Modeling external events in the three-level analysis of multiple-baseline across-participants designs: A simulation study

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

Published online: 12 December 2012
© Psychonomic Society, Inc. 2012

Abstract In this study, we focus on a three-level meta-analysis for combining data from studies using multiple-baseline across-participants designs. A complicating factor in such designs is that results might be biased if the dependent variable is affected by not explicitly modeled external events, such as the illness of a teacher, an exciting class activity, or the presence of a foreign observer. In multiple-baseline designs, external effects can become apparent if they simultaneously have an effect on the outcome score (s) of the participants within a study. This study presents a method for adjusting the three-level model to external events and evaluates the appropriateness of the modified model. Therefore, we use a simulation study, and we illustrate the new approach with real data sets. The results indicate that ignoring an external event effect results in biased estimates of the treatment effects, especially when

there is only a small number of studies and measurement occasions involved. The mean squared error, as well as the standard error and coverage proportion of the effect estimates, is improved with the modified model. Moreover, the adjusted model results in less biased variance estimates. If there is no external event effect, we find no differences in results between the modified and unmodified models.

Keywords Multiple baseline across participants · Three-level meta-analysis · Effect sizes · External event effect

M. Moeyaert · M. Ugille
University of Leuven,
Leuven, Belgium

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

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

W. Van den Noortgate
University of Leuven,
Leuven, Belgium

M. Moeyaert (✉)
Faculty of Psychology and Educational Sciences,
University of Leuven,
Andreas Vesaliusstraat 2, Box 3762, 3000 Leuven, Belgium
e-mail: mariola.moeyaert@ped.kuleuven.be

A multiple-baseline design (MBD) is one of the variants of single-subject experimental designs (SSEDs). SSED researchers observe and measure a participant or case repeatedly over time. Observations are obtained during at least one baseline phase (when no intervention is present) and at least one treatment phase (when an intervention is present). By comparing scores from both kinds of phases, SSED researchers can assess whether the outcome scores on the dependent variable changed, for instance, in level or in slope when the treatment was present (Onghena & Edgington, 2005).

In an MBD, an AB phase design (with one baseline phase, A, and one treatment phase, B) is implemented simultaneously to different participants, behaviors, or settings (Barlow & Hersen, 1984; Ferron & Scott, 2005; Onghena, 2005; Onghena & Edgington, 2005). MBDs are popular among SSED researchers (Shadish & Sullivan, 2011) because the intervention is introduced sequentially over the participants (or settings and behaviors), which entails the advantage that the researchers can more easily disentangle effects of the intervention and effects of some external events, such as the illness of a teacher, an exciting class activity, the presence of a foreign observer, or

a teacher intern (Baer, Wolf, & Risley, 1968; Barlow & Hersen, 1984; Kinugasa, Cerin, & Hooper, 2004; Koehler & Levin, 2000). This is because, if an external event occurs at certain points in time, the outcome scores for all participants in that study might be simultaneously influenced. Figure 1 gives a graphical presentation of possible consequences for the occurrence of an external event in a multiple-baseline across-participants design. In Fig. 1a, the external event has a constant effect on the dependent variable on subsequent measurements—for instance, the teacher is ill during subsequent days, or there is a foreign observer during some measurement occasions. Figure 1b illustrates a gradually fading away external event effect. For instance, the influence of a teacher intern on the behavior of the students may be reduced over time.

Van den Noortgate and Onghena (2003) proposed the use of multilevel models to synthesize data from multiple SSED studies, allowing investigation of the generalizability of the results and exploration of potential moderating effects. In previous research evaluating this multilevel meta-analysis of MBD data (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Ferron, Farmer, & Owens, 2010; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2012a, 2012b; Owens & Ferron, 2012), the data were typically simulated with a treatment effect and random noise only. Potential

confounding events that could have a simultaneous effect on all participants within a study were not taken into account. In this study, we evaluate the performance of the basic three-level model when there are effects of external events, as well as that of an extension of the model that tries to account for potential event effects. In the following, we first present the basic model and a possible extension to account for external events. Next, we evaluate the performance of both models, by means of a simulation study and an analysis of real data.

Three-level meta-analysis

A meta-analysis combines the results of several studies addressing the same research question (Cooper, 2010; Glass, 1976). Study results are typically first converted to a common standardized effect size before meta-analyzing them. The effect sizes may be reported in the primary studies or can be calculated afterward, using reported summary and/or test statistics.

One possible way to calculate effect sizes when SSEDs are used is to analyze the data using regression models and to use the regression coefficients as effect sizes. A

Fig. 1 Graphical display of a constant external event effect (a) and a gradually fading away external event effect (b) affecting the score on four subsequent moments (day 17, day 19, day 21, and day 23) for a

multiple-baseline design across 3 participants, with the treatment starting on day 6, day 16, and day 24, respectively

regression model of interest here is the one proposed by Center, Skiba, and Casey (1985–1986):

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 D_i + \beta_3 T'_i D_i + e_i. \quad (1)$$

The score of the dependent variable on measurement occasion i (Y_i) depends on a dummy coded variable (D_i) indicating whether the measurement occasion i belongs to the baseline phase ($D_i = 0$) or the treatment phase ($D_i = 1$); a time-related variable T_i that equals 1 on the first measurement occasion of the baseline phase; and an interaction term between the centered time indicator and the dummy variable, $T'_i D_i$, where T'_i is centered such that T'_i equals 0 on the first measurement occasion of the treatment phase. β_0 indicates the expected baseline level β_1 is the linear trend during the baseline, β_2 refers to the immediate treatment effect, and β_3 refers to the effect of the treatment on the time trend.

Van den Noortgate and Onghena (2003) proposed using the ordinary least squares estimates for β_2 and β_3 from Eq. 1 as effect sizes in the three-level meta-analysis. At the first level, the estimated effect sizes of the immediate treatment effect, b_{2jk} , and the treatment effect on the time trend, b_{3jk} , for participant j from study k are equal to the unknown population effect sizes, β_{2jk} and β_{3jk} , respectively, plus random deviation s, r_{2jk} and r_{3jk} , that are assumed to be normally distributed with a mean of zero:

$$\begin{aligned} b_{2jk} &= \beta_{2jk} + r_{2jk} \quad \text{with} \quad r_{2jk} \sim N(0, \sigma_{r_{2jk}}^2) \\ b_{3jk} &= \beta_{3jk} + r_{3jk} \quad \text{with} \quad r_{3jk} \sim N(0, \sigma_{r_{3jk}}^2). \end{aligned} \quad (2)$$

