$\gamma_{200} = 0$				$\gamma_{200} = 2$			
			$\sigma^2_{v_2} = 0.5$		$\sigma^2_{v_2} = 8$		$\sigma^2_{v_2} = 0.5$		$\sigma^2_{v_2} = 8$	
			$\sigma^2_{u_2} = 0.5$	$\sigma^2_{u_2} = 8$	$\sigma^2_{u_2} = 0.5$	$\sigma^2_{u_2} = 8$	$\sigma^2_{u_2} = 0.5$	$\sigma^2_{u_2} = 8$	$\sigma^2_{u_2} = 0.5$	$\sigma^2_{u_2} = 8$
U	10	10	.950	**.962**	.950	.953	.948	.957	.947	.949
		20	.949	**.962**	.948	.949	.953	.960	.955	.951
	30	10	.949	.954	.944	.949	.950	.957	.952	.950
		20	.951	.955	.952	.953	.951	.955	.948	.949
S	10	10	.951	**.963**	.952	.954	**.914**	.944	.944	.944
		20	.949	**.964**	.949	.951	.946	.960	.955	.948
	30	10	.950	.955	.947	.952	**.775**	**.889**	**.924**	**.935**
		20	.952	.957	.951	.952	**.920**	.945	.946	.949
	K	I	$\gamma_{300} = 0$				$\gamma_{300} = 0.2$			
			$\sigma^2_{v_3} = 0.05$		$\sigma^2_{v_3} = 0.08$		$\sigma^2_{v_3} = 0.05$		$\sigma^2_{v_3} = 0.08$	
			$\sigma^2_{u_3} = 0.05$	$\sigma^2_{u_3} = 0.08$	$\sigma^2_{u_3} = 0.05$	$\sigma^2_{u_3} = 0.08$	$\sigma^2_{u_3} = 0.05$	$\sigma^2_{u_3} = 0.08$	$\sigma^2_{u_3} = 0.05$	$\sigma^2_{u_3} = 0.08$
U	10	10	.958	.955	.952	.954	.954	.956	.952	.954
		20	.952	.951	.950	.953	.952	.953	.945	.950
	30	10	.950	.956	.943	.952	.952	.957	.954	.953
		20	.950	.949	.948	.952	.947	.950	.948	.948
S	10	10	**.961**	.959	.956	.960	.954	.956	.953	.955
		20	.954	.952	.951	.954	.955	.950	.946	.952
	30	10	.950	.957	.944	.952	.942	.944	.941	**.938**
		20	.953	.948	.950	.952	.942	.950	.946	.948

Coverage proportions that are significant different from .950 ($\alpha = .05$) appear in bold

Table 5 Power of γ_{200} for $\gamma_{300} = 0.2$, for the unstandardized effect sizes

K	J	I	$\gamma_{200} = 0$				$\gamma_{200} = 2$			
			$\sigma^2_{v_2} = 0.5$		$\sigma^2_{v_2} = 8$		$\sigma^2_{v_2} = 0.5$		$\sigma^2_{v_2} = 8$	
			$\sigma^2_{u_2} = .5$	$\sigma^2_{u_2} = 8$	$\sigma^2_{u_2} = .5$	$\sigma^2_{u_2} = 8$	$\sigma^2_{u_2} = .5$	$\sigma^2_{u_2} = 8$	$\sigma^2_{u_2} = .5$	$\sigma^2_{u_2} = 8$
10	3	10	.045	.039	.047	.041	**.993**	**.818**	.467	.378
		20	.041	.044	.049	.055	**.998**	**.851**	.515	.394
		40	.047	.034	.051	.047	**1.000**	**.875**	.498	.389
	4	10	.055	.036	.045	.054	**.999**	**.891**	.465	.402
		20	.058	.033	.050	.051	**1.000**	**.922**	.506	.417
		40	.059	.040	.046	.046	**1.000**	**.937**	.528	.412
	7	10	.051	.042	.058	.047	**1.000**	**.985**	.480	.457
		20	.056	.038	.057	.049	**1.000**	**.988**	.506	.456
		40	.054	.047	.052	.050	**1.000**	**.991**	.494	.475
30	3	10	.053	.049	.049	.056	**1.000**	**1.000**	**.947**	**.874**
		20	.043	.042	.045	.051	**1.000**	**1.000**	**.956**	**.892**
		40	.047	.042	.054	.053	**1.000**	**1.000**	**.954**	**.904**
	4	10	.050	.049	.060	.046	**1.000**	**1.000**	**.950**	**.905**
		20	.052	.048	.049	.040	**1.000**	**1.000**	**.959**	**.911**
		40	.052	.049	.043	.042	**1.000**	**1.000**	**.956**	**.909**
	7	10	.052	.042	.059	.053	**1.000**	**1.000**	**.959**	**.948**
		20	.052	.047	.049	.052	**1.000**	**1.000**	**.962**	**.941**
		40	.055	.045	.054	.048	**1.000**	**1.000**	**.957**	**.936**

Values equal to or larger than .80 appear in bold

power. For conditions in which the number of studies being meta-analyzed was 30, the power was high. On the other hand, when there were only 10 studies and a lot of between-study variability, the power was found to be much lower.

