Article

Journal of Educational and Behavioral Statistics
2022, Vol. 47, No. 2, pp. 131–166
DOI: 10.3102/10769986211035507
Article reuse guidelines: sagepub.com/journals-permissions

Comparison of Within- and Between-Series Effect Estimates in the Meta-Analysis of Multiple Baseline Studies

Seang-Hwane Joo
The University of Kansas

Yan Wang
University of Massachusetts Lowell

John Ferron
University of South Florida

S. Natasha Beretvas
The University of Texas at Austin

Mariola Moeyaert
The State University of New York

Wim Van Den Noortgate
Katholieke Universiteit Leuven

Multiple baseline (MB) designs are becoming more prevalent in educational and behavioral research, and as they do, there is growing interest in combining effect size estimates across studies. To further refine the meta-analytic methods of estimating the effect, this study developed and compared eight alternative methods of estimating intervention effects from a set of MB studies. The methods differed in the assumptions made and varied in whether they relied on within- or between-series comparisons, modeled raw data or effect sizes, and did or did not standardize. Small sample functioning was examined through two simulation studies, which showed that when data were consistent with assumptions the bias was consistently less than 5% of the effect size for each method, whereas root mean squared error varied substantially across methods. When assumptions were violated, substantial biases were found. Implications and limitations are discussed.

Keywords: *multiple baseline design; single-case; mixed linear model; meta-analysis; standardized effect size*

Multiple baseline (MB) designs are a type of single-case experimental design that includes multiple time series, typically stemming from multiple cases. For case j (of the J cases) in a study, the dependent variable is repeatedly measured I_j times including A_j baseline observations and $I_j - A_j$ treatment observations. The transitions between the baseline and treatment phase are temporally staggered across the J cases (i.e., $A_j \neq A_{j'}$ for pair of cases $j \neq j'$; Baer et al., 1968). MB designs are becoming more prevalent in behavioral research and education research, and as they do, there is growing interest in estimating treatment effects and meta-analyzing the results from these designs (e.g., Hedges et al., 2013; Pustejovsky et al., 2014; Van den Noortgate & Onghena, 2003a, 2003b, 2008).

Estimation of a raw score effect size (ψ_t) at focal time t, or a corresponding standardized effect size (δ_t), often relies on the application of a mixed linear model. Typically, the model is based on a within-case comparison conceptualization that relies on assumptions about the form of the growth trajectories in the baseline and treatment phases and on the extrapolation of the baseline trajectories. As a consequence, the treatment effect estimates are susceptible to bias when there is misspecification of the form of the trajectory (e.g., specification of a linear trend when growth is nonlinear) or when the errors in the time series are not stationary, such as when some unmeasured variable leads to a shift in the level of the time series (Ferron et al., 2014). These concerns become more pronounced when observation series are short and thus the data available to evaluate competing models are limited. Concerns with potential bias in the treatment effect estimates have led to the consideration of different estimators of ψ_t that focus on contrasting the responses of the cases that are in treatment to the responses of the cases that are still in baseline (Ferron et al., 2014). Henceforth, we refer this approach as the *between-series estimators* of ψ_t because the estimator compares observations between the treatment and baseline phases across participants at certain time points.

The purpose of this study is to develop and compare alternative methods of estimating the average treatment effect in the meta-analysis of MB studies. We will consider eight approaches that differ in whether the effect estimates are based on within- or between-series comparisons, whether individual participant data or effect sizes are meta-analyzed, and whether effects are or are not standardized (see Figure 1). For the methods based on within-case comparisons, our methods involve just minor adaptions to methods that have already been examined. The methods based on between-case comparisons are more novel because neither meta-analytic effect estimates nor standardized effect estimates have been previously developed. Because approximate small sample size adjustments will be used in our methods, simulations will be used to compare the alternative methods of estimating ψ_t and δ_t in terms of parameter bias and root mean squared error (RMSE).