The sampling variances of the observed effects, $\sigma_{r_{2jk}}^2$ and $\sigma_{r_{3jk}}^2$, are the squared standard errors that are typically reported by default when a regression analysis is performed. These variances depend to a large extent on the number of observations and the variance of these observations and, therefore, can be participant and study specific. At the second level, the population effect sizes β_{2jk} and β_{3jk} from Eq. 2 can be modeled as varying over participants around the study-specific mean effect, θ_{20k} and θ_{30k} (Eq. 3):

$$\begin{aligned} \beta_{2jk} &= \theta_{20k} + u_{2jk} \quad \text{with} \quad u_{2jk} \sim N(0, \sigma_{u_{2jk}}^2) \\ \beta_{3jk} &= \theta_{30k} + u_{3jk} \quad \text{with} \quad u_{3jk} \sim N(0, \sigma_{u_{3jk}}^2). \end{aligned} \quad (3)$$

The population effects for studies can vary between studies (third level, Eq. 4):

$$\begin{aligned} \theta_{20k} &= \gamma_{200} + v_{20k} \quad \text{with} \quad v_{20k} \sim N(0, \sigma_{v_{20k}}^2) \\ \theta_{30k} &= \gamma_{300} + v_{30k} \quad \text{with} \quad v_{30k} \sim N(0, \sigma_{v_{30k}}^2). \end{aligned} \quad (4)$$

The model parameters that we are typically interested in when using a multilevel model are the fixed effects regression coefficients (i.e., γ_{200} , referring to the average immediate treatment effect over participants and studies, and γ_{300} , referring to the average treatment effect on the linear trend over participants and studies in Eq. 4) and the variances (i.e., $\sigma_{v_{20k}}^2$, referring to the between-study variance for the estimated immediate treatment effect; $\sigma_{v_{30k}}^2$, indicating the between-study variance for the estimated treatment effect on the time trend; $\sigma_{u_{2jk}}^2$, the between-case variance for the estimated immediate treatment effect; and $\sigma_{u_{3jk}}^2$, referring to the between-case variance of the estimated treatment effect on the time trend).

Correcting effect sizes for external events

External events in a multiple-baseline across-participants design can have an effect on the outcome score(s) of all participants within a study. These external event effects are common in SSEDs, because practitioners often implement these designs in their everyday setting (for example, in the home, school, etc.), where they cannot control for outside experimental factors (Christ, 2007; Kratochwill et al., 2010; Shadish, Cook, & Campbell, 2002). If we do not model these external events, the results might be biased. For instance, suppose that a researcher is interested in a change in challenging behavior and staggers the beginning of the treatment across 3 participants. The 3 participants receive the treatment at day 6, day 16, and day 24, respectively (see Fig. 1) and are observed every 2 days. On days 17, 19, 21, and 23, the teacher is ill, and as a consequence, a substitute teacher takes his or her place, and the participants exhibit more challenging behavior. In this situation, the estimated treatment effect for participants 1 and 2 will be smaller, and the estimated treatment effect for participant 3 will be larger, and therefore differences between participants in the treatment effects are also likely to be overestimated, unless we correct the effect sizes for possible external events.

A possible way to calculate effect sizes corrected for an external event in an SSED is by estimating effect sizes for participants per study, by performing a regression analysis with a model including possible event effects, and by assuming that external events simultaneously affect all participants in a study. Thereafter, the corrected effect sizes can be combined over studies in the three-level meta-analysis.

For the first step, we propose to use an extension of the Center et al. (1985–1986) model, including dummy variables for measurement occasions:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j} T_{ij} + \beta_{2j} D_{ij} + \beta_{3j} D_{ij} T'_{ij} \\ &+ \sum_{m=2}^{I-1} \beta_{(m+2)} M_{mi} + e_{ij}. \end{aligned} \quad (5)$$

The score on the dependent variable Y on measurement occasion i ($= 1, 2, \dots, I$) from participant j ($= 1, 2, \dots, J$) is modeled as a linear function of the dummy-coded variable (D_{ij}) indicating whether the measurement occasion i from participant j belongs to the baseline phase ($D_{ij} = 0$) or the treatment phase ($D_{ij} = 1$); a time-related variable T_{ij} , which equals 1 at the start of the baseline phase; an interaction term between the dummy variable indicating the phase and the time indicator centered around its value at the start of the treatment phase, $D_{ij}T'_{ij}$; and finally, dummy-coded variables indicating the moment ($M_{mi} = 1$ if $m = i$, zero otherwise). By including the effects of individual moments, coefficients β_{2j} and β_{3j} can be interpreted as the treatment effects, corrected for possible external events.

We do not include a dummy variable for one measurement moment in the baseline phase and one measurement moment in the treatment phase. This is to ensure that the model is identified; if we included these parameters as well, an increase in the effects for each moment in the baseline phase could be compensated for by a decrease of the intercept, illustrating that without constraining these parameters, there would be an infinite number of equivalent solutions. For our study, we select the first and last moments as the times at which to set the moment effects to zero, but different moments can be chosen if we suspect a moment effect during one of these times.

While the baseline level and slope (β_{0j} and β_{1j}) and both treatment effects (β_{2j} and β_{3j}) are participant specific, the moment effects are assumed to be the same for all participants from the same study and, therefore, have to be estimated for each study, using all data from that study. To this end, we propose to extend Eq. 5 by including a set of dummy participant indicators. For 2 participants, using dummy participant indicators P_1 and P_2 , respectively, this results in Eq. 6:

$$Y_{ij} = \beta_{01}P_{1j} + \beta_{02}P_{2j} + \beta_{11}T_{i1}P_{1j} + \beta_{12}T_{i2}P_{2j} + \beta_{21}D_{i1}P_{1j} + \beta_{22}D_{i2}P_{2j} + \beta_{31}D_{i1}T'_{i1}P_{1j} + \beta_{32}D_{i2}T'_{i2}P_{2j} + \sum_{m=2}^{I-1} \beta_{(m+2)}M_{mi} + e_{ij}. \quad (6)$$

After using Eq. 6 for each study to estimate the corrected effect sizes (β_{2j} and β_{3j}) for each participant, we can use the three-level meta-analysis (see Eqs. 2–4) to combine the corrected effect size estimates from multiple participants. In principle, we could also use a two-level model per study to estimate the participant-specific effects, but given the typically very small number of participants per study, using a multilevel model might not be recommended.