Table 6 contains the power of the estimates of γ_{300} for the unstandardized effects. The estimated power was lower, as compared with the power of γ_{200}. The power was above .80 only when $K = 30$ and $I > 10$, except if the number of cases equaled 7. Here again, the results for the estimates of the standardized effects were very similar.

Variance estimates

In the meta-analyses, the between-study and within-study residuals' variances were estimated for both the effect on the intercept and the effect on the slope parameters. We will examine the relative deviations, which are the deviations from the true value divided by the value of the population parameter. In this study, the same problem arises for the estimates of the four variances and for the meta-analyses of both the unstandardized and the standardized effect sizes: namely, the distribution is positively skewed, with a long tail at the right side. For example, for the meta-analyses on the standardized effect sizes, the relative bias (i.e., the mean relative deviation) was larger than 100 % for 13.39 % of the estimates of the between-study variance of the effect on the intercept, for 12.52 % of the estimates of the between-study variance of the effect on the slope, for 25.07 % of the estimates of the within-study variance of the effect on the intercept, and for 26.94 % of the estimates of the within-study variance of the effect on the slope. Table 7 shows the mean and median of the relative deviations of the variance estimates.

Table 7 indicates that the relative bias of the variance estimates was larger when the meta-analysis was performed on standardized effect sizes. The number of measurements also had a large effect on the bias of the four estimated variances. The within-study variance exhibited the highest bias when the number of measurement occasions was 10. In general, the median relative deviation of the estimates of the within-study variance was larger than that found for the between-study variance estimates. The median of the relative deviation of the within-study variance of γ_{200} was especially large when the between-study variance was large and the within-study variance was small.

Discussion

In this study, we examined whether single-case studies can be combined using a multilevel meta-analysis, with

Table 6 Power of γ_{300} for $\gamma_{200} = 2$, for the unstandardized effect sizes

K	J	I	$\gamma_{300} = 0$				$\gamma_{300} = 0.2$			
			$\sigma^2_{v_3} = 0.05$		$\sigma^2_{v_3} = .08$		$\sigma^2_{v_3} = 0.05$		$\sigma^2_{v_3} = .08$	
			$\sigma^2_{u_3} = .05$	$\sigma^2_{u_3} = .08$	$\sigma^2_{u_3} = .05$	$\sigma^2_{u_3} = .08$	$\sigma^2_{u_3} = .05$	$\sigma^2_{u_3} = .08$	$\sigma^2_{u_3} = .05$	$\sigma^2_{u_3} = .08$
10	3	10	.042	.042	.049	.037	.222	.204	.200	.191
		20	.045	.047	.048	.052	.520	.453	.395	.375
		40	.053	.052	.050	.052	.570	.518	.429	.403
	4	10	.039	.043	.052	.042	.275	.273	.240	.250
		20	.045	.049	.047	.048	.557	.498	.397	.390
		40	.057	.050	.048	.053	.591	.552	.472	.443
	7	10	.046	.050	.045	.059	.400	.402	.307	.346
		20	.055	.051	.055	.041	.619	.592	.439	.453
		40	.060	.051	.058	.060	.663	.601	.480	.463
30	3	10	.044	.040	.053	.051	.643	.631	.579	.579
		20	.045	.050	.051	.050	**.957**	**.946**	**.874**	**.868**
		40	.047	.052	.057	.049	**.984**	**.963**	**.930**	**.885**
	4	10	.048	.046	.061	.048	.747	.744	.669	.664
		20	.058	.049	.054	.047	**.974**	**.965**	**.909**	**.892**
		40	.048	.043	.054	.055	**.986**	**.979**	**.930**	**.913**
	7	10	.060	.047	.059	.046	**.905**	**.894**	**.818**	**.801**
		20	.047	.055	.053	.049	**.990**	**.986**	**.935**	**.917**
		40	.055	.046	.047	.046	**.994**	**.990**	**.948**	**.944**

Values equal or larger than .80 appear in bold

unstandardized or standardized regression coefficients as effect sizes. Several realistic conditions were simulated and analyzed.

The multilevel approach works well when unstandardized effect sizes are used. The approach is also suitable for standardized effect sizes when there are many studies (30 or more), when there are a lot of measurement occasions for each subject (20 or more), and when the studies are rather homogeneous (which corresponds with a small amount of between-study variance). In these situations, the effects are well estimated, the mean squared error is small, the coverage proportion of the 95 % confidence interval is around 95 %, and the power of each effect is above 80 %.