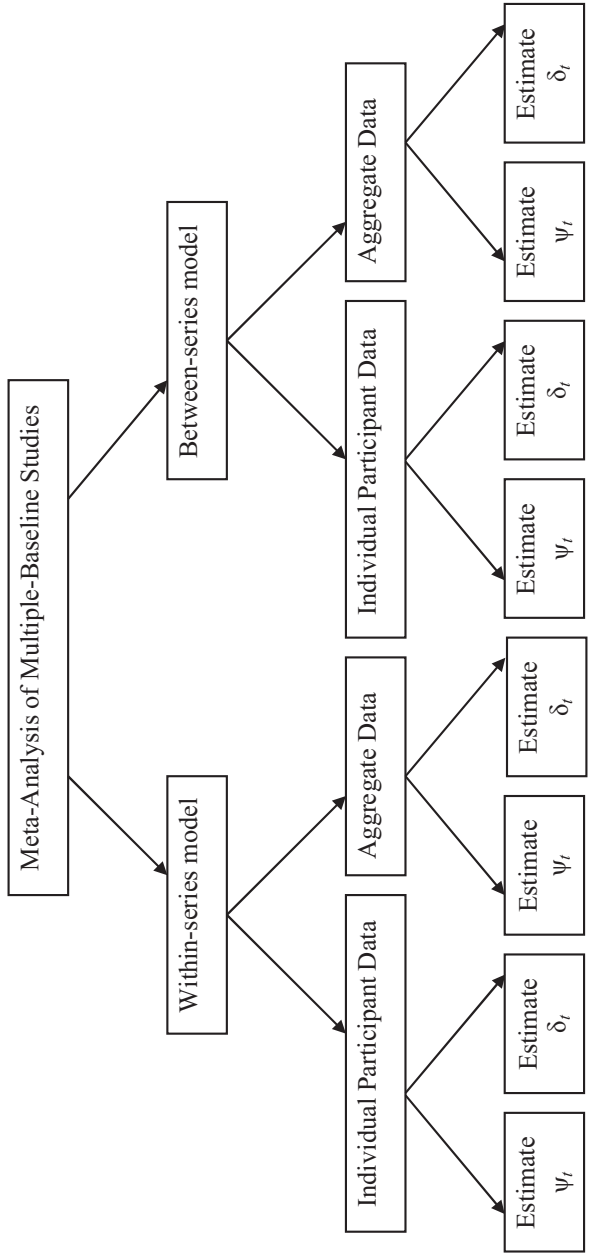

FIGURE 1. *Diagram of the eight meta-analytic methods for estimating effects* ψ_t *and* δ_t.

To illustrate the difference between the within-series and between-series approaches, see Table 1. All values of the MB study would be used in the within-series approach, which focuses on contrasting the treatment values (B) to the baseline values (A) within each row. The between-series approach differs in that it uses only the observations from the between-subject subexperiments, those that are enclosed in boxes in Table 1, and then contrasts the observations of those in treatment for some specific amount of time to the observations at the same time point of those individuals who are still in baseline (i.e., the B values to the A values within the same column of the enclosed boxes in Table 1).

More formally, the design matrix for the fixed effects, $\mathbf{X}$, from the mixed linear model equation, $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zv} + \boldsymbol{\epsilon}$, is defined with two sets of dummy variables. Suppose there are $l = 1, \ldots, L$ subexperiments for the MB design (e.g., $L = 2$ in Table 1, each subexperiment is enclosed within a box) and $m = 1, \ldots, M$ observations after the start of treatment in the L subexperiments (e.g., if interested in the effect at one, two, and three observations into treatment, then $M = 3$, as in Table 1). Then, there are LM indicator variables included to gauge the mean baseline level in each of the L subexperiments at each of M times of interest following treatment. Second, there are M treatment indicator variables to index the effect at each of the M points in time after the start of treatment, where the treatment effects are assumed the same across subexperiments.

If participants are randomly assigned to conditions with shorter versus longer baselines, the between-subject subexperiments are randomized experiments and thus the treatment effect can be shown to be unbiased with fewer assumptions than needed when using the within-series approach. Specifically, the between-series comparison approach does not rely on assumptions about the form of growth trajectories or extrapolation, and thus unbiased treatment effect estimates can be obtained across a broader range of contexts. However, the between-series estimates of ψ_t are less precise and if the participants are not randomly assigned to baselines the biases may be more severe (Ferron et al., 2014).