A simulation study

Simulating three-level data

To evaluate the performance of the basic model and its extension, we performed a simulation study. We simulated raw data using a three-level model. At level 1, we used the following model:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}T_{ijk} + \beta_{2jk}D_{ijk} + \beta_{3jk}T_{ijk}D_{ijk} + e_{ijk} \quad (7)$$

with $e_{ijk} \sim N(0, \sigma_e^2)$,

with measurement occasions nested within participants, which form the units at level two:

$$\begin{cases} \beta_{0jk} = \theta_{00k} + u_{0jk} \\ \beta_{1jk} = \theta_{10k} + u_{1jk} \\ \beta_{2jk} = \theta_{20k} + u_{2jk} \\ \beta_{3jk} = \theta_{30k} + u_{3jk} \end{cases} \quad \text{with} \quad \begin{bmatrix} u_{0jk} \\ u_{1jk} \\ u_{2jk} \\ u_{3jk} \end{bmatrix} \sim N(0, \Sigma_u). \quad (8)$$

The participants are, in turn, clustered within studies at the third level:

$$\begin{cases} \theta_{00k} = \gamma_{000} + v_{00k} \\ \theta_{10k} = \gamma_{100} + v_{10k} \\ \theta_{20k} = \gamma_{200} + v_{20k} \\ \theta_{30k} = \gamma_{300} + v_{30k} \end{cases} \quad \text{with} \quad \begin{bmatrix} v_{00k} \\ v_{10k} \\ v_{20k} \\ v_{30k} \end{bmatrix} \sim N(0, \Sigma_v). \quad (9)$$

Varying parameters

On the basis of a thorough overview of 809 SSED studies, Shadish and Sullivan (2011) enumerated some parameters that characterize SSEDs. On the basis of their results and our reanalyses of meta-analyses of SSEDs (Alen, Grietens, & Van den Noortgate, 2009; Denis, Van den Noortgate, & Maes, 2011; Ferron et al., 2010; Kokina & Kern, 2010; Shadish & Sullivan, 2011; Shogren, Faggella-Luby, Bae, & Wehmeyer, 2004; Wang, Cui, & Parrila, 2011), we decided to vary the following parameters that can have a significant influence on the quality of model estimation:

- γ_{200} , represents the immediate treatment effect on the outcome and had values 0 (no effect) or 2.
- The treatment effect on the time trend, defined by γ_{300} , was varied to have values 0 (no effect) or 0.2.
- The regression coefficients of the baseline, γ_{000} and γ_{100} , did not vary and were set at 0, because the interest is in the overall treatment effects (e.g., the immediate treatment effect and the treatment effect on the time trend).
- The number of simulated participants, J , equaled 4 or 7.
- The number of measurements within a participant, I , was 15 or 30. We chose to keep I constant for all participants within the same study.

- The number of studies, K , was 10 or 30.
- The between-case covariance matrix: Covariances between pairs of regression coefficients were set to zero. Therefore, Σ_u is a diagonal matrix: $\Sigma_u = \text{diag}(\sigma_{u_0}^2, \sigma_{u_1}^2, \sigma_{u_2}^2, \sigma_{u_3}^2) = \text{diag}(2, 0.2, 2, 0.2)$ or $\Sigma_u = \text{diag}(\sigma_{u_0}^2, \sigma_{u_1}^2, \sigma_{u_2}^2, \sigma_{u_3}^2) = \text{diag}(0.5, 0.05, 0.5, 0.05)$.
- The between-study covariance matrix: Covariances between pairs of regression coefficients were set to zero. Therefore, Σ_v is a diagonal matrix: $\Sigma_v = \text{diag}(\sigma_{v_0}^2, \sigma_{v_1}^2, \sigma_{v_2}^2, \sigma_{v_3}^2) = \text{diag}(2, 0.2, 2, 0.2)$ or $\Sigma_v = \text{diag}(\sigma_{v_0}^2, \sigma_{v_1}^2, \sigma_{v_2}^2, \sigma_{v_3}^2) = \text{diag}(0.5, 0.05, 0.5, 0.05)$.
- The moment of introducing a treatment effect was staggered across participants within a study (see Table 1), depending on the number of measurements.

In a first scenario, a constant external event was added to influence four subsequent scores of all the participants within a study (as in Fig. 1a). The moment was randomly generated from a uniform distribution for each study separately. Because we did not include a moment effect for the first and the last moments to make the model identified, the external event effect did not occur on these moments. The external event effect was 0 or 2, representing a null and a large external event effect, respectively.

In a second scenario, the effect of the external event effect was added, which fades away gradually (see Fig. 1b) for all the participants within a study. The effect across four time points was 3.5, 2.5, 1.5, 0.5, and 0, respectively, so that, on average, the overall effect was the same as in the first scenario. The start of the event effect was generated completely at random from a uniform distribution for each study separately, so that the external event effect did not occur on the first or last measurement occasion. Data were generated using SAS.

Analysis

We had a total of 2^9 (= 512) experimental conditions. We simulated 400 replications of each condition, resulting in 204,800 data sets to analyze. We analyzed the data twice and compared the results. First, we combined the uncorrected effect sizes in the three-level meta-analysis. Next, we analyzed the

three-level data by estimating the corrected effect sizes, β_{2j} and β_{3j} , using the regression analysis per study (see Eq. 5) before combining them in the three-level meta-analysis (see Eqs. 2–4).

In the two approaches, we used the SAS proc MIXED (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006) procedure to estimate the participant-specific effect sizes, β_{2jk} and β_{3jk} . In the first approach, the effect sizes were uncorrected for the external event effect, whereas the effect sizes in the second approach were corrected.

SAS proc MIXED was also used for the three-level meta-analysis. The Satterthwaite approach to estimating the degrees of freedom method was applied because this method provides more accurate confidence intervals for estimates of the average treatment effect for two-level analyses of multiple-baseline data (Ferron et al., 2009).

In order to evaluate the appropriateness of both models, uncorrected and corrected for external events, we calculated the deviations of the estimated immediate treatment effect, $\hat{\gamma}_{200}$, from its population value, γ_{200} , and the deviations of the estimated treatment effect on the time trend, $\hat{\gamma}_{300}$, from its population value, γ_{300} . The mean deviation gives us an idea of the bias. Next, we calculated the mean squared deviation (the *mean squared error* [MSE]), which gives information about the variance of both estimated treatment effects ($\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$) around the corresponding population effect (γ_{200} and γ_{300}). Furthermore, we discuss the standard error and the 95 % confidence interval coverage proportion (CP) of the estimated immediate treatment effect and the treatment effect on the time trend. We also evaluate the bias of the point estimates of the between-study and between-case variances.

We used ANOVAs to evaluate whether there were significant effects ($\alpha = .01$) of each model type (e.g., model using effect sizes corrected vs. uncorrected for external event effects) and of the simulation design parameters (γ_{200} , γ_{300} , K , I , J , $\sigma_{u_2}^2$, $\sigma_{v_2}^2$) on the bias, the MSE, the standard error, and the CP.

Results of the simulation study

We present the results in two sections. In the first section, we discuss the constant external effect over four subsequent measurement occasions. The second section considers the case where the external effect gradually fades away over

Table 1 Time of introducing the treatment

I	Start of intervention						
	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5	Participant 6	Participant 7
15	5	6	7	8	9	11	13
30	5	8	11	14	17	20	23

four subsequent measurements. Each section presents the results of the three-level analysis of uncorrected and corrected effect sizes.

When there is no external event effect, the results of the three-level meta-analysis (i.e., bias in the fixed effects, *MSE* of the fixed effects, estimated standard errors of the fixed effects, *CP* for the fixed effects, and bias in the variance components) were found to be independent of the model type (corrected or not corrected for external events).

We found no significant bias for $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$ when using the corrected or uncorrected model. Therefore, we discuss the results of the analyses of the data including only external event effects conditions.