Table 7 Mean and median of relative deviations of the variance estimates for $\gamma_{200} = 2$, $\gamma_{300} = 0.2$, $K = 10$, $J = 4$, $\sigma^2_{v_2} = 2$, and $\sigma^2_{u_2} = 2$ conditions

	I	Unstandardized		Standardized	
		Mean	Median	Mean	Median
$\widehat{\sigma^2_{v_2}}$	10	0.037	−0.063	0.367	0.155
	20	−0.006	−0.109	0.089	−0.032
	40	0.011	−0.097	0.0541	−0.055
$\widehat{\sigma^2_{u_2}}$	10	0.173	0.139	1.521	1.237
	20	0.024	−0.008	0.334	0.281
	40	0.007	−0.024	0.123	0.091
$\widehat{\sigma^2_{u_3}}$	10	−0.013	−0.142	0.314	0.086
	20	0.008	−0.096	0.111	−0.006
	40	−0.013	−0.100	0.030	−0.068
$\widehat{\sigma^2_{u_3}}$	10	0.434	0.369	1.647	1.338
	20	0.008	−0.019	0.221	0.174
	40	−0.009	−0.029	0.076	0.056

However, when these criteria are not met, problems may occur. In this study, it became clear that difficulties are encountered in particular when there are only 10 measurement occasions for each subject. In this situation, the overall effect estimate is biased as a result of biased estimates of the individual standardized regression coefficients. Because standardizing is often needed in order to make study results comparable, further research should focus on this problem of standardization when the number of measurements is rather small. A possible solution is the use of iterative bootstrap procedures (Goldstein, 1996), or through correcting the individual regression coefficients for bias (Hedges, 1981). Another solution might be to use the correction procedures for biased standardized regression coefficients that were proposed by Yuan and Chan (2011), but they should be adapted, because the procedures were developed for regression coefficients that are standardized in the

traditional way—namely, by multiplying them by the standard deviation of the independent variable and then dividing by the standard deviation of the dependent variable. In this study, we suggest standardizing based on the standard deviation of the dependent variable only, and only in so far as this dependent variable is not explained by the predictors (i.e., we propose to use the residual standard deviation).

An important question for a researcher who wants to conduct a multilevel meta-analysis of SSEDs has to do with the power of different scenarios. The results are the same for the unstandardized and the standardized effect sizes. When the immediate effect of the treatment is estimated, the power is acceptable when the studies are homogeneous, regardless of the number of studies, cases, or measurement occasions. On the other hand, when the studies' effect sizes are more heterogeneous, a power of 80 % or more can be reached only when there are a lot of studies (e.g., 30) being meta-analyzed. On the other hand, the effect of the treatment on the slope is estimated with enough power only when the number of studies is 30 or more and the number of measurement occasions per subject is 20 or more.

The major advantage of multilevel models is that they result not only in an overall estimate of the effect, but also in an estimate of the between-study and within-study variance. But these estimates are sometimes seriously biased. The estimates are worse for meta-analyses of standardized effects, and the estimates of the within-study variance is especially biased when the number of measurements is rather small.

We should also note that the conclusions are, in principle, limited to the conditions that were simulated. These data were balanced and were sampled from a normal distribution, there was no correlation between consecutive observations, there was no covariance, and the trajectories were not nonlinear. In practice, however, data will probably not be balanced, autocorrelation will likely occur when observations are taken in quick succession, there can be another underlying distribution, there might be correlation between the effects, and the baseline and/or intervention phase trends might be nonlinear. The purpose of this study was to discover which parameters really matter and to identify in which conditions problems occurred. Our aim was to get a preliminary insight into the empirical functioning of the multilevel model for SSED meta-analysis, but in future research, we will also want to investigate more complex situations.

In this study, we performed two separate meta-analyses for the effects b_{2jk} and b_{3jk}, whereby we assumed that these two effects are independent of each other. This assumption may not be realistic in applied settings. If a covariance at the second and/or third level can be expected, a multivariate meta-analysis might be more powerful (Kalaian & Raudenbush, 1996). In a pilot study, we already explored the performance of the multivariate approach, with simulated effect sizes that did not covary at the second and third level and showed sampling covariance only at the first level. The results were similar to the ones of this study. In future research, we want to explore the operating characteristics of the multivariate multilevel model for meta-analytic data of SSEDs in situations where there is nonzero covariance between parameters. We expect that the gains of such a multivariate model increases (e.g., higher power and accuracy) when the covariance increases.