To get more precise estimates of ψ_t and to capitalize on the increase in the prevalence of MB studies, there is growing interest in combining estimates across studies. When all studies use the same dependent variables measured on the same scale (Moeyaert, Ferron, et al., 2014), the meta-analytic extensions of the estimators for ψ_t are relatively straightforward. Meta-analytic estimators of ψ_t based on the within-series comparison approach have been proposed and studied, whereas meta-analytic estimators of ψ_t based on the between-series comparison approach have not been previously developed. In either case, ψ_t could be directly estimated from a mixed linear model of the individual participant data (for the within-series comparison approach, see Moeyaert, Ugille, et al., 2014; Owens & Ferron, 2012; Van den Noortgate & Onghena, 2008) or could be estimated using a traditional meta-analytic approach where effect estimates are first obtained for each individual or each study and then these estimates are combined within and across studies (for the within-series comparison approach,

TABLE 1.
Schematic Diagram of Observations in a Six Case Multiple Baseline Design

${}^{-5}A_{1,1}$	${}^{-4}A_{2,1}$	${}^{-3}A_{3,1}$	${}^{-2}B^{1,1}_{4,1}$	${}^{-1}B^{2,1}_{5,1}$	${}^{0}B^{3,1}_{6,1}$	${}^{1}B_{7,1}$	${}^{2}B_{8,1}$	${}^{3}B_{9,1}$	${}^{4}B_{10,1}$	${}^{5}B_{11,1}$	${}^{6}B_{12,1}$
${}^{-5}A_{1,2}$	${}^{-4}A_{2,2}$	${}^{-3}A_{3,2}$	${}^{-2}B^{1,1}_{4,2}$	${}^{-1}B^{2,1}_{5,2}$	${}^{0}B^{3,1}_{6,2}$	${}^{1}B_{7,2}$	${}^{2}B_{8,2}$	${}^{3}B_{9,2}$	${}^{4}B_{10,2}$	${}^{5}B_{11,2}$	${}^{6}B_{12,2}$
${}^{-8}A_{1,3}$	${}^{-7}A_{2,3}$	${}^{-6}A_{3,3}$	${}^{-5}A^{1,1}_{4,3}$	${}^{-4}A^{2,1}_{5,1}$	${}^{-3}A^{3,1}_{6,3}$	${}^{-2}B^{1,2}_{7,3}$	${}^{-1}B^{2,2}_{8,3}$	${}^{0}B^{3,2}_{9,3}$	${}^{1}B_{10,3}$	${}^{2}B_{11,3}$	${}^{3}B_{12,3}$
${}^{-8}A_{1,4}$	${}^{-7}A_{2,4}$	${}^{-6}A_{3,4}$	${}^{-5}A^{1,1}_{4,4}$	${}^{-4}A^{2,1}_{5,4}$	${}^{-3}A^{3,1}_{6,4}$	${}^{-2}B^{1,2}_{7,4}$	${}^{-1}B^{2,2}_{8,4}$	${}^{0}B^{3,2}_{9,4}$	${}^{1}B_{10,4}$	${}^{2}B_{11,4}$	${}^{3}B_{12,4}$
${}^{-11}A_{1,5}$	${}^{-10}A_{2,5}$	${}^{-9}A_{3,5}$	${}^{-8}A^{1,1}_{4,5}$	${}^{-7}A^{2,1}_{5,5}$	${}^{-6}A^{3,1}_{6,5}$	${}^{-5}A^{1,2}_{7,5}$	${}^{-4}A^{2,2}_{8,5}$	${}^{-3}A^{3,2}_{9,5}$	${}^{-2}B_{10,5}$	${}^{-1}B_{11,5}$	${}^{0}B_{12,5}$
${}^{-11}A_{1,6}$	${}^{-10}A_{2,6}$	${}^{-9}A_{3,6}$	${}^{-8}A^{1,1}_{4,6}$	${}^{-7}A^{2,1}_{5,6}$	${}^{-6}A^{3,1}_{6,6}$	${}^{-5}A^{1,2}_{7,6}$	${}^{-4}A^{2,2}_{8,6}$	${}^{-3}A^{3,2}_{9,6}$	${}^{-2}B_{10,6}$	${}^{-1}B_{11,6}$	${}^{0}B_{12,6}$