Constant external event over four subsequent measurement occasions

Overall treatment effect

Bias When we estimate γ_{200} and the effect sizes are uncorrected, the estimated treatment effect is, on average, significantly larger than the population value ($\gamma_{200} = 0$ or 2). Over all conditions, the bias equals 0.032, $t(51199) = 17.32$, $p < .0001$, whereas there is no significant bias for the corrected effect sizes (-0.0015), $t(51199) = -0.96$, $p = .34$. Table 2 presents the bias estimates for $\hat{\gamma}_{200}$, when $\gamma_{200} = 2$ and $\gamma_{300} = 0.2$.

Similar results are obtained for $\hat{\gamma}_{300}$. The bias is significantly negative for the uncorrected effect sizes and equals -0.20 , $t(51199) = -255.27$, $p < .0001$, whereas the bias is not significant for the corrected effect sizes, $t(51199) = -0.00020$, $p = .79$. Moreover, an analysis of variance on the deviations reveals a significant difference between the two different models, for both $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$ [$F(1, 102398) = 192.06$, $p < .0001$ for $\hat{\gamma}_{200}$, and $F(1, 102398) = 33,695.1$, $p < .0001$ for $\hat{\gamma}_{300}$]. The differences are largest when there is a small number of measurement occasions ($I = 15$) and studies ($K = 10$). In the following condition, the

largest difference was identified: $\gamma_{200} = 2$, $\gamma_{300} = 0$, $K = 10$, $I = 15$, $J = 4$, $\sigma_{u_2}^2 = 0.5$, and $\sigma_{v_2}^2 = 2$ (with a difference of 0.23).

MSE Similar to the bias, the *MSE* of the estimated treatment effect depends significantly on the model type; using an analysis of variance on the squared deviations, $F(1, 102398) = 882.77$, $p < .0001$ for $\hat{\gamma}_{200}$ and $F(1, 102398) = 7,076.91$, $p < .0001$ for $\hat{\gamma}_{300}$. When using the corrected model, the *MSE* for $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$ equals 0.12 and 0.028, respectively, whereas it is 0.18 and 0.070, respectively, for the uncorrected effect sizes. Differences between both models are larger if the number of observations and the number of studies are small (see Table 3 for $\hat{\gamma}_{200}$; similar results are obtained for $\hat{\gamma}_{300}$). So especially in these conditions, the modified model is recommended.

Estimates of the standard errors In order to evaluate inferences regarding the treatment effects, we constructed confidence intervals around the estimated treatment effects, $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$. Therefore, we needed to estimate the standard errors of the estimated treatment effects. Because we obtained 400 estimates of the effects in each condition, the standard deviations of the effect estimates can be regarded as a relatively good estimate of the standard deviation of the sampling distribution and can, therefore, be used as a criterion to evaluate the standard error. We looked at the relative standard error biases, which are the differences between the median standard error estimates and the standard deviation of the estimates of the effect divided by the standard deviation of the estimates of $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$. The relative differences are negative for $\hat{\gamma}_{200}$, which means that the median standard error estimates are smaller than expected. For $\hat{\gamma}_{300}$, these differences are positive, referring to median standard error estimates larger than expected. The relative standard error biases for both $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$ are, on average, larger across the conditions for the uncorrected effect sizes, in comparison with

Table 2 The bias of $\hat{\gamma}_{200}$ in the $\gamma_{200} = 2$, and $\gamma_{300} = 0.2$ conditions for the constant external event effect over four subsequent measurement occasions

K	J	$\sigma_{u_2}^2$	Corrected				Uncorrected			
			I = 15		I = 30		I = 15		I = 30	
			$\sigma_{v_2}^2 = 0.5$	$\sigma_{v_2}^2 = 2$	$\sigma_{v_2}^2 = 0.5$	$\sigma_{v_2}^2 = 2$	$\sigma_{v_2}^2 = 0.5$	$\sigma_{v_2}^2 = 2$	$\sigma_{v_2}^2 = 0.5$	$\sigma_{v_2}^2 = 2$
10	4	0.5	-0.003	0.007	0.025	-0.036	0.213	0.208	-0.027	0.027
		2	0.015	0.002	-0.017	0.014	0.129	0.196	0.012	0.035
	7	0.5	-0.026	-0.057	0.024	0.005	-0.093	-0.058	-0.019	-0.074
		2	-0.028	-0.015	-0.011	-0.003	-0.099	-0.060	-0.016	-0.026
30	4	0.5	0.009	0.028	0.004	-0.005	0.219	0.185	-0.008	0.013
		2	0.018	0.021	0.004	-0.011	0.210	0.222	0.008	0.035
	7	0.5	0.023	0.005	0.002	-0.009	-0.075	-0.105	-0.004	-0.016
		2	0.001	0.026	-0.006	-0.012	-0.077	-0.088	-0.003	0.006

“Corrected” and “uncorrected” refer, respectively, to corrected effect size and uncorrected effect size for external event effects

Table 3 The *MSE* of $\hat{\gamma}_{200}$ in the $\gamma_{200} = 2$, and $\gamma_{300} = 0.2$ conditions for the constant external event effect over four subsequent measurement occasions

<i>K</i>	<i>J</i>	$\sigma_{u_2}^2$	Corrected				Uncorrected			
			<i>I</i> = 15		<i>I</i> = 30		<i>I</i> = 15		<i>I</i> = 30	
			$\sigma_{v_2}^2 = 0.5$	$\sigma_{v_2}^2 = 2$	$\sigma_{v_2}^2 = 0.5$	$\sigma_{v_2}^2 = 2$	$\sigma_{v_2}^2 = 0.5$	$\sigma_{v_2}^2 = 2$	$\sigma_{v_2}^2 = 0.5$	$\sigma_{v_2}^2 = 2$
10	4	0.5	0.17	0.28	0.11	0.26	0.32	0.43	0.14	0.25
		2	0.20	0.32	0.14	0.28	0.31	0.49	0.16	0.36
	7	0.5	0.09	0.24	0.07	0.23	0.18	0.31	0.09	0.22
		2	0.11	0.26	0.09	0.24	0.20	0.31	0.09	0.28
30	4	0.5	0.06	0.10	0.04	0.09	0.14	0.19	0.04	0.10
		2	0.06	0.11	0.04	0.09	0.15	0.20	0.05	0.12
	7	0.5	0.03	0.07	0.03	0.08	0.06	0.10	0.03	0.09
		2	0.04	0.08	0.03	0.08	0.07	0.10	0.04	0.08

“Corrected” and “uncorrected” refer, respectively, to corrected effect size and uncorrected effect size for external event effects

the corrected effect sizes. For $\hat{\gamma}_{200}$, the average relative standard error biases equal -1.8% and -2.0% for the corrected and uncorrected models, respectively. The average relative standard error biases difference for $\hat{\gamma}_{300}$ for the uncorrected model is 2% , whereas it is substantial (more than 10% ; Hoogland & Boomsma; 1998) for the uncorrected model (25.7%). So the difference between the model types becomes more apparent when γ_{300} is estimated, $F(1, 254) = 38.9$, $p < .0001$. The conditions with the largest relative standard error bias when the uncorrected model for $\hat{\gamma}_{300}$ was used tended to coincidence with the conditions where 30 studies, an immediate treatment effect of 2, and a treatment effect on the time trend of 0.2 were involved, with the bias mounting to 107% in the condition where $\gamma_{200} = 2$, $\gamma_{300} = 0.2$, $K = 30$, $J = 7$, $I = 30$, $\sigma_{v_2}^2 = 0.5$, and $\sigma_{u_2}^2 = 0.5$.