We also assumed that the repeated measures were uncorrelated. This is probably too strong an assumption in some real situations, because a typical characteristic of single-case data is that the measurements are taken in rapid succession. Shadish and Sullivan (2011) showed that the size of autocorrelation varies substantially over studies. In previous research (Ferron et al., 2009), where a two-level model was used to analyze SSEDs, it was found that not modeling an existing autocorrelation results in biased parameter estimates. In this study, we modeled the level one errors as $\sigma^2\boldsymbol{I}$, but there are many other covariance structures possible, of which the first-order autoregressive type is often used to model autocorrelation. Thus, future research should explore scenarios where the repeated measures are autocorrelated and assess the impact of failing to model this autocorrelation, as well as evaluating recovery of model parameters, given the small number of data points typically encountered in single-case design research.

Can single-case studies be combined with a multilevel meta-analysis? The answer is yes when it is not necessary to standardize effect sizes. And even if effect sizes should be standardized because studies' outcomes are on different scales, there are no real problems as long as the studies' effects are reasonably homogeneous or when there are a lot of measurements per individual and there are a lot of studies being meta-analyzed. But the method does not work well for standardized effect sizes when there are only a few measurements for each subject; this situation calls for an adaptation of the method and additional research into alternative estimation procedures.

Author note Maaike Ugille, Faculty of Psychology and Educational Sciences, University of Leuven, Belgium; Mariola Moeyaert, Faculty of Psychology and Educational Sciences, University of Leuven, Belgium; S. Natasha Beretvas, Department of Educational Psychology, University of Texas at Austin; John Ferron, Department of Educational Measurement & Research, University of South Florida; Wim Van den Noortgate, Faculty of Psychology and Educational Sciences, ITEC-IBBT Kortrijk, University of Leuven, Belgium.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D110024 to Katholieke Universiteit Leuven, Belgium. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. For the simulations, we used the infrastructure of the VSC–Flemish Supercomputer Center, funded by the Hercules foundation and the Flemish Government–Department EWI.

References

Alen, E., Grietens, H., & Van den Noorgate, W. (2009). *Meta-analysis of single-case studies: An illustration for the treatment of anxiety disorders*. Unpublished master's thesis, Katholieke Universiteit Leuven, Leuven, Belgium.

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

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

Beretvas, S. N., & Chung, H. (2008). A review of single-subject experimental design meta-analyses: Methodological issues and practice. *Evidence-Based Communication and Assessment and Intervention, 2,* 129–141.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28,* 234–246.

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

Cooper, H. (2010). *Research synthesis and meta-analysis* (4th ed.). London: Sage.

Denis, J., Van den Noortgate, W., & Maes, B. (2011). Self-injurious behavior in people with profound intellectual disabilities: A meta-analysis of single-case studies. *Research in Developmental Disabilities, 32,* 911–923.

Farmer, J., Owens, C. M., Ferron, J., & Allsopp, D. (2010). *A review of social science single-case meta-analyses*. Manuscript in preparation.

Ferron, J. M., Bell, B. A., Hess, M. F., 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. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study of multilevel-modeling approaches. *Behavior Research Methods, 42,* 930–943.

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.

Goldstein, H. (1996). Consistent estimators for multilevel generalized linear models using an interated bootstrap. *Multilevel Modelling Newsletter, 8,* 3–6.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.

Heyvaert, M., Maes, B., Van Den Noortgate, W., Kuppens, S., & Onghena, P. (in press). A multilevel meta-analysis of single-case and small-n research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in Developmental Disabilities*.

Kalaian, H. A., & Raudenbush, S. W. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods, 1,* 227–235.

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

Kokina, A., & Kern, L. (2010). Social story interventions for students with autism spectrum disorders: A meta-analysis. *Journal of Autism and Developmental Disorders, 40,* 812–826.

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS© system for mixed models* (2nd ed.). Cary, NC: SAS Institute.

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011a). A quantitative synthesis of methodology in the meta-analysis of single-subject reserach for students with disabilities: 1985–2009. *Exceptionality, 19,* 109–135.

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

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35,* 303–322.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (Vol. 2)*. London: Sage.

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

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43,* 971–980.

Shogren, K. A., Fagella-Luby, M. N., Bae, J. S., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior. *Journal of Positive Behavior Interventions, 6,* 228–237.

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.

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

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35,* 1–10.

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

Wang, S., Cui, Y., & Parrila, R. (2011). Examining the effectiveness of peer-mediated and video-modeling social skills interventions for children with autism spectrum disorders: A meta-analysis in single-case research using HLM. *Research in Autism Spectrum Disorders, 5,* 562–569.

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality, 59,* 602–662.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitaively syntesizing single-subject data. *Journal of Special Education, 44,* 18–28.

Yuan, K., & Chan, W. (2011). Biases and standard errors of standardized regression coefficients. *Psychometrika, 76,* 670–690.