Note. *A* indicates baseline observation, *B* indicates treatment phase observation, and ${}^{t}B^{m,l}_{i,j}$ indicates measurement occasion *i* for case *j* which occurs when time (centered to estimate the effect 3 observations into treatment) is *t* and *m* observations after the start of treatment in Subexperiment l. Observations enclosed in boxes are included in the between-series estimate of the treatment effect.

see Ugille et al., 2012; Van den Noortgate & Onghena, 2008). The relative advantages of meta-analyses based on aggregated data versus individual participant data have been discussed in the literature on meta-analyzing group studies (Cooper & Patall, 2009; Riley et al., 2010), but not studied for meta-analysis of single-case studies. Both options are generally available when synthesizing MB studies because the individual data are provided graphically in standard reports of MB studies.

A potential difficulty that arises when effect estimation is based on data from multiple studies is that there may be different operationalizations of the dependent variable from study to study. Meta-analysts typically deal with scale variation by using a standardized effect size measure. The relative advantages and limitations of meta-analysis of standardized versus raw score mean differences have been discussed in the context of group comparison meta-analyses (Bond et al., 2003), and a variety of methods for standardizing effects have been considered for single-case research (Ugille et al., 2012; Van den Noortgate & Onghena, 2008). The approach we follow here is to choose the standardized mean difference, δ, as the effect size measure. For δ, both the within- and between-case variance is used in the standardization and thus it is comparable to the standardized mean difference commonly used in meta-analysis of group comparison studies' results. Estimators of δ that assume the effect is consistent over time have been developed (Hedges et al., 2013), and estimators of δ at time t (δ_t) based on the within-series comparison approach have been defined for estimating effects in a single MB study (Pustejovsky et al., 2014) and when using data from a group of MB studies (Van den Noortgate & Onghena, 2008). However, small sample estimates of δ_t using the within-series approach have only been examined in the context of primary studies (Pustejovsky et al., 2014), not meta-analysis. Furthermore, between-series estimators of δ_t have not been previously developed and investigated.

Meta-Analytic Methods of Estimating ψ_t and δ_t

A variety of purposes may motivate the meta-analysis of multiple-baseline studies. In some situations, the goal is to use the data from all the studies to get an estimate of the average treatment effect, whereas in other situations, the purpose is to examine variation in the treatment effect across cases and identify moderators of the individual treatment effects. Our focus here is on the former. By focusing on estimating the average effect, it opens up the opportunity to consider between-series approaches (Ferron et al., 2014) and design comparable effect sizes (i.e., those that standardize the raw score effect in a manner comparable to group comparison studies; Hedges et al., 2013; Pustejovsky et al., 2014).

Within-Series Approach to Estimating ψ_t Using Individual Participant Data ($\hat{\psi}^{W}_{IPD}$)

The unstandardized effect size estimate ψ_t can be obtained by specifying and estimating a three-level mixed linear model for the individual participant data that reflect the meta-analyst's assumptions regarding variation in the dependent variable within individuals, across individuals within a study, and across studies. Consider Y_{ijk} is an observation at the *i*th measurement occasion, for the *j*th case in the *k*th study. As described previously, let Tx_{ijk} be a dummy variable such that $Tx_{ijk} = 1$ if Y_{ijk} is in the treatment phase, and $Tx_{ijk} = 0$ if Y_{ijk} is in the baseline phase, and let $Time_{ijk}$ be a time variable centered per individual so that $Time_{ijk} = 0$ at the time of interest. A common specification of the within individual, or level-1 model, is based on the assumption that trends are linear within phase:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk} Tx_{ijk} + \pi_{2jk} Time_{ijk} + \pi_{3jk} Tx_{ijk} Time_{ijk} + e_{ijk}, \tag{1}$$