Coverage proportion We estimated the *CP* of the 95% confidence intervals, which allowed us to evaluate the interval estimates of $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$. The confidence intervals were estimated by using the standard errors and the Satterthwaite estimated degrees of freedom. The *CP* of these confidence intervals was estimated for each of the combinations. A positive significant difference between the corrected model and the uncorrected model in the *CP* is found for $\hat{\gamma}_{200}$, $F(1, 254) = 27.56$, $p < .0001$ (see Table 4). Also, for $\hat{\gamma}_{300}$, the mean *CP* depends significantly on the model type, $F(1, 254) = 20.96$, $p < .0001$ (see Table 4). The conditions with a *CP* less than $.93$ all have 15 measurements in common and occur when the effect sizes are uncorrected, for both $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$. Moreover, for $\hat{\gamma}_{300}$, the *CP* is not only too small when $I = 15$ and $K = 30$, but also too

Table 4 The coverage proportion of $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$ in the $\gamma_{200} = 2$, $\gamma_{300} = 0.2$, and $\sigma_{u_2}^2 = 2$ conditions for the constant external event effect over four subsequent measurement occasions

<i>K</i>	<i>J</i>	$\sigma_{v_2}^2$	$\hat{\gamma}_{200}$				$\hat{\gamma}_{300}$			
			Corrected		Uncorrected		Corrected		Uncorrected	
			<i>I</i> = 15	<i>I</i> = 30	<i>I</i> = 15	<i>I</i> = 30	<i>I</i> = 15	<i>I</i> = 30	<i>I</i> = 15	<i>I</i> = 30
10	4	0.5	.96	.96	.96	.96	.94	1.00	.97	1.00
		2	.95	.95	.92	.95	.96	.98	.93	1.00
	7	0.5	.96	.95	.94	.97	.99	1.00	.97	1.00
		2	.97	.96	.95	.95	.96	.97	.84	.99
30	4	0.5	.97	.96	.89	.97	.97	1.00	.90	1.00
		2	.97	.96	.91	.94	.96	.98	.49	1.00
	7	0.5	.94	.94	.92	.96	.98	1.00	.93	1.00
		2	.96	.95	.96	.96	.96	.97	.26	.96

Values smaller than $.93$ and larger than $.97$ appear in bold. “Corrected” and “uncorrected” refer, respectively, to corrected effect size and uncorrected effect size for external event effects

large when $I = 30$ (values for the CP range from .99 to 1.00). When the effect sizes are uncorrected, the CP is well estimated when $I = 30$ for $\hat{\gamma}_{200}$ and $I = 15$ and $K = 10$ for $\hat{\gamma}_{300}$. The difference in CP for $\hat{\gamma}_{200}$ is largest when there is only a small number of measurements ($I = 15$) and a large number of studies ($I = 30$).

Variance components

In the three-level analyses, the between-study and between-case variances were estimated for both the immediate treatment effect and the treatment effect on the trend. Because variance estimates are expected to be positively skewed, due to truncation of negative estimates to zero, we calculated the median (relative) deviation of the estimates from the population value, rather than the mean (relative) deviation, to evaluate the (relative) bias in the estimates. We discuss only the between-case variance and the between-study variance of the immediate treatment effect ($\sigma_{u_2}^2$ and $\sigma_{v_2}^2$), because similar results are obtained for the treatment effect on the time trend ($\sigma_{u_3}^2$ and $\sigma_{v_3}^2$). The bias of the estimated between-study variance and the estimated between-case variance of the immediate effect is larger when there are only 10 studies and 15 measurement occasions involved. The conditions with the largest relative bias all had 15 measurements, 4 participants, and a small between-study variance ($\sigma_{v_2}^2 = 0.5$) in common. If the effect sizes are corrected and we estimate the between-study variance of the immediate treatment effect, we find relative parameter bias values across conditions ranging from 17 % to 55 %, while the relative bias goes up to a value of 313 % when the effect sizes are uncorrected. Similar results are found for $\hat{\sigma}_{u_2}^2$, where the relative bias in a condition is maximum 119 % for the corrected effect sizes and 326 % for the uncorrected effect sizes (see Table 5). Overall, the adjusted model results in less biased variance estimates.

External event fades away gradually over four subsequent measurement occasions

Overall treatment effect

Bias The bias of $\hat{\gamma}_{200}$ for uncorrected and corrected effect sizes is, respectively, -0.0073 , $t(51199) = -418$, $p < .001$, and 0.00057 , $t(51199) = 38$, $p = .74$. This means that there is a significant negative bias for the uncorrected effect sizes, whereas this is not the case for the corrected effect sizes, and the models differ significantly, $F(1, 102398) = 0.009$, $p = .77$. The bias for $\hat{\gamma}_{300}$ depends largely on the model type, $F(1, 102398) = 30,476.1$, $p < .0001$. The bias for the uncorrected effect sizes is significant (-0.19), $t(51199) = -246.23$, $p < .0001$, whereas this is not the case for the corrected

Table 5 Median of relative deviation of the variance estimates of γ_{200} in the $\gamma_{200} = 2$, and $\gamma_{300} = 0.2$ conditions for the constant external event effect over four subsequent measurement occasions

		$\hat{\sigma}_{u_2}^2$											
		Corrected						Uncorrected					
		Corrected			Uncorrected			Corrected			Uncorrected		
		$K = 10$			$K = 30$			$K = 10$			$K = 30$		
		$\sigma_{u_2}^2 = 0.5$	$\sigma_{u_2}^2 = 2$	$\sigma_{u_2}^2 = 0.5$	$\sigma_{u_2}^2 = 2$	$\sigma_{u_2}^2 = 0.5$	$\sigma_{u_2}^2 = 2$	$\sigma_{u_2}^2 = 0.5$	$\sigma_{u_2}^2 = 2$	$\sigma_{u_2}^2 = 0.5$	$\sigma_{u_2}^2 = 2$	$\sigma_{u_2}^2 = 0.5$	$\sigma_{u_2}^2 = 2$
I	J	$\sigma_{v_2}^2$	$\sigma_{u_2}^2$	$\sigma_{v_2}^2$	$\sigma_{u_2}^2$	$\sigma_{v_2}^2$	$\sigma_{u_2}^2$	$\sigma_{v_2}^2$	$\sigma_{u_2}^2$	$\sigma_{v_2}^2$	$\sigma_{u_2}^2$	$\sigma_{v_2}^2$	$\sigma_{u_2}^2$
15	4	0.5	0.47	0.61	0.55	0.58	0.58	2.7	3.09	3.11	2.91	0.24	1.19
		2	0.05	-0.03	0.13	0.12	0.12	0.71	0.57	0.74	0.65	0.20	1.19
	7	0.5	-0.07	-0.11	0.09	0.06	0.06	0.99	0.99	1.10	0.98	0.25	1.20
		2	-0.02	-0.08	0.01	-0.02	-0.02	0.15	0.16	0.29	0.28	0.27	1.20
30	4	0.5	-0.12	-0.09	-0.02	0.05	0.05	0.33	0.26	0.45	0.48	0.21	0.80
		2	-0.08	-0.06	-0.01	-0.04	-0.04	0.14	0.10	0.08	0.10	0.16	0.85
	7	0.5	-0.17	-0.12	0.02	-0.03	-0.03	-0.08	-0.07	0.04	0.04	-0.06	-0.10
		2	-0.11	-0.12	-0.01	-0.04	-0.04	-0.07	-0.02	0.00	0.00	-0.05	-0.12