and that the error term e_{ijk} is normally distributed with zero mean and a first-order autoregressive covariance structure. That is, $\text{Cov}(e_{ijk}, e_{i'jk}) = \rho^{|i-i'|}\sigma_e^2$ for $i \neq i'$ and thus if $\rho = 0$, then $e_{ijk} \sim N(0,\ \sigma_e^2)$. The coefficients of the within-individual model are typically assumed to vary randomly across individuals within a study,

$$\begin{aligned} \pi_{0jk} &= \gamma_{00k} + u_{0jk}, \\ \pi_{1jk} &= \gamma_{10k} + u_{1jk}, \\ \pi_{2jk} &= \gamma_{20k} + u_{2jk}, \\ \pi_{3jk} &= \gamma_{30k} + u_{3jk}, \end{aligned} \tag{2}$$

and the residual vector, $\mathbf{U} = [u_{0jk},\ u_{0jk},\ u_{0jk},\ u_{0jk}]$, is assumed distributed as $MVN(0, \mathbf{\Sigma}_U)$, where $\mathbf{\Sigma}_U$ is typically assumed unstructured (Pustejovsky et al., 2014) or diagonal (Moeyaert et al., 2015) with variances σ^2_{u0}, σ^2_{u1}, σ^2_{u2}, and σ^2_{u3}. The coefficients from the second level of the mixed linear model are then assumed to vary randomly across studies,

$$\begin{aligned} \gamma_{00k} &= \beta_{000} + v_{00k}, \\ \gamma_{10k} &= \beta_{100} + v_{10k}, \\ \gamma_{20k} &= \beta_{200} + v_{20k}, \\ \gamma_{30k} &= \beta_{300} + v_{30k}, \end{aligned} \tag{3}$$

and the residual vector, $\mathbf{V} = [v_{0jk},\ v_{0jk},\ v_{0jk},\ v_{0jk}]$, is assumed distributed as $MVN(0, \mathbf{\Sigma}_V)$, where $\mathbf{\Sigma}_V$ is typically assumed unstructured (Moeyaert et al., 2015) or diagonal (Moeyaert, Ferron, et al., 2014) with variances $\sigma^2_{\upsilon 0}$, $\sigma^2_{\upsilon 1}$, $\sigma^2_{\upsilon 2}$, and $\sigma^2_{\upsilon 3}$. Then, the effect size estimate at the time of interest is

$$\hat{\psi}^{W}_{IPD} = \hat{\beta}_{100} \tag{4}$$

The subscript IPD indicates individual participant data and the superscript W stands for the within-series approach. Restricted maximum likelihood (REML) estimation of effects using this approach has been previously examined (Moeyaert, Ugille, et al., 2014). Because this estimator is equal to a fixed effect of a REML estimated mixed linear model, it is the empirically best linear unbiased estimator (EBLUE), and under relatively general conditions would be unbiased (Robinson, 1991). Simulation studies have shown little to no bias of this estimator for MB data, assuming the function form of the model is correctly specified (Moeyaert, Ugille, et al., 2014). However, when the functional form is misspecified, the estimate may be substantially biased, and inferences may be inaccurate. The approach also relies on assumptions of homogeneity of error variances and random effect variances across studies, and if these homogeneity assumptions do not hold, the inferences may be jeopardized. In addition, because the treatment variable is a level-1 variable (i.e., a variable that varies within participants), we expect the asymptotic order of magnitude of the error variance of $\hat{\psi}^{W}_{IPD}$ would be inversely related to the total sample size based on the analytic work on mixed linear models of Snijders and Bosker (1993). Thus, for this estimator, we would expect that longer series lengths, more participants within a study, and more studies would lead to more precise estimates.

Between-Series Approach to Estimating ψ_t Using Individual Participant Data ($\hat{\psi}^{B}_{IPD}$)

In addition, the between-series approach can be used to obtain unstandardized effect size estimate ψ_t using individual participant data. The steps are as follows: (1) select for analysis the treatment observations gathered 1 to M observations after the start of treatment and the baseline observations at the same points in time in each of the L subexperiments from each of the K studies, and then (2) specify and estimate a mixed linear model consistent with assumptions regarding variation in the dependent variable within the subexperiments of a study and across studies. The mixed linear model presented here is an extension of the single study model discussed by Ferron et al. (2014).

Suppose that in the K studies, there are $L + 1$ baseline lengths and thus a total of L subexperiments, where the individuals in treatment for 1 to M observations can be compared to the individuals in baseline at those same points in time in each of the L subexperiments. Suppose $LM + M$ is the number of dummy variables, where the first LM dummy variables indicate at which of the LM time points the observation is taken, such that $P^{lm}_{ijk} = 1$ if Y_{ijk} is a mth observation ($m = 1, \ldots, M$) from lth subexperiment ($l = 1, \ldots, L$), and $P^{lm}_{ijk} = 0$, otherwise. Furthermore, M is the number of dummy variables indicating at which of the M points in time after the start of treatment the observation is taken, such that

$P_{ijk}^{LM+m} = 1$ if Y_{ijk} is from time m after the start of treatment, and $P_{ijk}^{LM+m} = 0$, otherwise. Then, the first level of the mixed linear model can be described as

$$Y_{ijk} = \sum_{l=1}^{L}\sum_{m=1}^{M}\gamma_{lmk}P_{ijk}^{lm} + \sum_{m=1}^{M}\gamma_{(LM+m)k}P_{ijk}^{LM+m}Tx_{ijk} + e_{ijk}, \tag{5}$$

where the errors are assumed homogeneous across subexperiments and studies but heterogeneous across phases such that for each subexperiment $e_{ijk} \sim MVN(0, \boldsymbol{\Sigma}_{eA})$ for baseline phase observation errors and $e_{ijk} \sim MVN(0, \boldsymbol{\Sigma}_{eB})$ for treatment phase observation errors and where $\boldsymbol{\Sigma}_{eA}$ and $\boldsymbol{\Sigma}_{eB}$ can be assumed to follow the same structure used in the within-series model. In general, the heterogeneous variance structure across phases for the between-series model (e.g., $\boldsymbol{\Sigma}_{eA}$ and $\boldsymbol{\Sigma}_{eB}$) is considered as the correctly specified model if treatment effect variance is nonzero. Previous study (Joo & Ferron, 2019) has shown that the between-series model assuming a homogeneous variance structure across phases yielded underestimated standard errors of the treatment effect estimate and consequently inflated power and Type I error rates. On the other hand, the between-series model with the heterogeneous variance structure showed unbiased standard error estimates, and inferences for the treatment effect were accurate. By selecting observation m into treatment such that it corresponds to the time t of interest, the treatment effect of interest for study k is $\gamma_{(LM+m)k}$, which is assumed constant across subexperiments.

This single study model is extended to account for multiple studies by assuming the coefficients from this model vary randomly across studies:

$$\begin{aligned}
&\gamma_{1k} = \beta_{10} + u_{1k},\\
&\gamma_{2k} = \beta_{20} + u_{2k},\\
&\ldots\\
&\gamma_{(LM+M)k} = \beta_{(LM+M)0} + u_{(LM+M)k}.
\end{aligned} \tag{6}$$

The residual vector, $\mathbf{U} = [u_{1k}, u_{2k}, \ldots, u_{(LM+M)k}]$, is assumed distributed as $MVN(0, \boldsymbol{\Sigma}_U)$, where $\boldsymbol{\Sigma}_U$ is a diagonal matrix with elements, $\sigma^2_{u_{11}}$, $\sigma^2_{u_{22}}, \ldots,$ $\sigma^2_{u_{(LM+M)(LM+M)}}$. Then, the effect estimate equals

$$\hat{\psi}^{B}_{IPD} = \hat{\beta}_{(LM+M)0}. \tag{7}$$

The superscript B indicates that the effect size estimate is from the between-series approach.