“Corrected” and “uncorrected” refer, respectively, to corrected effect size and uncorrected effect size for external event effects

effect sizes (0.000179), $t(51199) = 0.24$, $p = .81$. For both $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$, the difference is largest when there are a small number of measurements ($I = 15$) involved.

MSE For both estimated treatment effects, the *MSEs* are larger for the uncorrected effect sizes, in comparison with the corrected effect sizes (see Table 6). For both $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$, the model type has a significant influence on the *MSE*, $F(1, 102398) = 724.69$, $p \leq .0001$ for $\hat{\gamma}_{200}$, and for $\hat{\gamma}_{300}$, $F(1, 102398) = 5,431.15$, $p < .0001$. For both estimated treatment effects, the *MSE* is large when the studies are heterogeneous ($\sigma_{v_2}^2 = 2$) and a small number of measurement occasions ($I = 15$) and studies ($K = 10$) are used. The difference between the models is largest when a small number of measurements is used.

Estimates of the standard errors The difference between the average relative bias in the standard errors of the uncorrected effect sizes equals 0.02 for both uncorrected and corrected effect sizes when γ_{200} is estimated.

Similar to the constant external event effect results, the difference between the average relative bias in the standard errors of the uncorrected effect sizes ($M = 39.3$) and corrected effect sizes (0.06) for $\hat{\gamma}_{300}$ is larger and statistically significant, $F(1, 254) = 129.66$, $p = .0001$ (see Table 7). The difference in results due to the model type is more obvious if there are a small number of studies involved ($K = 10$).

Coverage proportion Similar to the *CP* for the constant external event effect, the mean *CP* for the uncorrected and corrected effect sizes for the estimate of the immediate treatment effect differ significantly at the 5 % significance level for both $\hat{\gamma}_{200}$, $F(1, 254) = 3.92$, $p = .05$, and $\hat{\gamma}_{300}$, $F(1,$

Table 7 Difference between the median of the standard error estimates and the standard deviation of $\hat{\gamma}_{300}$ in the $\gamma_{200} = 2$, $\gamma_{300} = 0.2$, and $\sigma_{u_3}^2 = 0.05$ conditions for the external event effect fading away gradually over four subsequent measurement occasions

<i>K</i>	<i>J</i>	$\sigma_{v_3}^2$	Corrected		Uncorrected	
			<i>I</i> = 15	<i>I</i> = 30	<i>I</i> = 15	<i>I</i> = 30
10	4	0.05	0.010	0.037	0.076	0.103
		0.2	−0.031	−0.002	0.029	0.050
	7	0.05	0.004	0.035	0.069	0.068
		0.2	−0.001	−0.007	0.010	0.012
30	4	0.05	−0.018	0.022	0.039	0.061
		0.2	−0.009	0.0003	0.022	0.024
	7	0.05	0.002	0.021	0.038	0.040
		0.2	−0.002	0.001	0.004	0.003

“Corrected” and “uncorrected” refer, respectively, to corrected effect size and uncorrected effect size for external event effects

254) = 3.25, $p = .007$. The *CP* with values smaller than .93 all have 15 measurement occasions, have a large between-study variance ($\sigma_{v_3}^2 = 2.0$), and occur when the effect sizes are uncorrected (for both $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$). Similar to the constant external event effect, the *CP* is overestimated for $\hat{\gamma}_{300}$ and when the effect sizes are uncorrected in the condition where 30 measurement occasions are included. In the condition where $I = 15$ and $\sigma_{v_3}^2 = 2.0$, the difference between corrected and uncorrected effect sizes is largest.

Variance components

The results are similar to the results of the constant external event effect, and results are less biased using the adjusted

Table 6 *MSE* of $\hat{\gamma}_{200}$ and $\hat{\gamma}_{300}$ in the $\gamma_{200} = 2$, $\gamma_{300} = 0.2$, and $\sigma_{u_2}^2 = 0.5$ conditions for the external event effect fading away gradually over four subsequent measurement occasions

<i>K</i>	<i>J</i>	$\sigma_{v_2}^2$	$\hat{\gamma}_{200}$		$\hat{\gamma}_{300}$		$\hat{\gamma}_{200}$		$\hat{\gamma}_{300}$	
			Corrected		Uncorrected		Corrected		Uncorrected	
			<i>I</i> = 15	<i>I</i> = 30	<i>I</i> = 15	<i>I</i> = 30	<i>I</i> = 15	<i>I</i> = 30	<i>I</i> = 15	<i>I</i> = 30
10	4	0.5	0.14	0.11	0.34	0.12	0.09	0.01	0.12	0.01
		2	0.31	0.24	0.47	0.27	0.13	0.03	0.14	0.03
	7	0.5	0.09	0.06	0.18	0.09	0.01	0.01	0.10	0.01
		2	0.22	0.23	0.32	0.22	0.03	0.02	0.12	0.02
30	4	0.5	0.05	0.03	0.11	0.04	0.04	0.004	0.11	0.01
		2	0.11	0.09	0.17	0.09	0.04	0.01	0.12	0.01
	7	0.5	0.03	0.02	0.06	0.02	0.004	0.002	0.09	0.01
		2	0.08	0.08	0.10	0.07	0.01	0.01	0.1	0.01

“Corrected” and “uncorrected” refer, respectively, to corrected effect size and uncorrected effect size for external event effects

model. We only discuss the estimated variances for the immediate treatment effect, because the results are similar for the estimated treatment effect on the trend. When we estimate the between-study variance and the effect sizes are uncorrected, the bias ranges from -0.002 to 3.41 , while it ranges from 0.002 to 0.73 for the corrected effect sizes. So the estimated variances depend on the model type, $F(1, 102398) = 1,631$, $p < .0001$. Similar results are obtained for the estimate of the between-case variance. The maximum bias for the corrected effect sizes is 1.60 , while it is 3.21 for the uncorrected effect sizes, and these estimates depend on the model type, $F(1, 102398) = 5,628.62$, $p < .0001$.

Empirical illustration

In this section, we give empirical illustrations of the comparison of the modified three-level model in which external events are taken into account with the uncorrected model. Therefore, we used a part of the meta-analytic data set of Heyvaert, Saenen, Maes, and Onghena (2012) in which restraint interventions for challenging behavior among persons with intellectual disabilities was investigated. We give two empirical illustrations of the consequences of ignoring the external event effect in a multiple-baseline across-participants design. We illustrate first the consequences of ignoring external events in a single study, and next the consequences of ignoring external events in a three-level meta-analysis.

Ignoring external events in a single study

To illustrate the regression analysis of a multiple-baseline across-3-participants design, we use the study of Thompson, Iwata, Conners, and Roscoe (1999), which was included in the meta-analysis of Heyvaert et al. (2012). In their study, the effects of benign punishment on the self-injurious behavior of individuals who had been diagnosed with mental retardation was investigated. The 3 participants were measured repeatedly over time on 22 measurement occasions, and the intervention started on sessions 11, 13, and 20, respectively (see Fig. 2). From this figure, we might expect that there is an immediate reduction in challenging behavior when the treatment is introduced and that the effect of the treatment on the challenging behavior decreases over time (so there is a positive effect on the time trend during the treatment). We also see that the 3 participants' scores on measurement occasions 4 and 10 are possibly influenced by an external event.

Results

If we ignore possible external events in the regression analysis before combining the effect sizes in the two-level meta-

Fig. 2 Graphical display of a multiple-baseline design across-3-participants designs using data from the study of Thompson, Iwata, Conners, and Roscoe (1999)

analyses, the average immediate treatment effect over cases for that study equals -25.58 , and the average treatment effect on the time trend over cases from that study equals -2.58 . If we take the external event into account by correcting the effect sizes before combining them, the immediate treatment effect equals -23.23 , and the treatment effect on the time trend is 1.24 . This means that $\hat{\gamma}_{200}$ is 9.19% smaller when the effect sizes are corrected, in comparison with the uncorrected effect sizes. Moreover $\hat{\gamma}_{300}$ is positive for the corrected effect sizes, whereas it is negative for the uncorrected, which means that the effect of the treatment over time decreases for the corrected effect sizes, whereas it increases for the uncorrected.

Ignoring external events in a three-level meta-analysis

The three-level analysis of SSED data includes summarizing the immediate treatment effect and the treatment effect on the time trend over participants and over studies.

We estimate the immediate treatment effect and the treatment effect on the time trend across seven studies. Again, we use the meta-analysis of Heyvaert et al. (2012) to randomly select multiple-baseline across-participants studies. We combined the multiple-baseline across-participants

study of Lindberg, Iwata, and Kahng (1999), Chung and Cannella-Malone (2010), Zhou, Goff, and Iwata (2000), Thompson et al., (1999), Hanley, Iwata, Thompson, and Lindberg (2000), Rolider, Williams, Cummings, and Van Houten (1991), and Roscoe, Iwata, and Goh (1998). In all these studies, the same dependent variable was measured—namely, the reduction in self-injurious behavior. Again, we compare the three-level meta-analysis of uncorrected and corrected effect sizes.

Results

With the uncorrected effect sizes in the three-level meta-analysis, the overall immediate treatment effect equals -33.14 , $t(6.39) = -3.44$, $p = .012$, and the overall treatment effect on the time trend equals -4.42 , $t(3.95) = -1.52$, $p = .19$. When the effect sizes are corrected before estimating the effects over participants, the immediate treatment effect equals -21.07 , $t(6.88) = -1.13$, $p = .30$, and the treatment effect on the time trend equals -0.43 , $t(1) = -0.28$, $p = .83$. This means that the immediate treatment effect of the corrected effect sizes is 36.42 % smaller, as compared with the uncorrected effect size, and the treatment effect on the time trend for the corrected effect size during the treatment is 90.27 % smaller.

This is consistent with the results of the simulation study, where we found that the estimated treatment effects are biased when the effect sizes are uncorrected before combining them in the three-level meta-analysis.

Discussion

External event effects are common in SSEDs because single-case researchers often implement these kinds of designs in everyday scenarios where they cannot control for outside factors (Christ, 2007; Kratochwill et al., 2010; Shadish et al., 2002). External events are not always anticipated by researchers, and thus, they may not be measured during the conduct of the study. Furthermore, the size of an event effect may be small, and researchers may be unaware of it even after the study has been completed. Whether researchers recognize an external event or not, the failure to account for the event in a meta-analysis can bias the estimate of the treatment effect. Thus, we searched for a method with which to model external events that could be applied even when the events had not been previously identified. Because we used a multiple-baseline across-participants design, there was a need to take into account the interdependence of the participants. Therefore, an external event that influenced the scores of 1 participant was assumed to influence the scores of the other participants in the same study.

We discussed two possible scenarios. In one scenario, the external event effect remains constant and influences the scores of all participants within a study on four subsequent moments. This occurs, for example, when a teacher is ill and a substitute teacher takes over the classroom or when a foreign observer is present on subsequent measurement occasions. In the second scenario, the external event's effect would likely gradually fade away over four subsequent moments. For instance, the influence of a teacher intern on the behavior of students reduces over time. Moreover, the model adjusted for external event effects takes into account that measurement occasions closer in time are more related than measurement occasions further in time.

We evaluated this approach using a large simulation study and gave some empirical examples. If there is an external event effect of zero, both models (the one that corrects for moment effects and the one that does not) are appropriate. If the external event influences subsequent scores for all the participants within a study, the three-level approach for uncorrected effect sizes is not recommended, because the estimates of both treatment effects (i.e., immediate effect on level and effect on time trend) are substantially biased. The *MSE*, standard error, and *CP* are better estimated when the modified model, which includes moment effects, is used. The difference between the corrected and uncorrected effect sizes is largest when there are a small number of studies and measurement occasions, so in this context, we advise using the adjusted model. Moreover, the adjusted model results in less biased variance estimates.

But, of course, we should be aware of some limitations. We assumed that all the participants within a study are influenced the same way by the external event effect. It is possible that different participants from the same study are at separate locations and, therefore, are not all influenced by the external event. Modeling event effects that are not common to all participants in a study is an important avenue for future research. We chose to keep the number of measurements within a study constant for all participants within the same study. Of course, it is possible that different participants of the same study have different series lengths. Furthermore, we cannot generalize these results to other conditions not involved in this simulation study, but we partially addressed this by simulating a large number of conditions and choosing realistic values for the parameters.

Another limitation is that we assumed linear trajectories in the treatment phase, which might not be true in some real situations. To simplify the simulation model, we further did not account for a possible dependence between regression coefficients, which can be accounted for in a multilevel analysis by estimating the covariance at the various levels.

In addition, subjects in MBDs are repeatedly measured, and succeeding measurements may be more related to each other than measurements further away in time. We did not account for this possible autocorrelation and suggest that this as a useful extension to the present study.

Kazdin (2010) argued that there needs to be a minimum of three measurement occasions between the participants in an MBD in order to show an experimental effect. We did not take this into account in the condition where the number of measurement occasions was 15, because it was not possible to do this and provide each of 7 participants a unique baseline. We could alter the intervention schedule to introduce the treatment for some participants (e.g., randomly selected pairs) at the same moment. Examining this strategy specifically and alternative intervention schedules more generally would allow further research to extend results to a wider range of multiple-baseline applications.

It can be difficult to attribute simultaneously unusual outcome scores for all participants within a study to an external event effect. If there is no external event effect, we can still use the corrected model, because both the corrected and uncorrected effect sizes will be unbiased and, thus, there is no need to identify before the analyses whether an external event effect occurred or not. We advise single-case researchers to first use both models in the sensitivity analysis and then decide which model to use. If researchers are interested in the occurrence of external event effects, we recommend that they keep a log in order to identify potential outside factors that may influence the scores at certain measurement occasions and include dummy indicator variables at least for these moments.

The extension of the three-level model for multiple-baseline across-participants designs to include modeling of potential external effects makes it even more appropriate and useful for the analysis of realistic SSED data sets. This study has indicated that the three-level model corrected for external event effects provides better results than does the uncorrected model for combining results from multiple-baseline across-participants data, especially if there is only a small number of observations ($I = 15$) and a small number of studies ($K = 10$) in the synthesis. As was found here, even when an external event effect is small, a failure to correct for it can lead to biased effect sizes. Thus, applied SSED researchers are encouraged to consider use of the three-level model that corrects for external event effects when synthesizing results of MBD data.

Author Note Mariola Moeyaert, Faculty of Psychology and Educational Sciences, University of Leuven, Belgium; Maaïke Ugille, Faculty of Psychology and Educational Sciences, University of Leuven, Belgium; John M. Ferron, Department of Educational Measurement and Research, University of South Florida. S. Natasha Beretvas, Department of Educational Psychology, University of Texas; Wim Van den Noortgate, Faculty

of Psychology and Educational Sciences, ITEC-IBBT Kortrijk, University of Leuven, Belgium.

This research is funded by the Institute of Education Sciences, U.S. Department of Education, Grant R305D110024. 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 Flemish Supercomputer Center, financed by the Department of Economy, Science and Innovation–Flemish Government and the Hercules Foundation.

References

- Alen, E., Grietens, H., & Van den Noortgate, W. (2009). *Meta-analysis of single-case studies: An illustration for the treatment of anxiety disorders*. Unpublished manuscript.
- Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91–97.
- Barlow, D. H., & Hersen, M. (1984). *Single-case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.
- 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.
- Christ, T. J. (2007). Experimental control and threats to internal validity of concurrent and nonconcurrent multiple-baseline designs. *Psychology in the Schools, 44*, 451–459.
- Chung, Y.-C., & Cannella-Malone, H. I. (2010). The effects of pre-session manipulations on automatically maintained challenging behavior and task responding. *Behavior Modification, 34*, 479–502. doi:10.1177/0145445510378380
- Cooper, H. M. (2010). *Research synthesis and meta-analysis: a step-by-step approach*. 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.
- 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.
- Ferron, J., & Scott, H. (2005). Multiple baseline designs. In B. Everitt & D. Howell (Eds.), *Encyclopedia of Behavioral Statistics* (Vol. 3, pp. 1306–1309). West Sussex, UK: Wiley & Sons Ltd.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3–8.
- Hanley, G. P., Iwata, B. A., Thompson, R. H., & Lindberg, J. S. (2000). A component analysis of "stereotypy as reinforcement" for alternative behavior. *Journal of Applied Behavior Analysis, 33*, 285–297.
- Heyvaert, M., Saenen, L., Maes, B., & Onghena, P. (2012). *Systematic review of restraint interventions for challenging behaviour among persons with intellectual disabilities: Experiences and effectiveness*. Manuscript submitted for publication.
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. An overview and a meta-analysis. *Sociological Methods & Research, 26*, 329–367.

- Kazdin, A. E. (2010). *Single-case Research Designs: Methods for Clinical and Applied Settings* (2nd ed.). New York: Oxford University Press.
- Kinugasa, T., Cerin, E., & Hooper, S. (2004). Single-Subject Research Designs and Data Analyses for Assessing Elite Athletes' Conditioning. *Sports Medicine*, 34, 1035–1050.
- Koehler, M. J., & Levin, J. R. (2000). RegRand: Statistical software for the multiple-baseline design. *Behavior Research Methods, Instruments, & Computers*, 32, 367–371.
- 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.
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf
- Lindberg, J. S., Iwata, B. A., & Kahng, S. W. (1999). On the relation between object manipulation and stereotypic self-injurious behavior. *Journal of Applied Behavior Analysis*, 32, 51–62.
- 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 Inc.
- Moeyaert, M., Ugille, M., Ferron, J., Beretvas, N., & Van den Noortgate, W. (2012a). Three-level analysis of single-case experimental data: Empirical validation. *Journal of Experimental Education*.
- Moeyaert, M., Ugille, M., Ferron, J., Beretvas, N., & Van den Noortgate, W. (2012b) *The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study*. Manuscript submitted for publication.
- Onghena, P. (2005). Single-case designs. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 4, pp. 1850–1854). Chichester: Wiley.
- Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *The Clinical Journal of Pain*, 21, 56–68.
- Owens, C. M., & Ferron, J. M. (2012). Synthesizing single-case studies: A Monte Carlo examination of a three-level meta-analytic model. *Behavior Research Methods*.
- Rolider, A., Williams, L., Cummings, A., & Van Houten, R. (1991). The use of a brief movement restriction procedure to eliminate severe inappropriate behavior. *Journal of Behavior Therapy and Experimental Psychiatry*, 22(1), 23–30. doi:10.1016/0005-7916(91)90029-5
- Roscoe, E. M., Iwata, B. A., & Goh, H.-L. (1998). A comparison of noncontingent reinforcement and sensory extinction as treatments for self-injurious behavior. *Journal of Applied Behavior Analysis*, 31, 635–646.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.
- Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, 971–980.
- Shogren, K. A., Faggella-Luby, M. N., Bae, S. J., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior: A meta-analysis. *Journal of Positive Behavior Interventions*, 4, 228–237.
- Thompson, R. H., Iwata, B. A., Connors, J., & Roscoe, E. M. (1999). Effects of reinforcement for alternative behavior during punishment of self-injury. *Journal of Applied Behavior Analysis*, 32, 317–328.
- Van den Noortgate, W., & Onghena, P. (2003). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, 18, 325–346.
- 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 Disorder*, 5, 562–569.
- Zhou, L. M., Goff, G. A., & Iwata, B. A. (2000). Effects of increased response effort on self-injury and object manipulation as competing responses. *Journal of Applied Behavior Analysis*, 33(1), 29–40.