Similar to the within-series approach, REML can be used to estimate the mixed linear model. Because $\hat{\psi}^{B}_{IPD}$ is a fixed effect estimate from a REML estimated mixed linear model, this between-series estimator would be the EBLUE and thus we anticipate it to be approximately unbiased, assuming correct model specification. This expectation is consistent with simulation studies that have examined the study specific effect size estimator $\gamma_{(LM+m)k}$ (Ferron et al.,

2014; Joo & Ferron, 2019). Of note, these simulations relied on random assignment of participants to baseline lengths, and without random assignment the estimates may be biased. Furthermore, like the previous approach, this approach relies on assumptions of homogeneity of error variances and random effect variances across studies. In addition, the fixed effect is associated with a variable that varies between individuals for each subexperiment (not within the individuals like $\hat{\psi}^{W}_{IPD}$), so we expect based on the analytic work of Snijders and Bosker (1993) that the asymptotic order of magnitude of the error variance of $\hat{\psi}^{B}_{IPD}$ would be inversely related to the number of participants in the subexperiments but not the series length. Thus, increasing the number of participants per multiple-baseline study, the number of subexperiments within the multiple-baseline study, and the number of studies in the meta-analysis would be expected to increase the precision of the estimate.

Within-Series Approach to Estimating δ_t Using Individual Participant Data ($\hat{\delta}^{W}_{IPD}$)

The estimate of the standardized effect δ_t can be obtained using the within-series comparison approach by utilizing the following three step process: (1) specify and estimate a two-level mixed linear model for each of the K studies based on researcher assumptions regarding variation in the dependent variable within individuals and across individuals within a study; (2) for each of the K studies, standardize the data; and (3) specify and estimate a three-level mixed linear model for the standardized data, where the first two levels are consistent with the model used for standardizing and the third is based on researcher assumptions about variability in effects across studies.

To define $\hat{\delta}^{W}_{IPD}$ more formally, suppose the two-level mixed linear model defined in Equations 1 and 2 is estimated for each of the K studies. Further assume the covariance structure at level-1 is assumed first order autoregressive and the covariance structure at level-2 is assumed unstructured. The variance for a mixed linear model is $\boldsymbol{V} = \boldsymbol{ZGZ}^T + \boldsymbol{R}$. The diagonal element of $\boldsymbol{ZGZ}^T$ that indexes the between-participant variance in study k when $Tx_{ijk} = 0$ and when $Time_{ijk} = 0$ is σ^2_{u0k}, and the corresponding diagonal element of $\boldsymbol{R}$, which indexes the within-participant variance, is σ^2_{ek}. Thus, the variance needed for standardization is $\sigma^2_{u0k} + \sigma^2_{ek}$.

Using the variance estimates from the two-level mixed linear model, the raw observations are standardized:

$$Y^{W(s)}_{ijk} = \frac{Y_{ijk}}{\sqrt{\hat{\sigma}^2_{ek} + \hat{\sigma}^2_{u0k}}} J(x), \tag{8}$$

where $\hat{\sigma}^2_{ek}$ is the estimated within-participant variation for study k, $\hat{\sigma}^2_{u0k}$ is the estimated between-participant variation in baseline levels estimated for kth study, and $J(x) = 1 - \frac{3}{4x-1}$ is an approximate bias correction used in estimating standardized effect sizes from this two-level mixed linear model (Hedges, 1981; Pustejovsky et al., 2014). We use $x = IJ - 4$ to estimate the degrees of freedom, following Van den Noortgate and Onghena (2008) and preliminary simulations that showed little bias when this relatively simple approach was used to estimate degrees of freedom (for a more complex alternative, see Pustejovsky et al., 2014). The superscript W(s) of Y in Equation 8 indicates standardized observations using the within-series model.

Following standardization, the standardized observations, $Y^{W(s)}_{ijk}$, can be placed into a single data set and then the three-level mixed linear model presented in Equations 1 through 3 can be specified and estimated using REML. Then, the effect size estimate at the time of interest is

$$\hat{\delta}^{W}_{IPD} = \hat{\beta}_{100}. \tag{9}$$

This estimator has not been previously defined or studied, and unlike $\hat{\psi}^{W}_{IPD}$, it is not known to what degree $\hat{\delta}^{W}_{IPD}$ is biased. We anticipate the bias will be relatively small when the model is correctly specified (e.g., correct functional form of the relationship between time and the outcome and homogeneity of error variances and random effect variances across studies) for a couple reasons. First, although it is challenging to correct for bias in the variance estimate from a single short time-series, we anticipate less problems because we are using a mixed linear model with multiple time series. With a single short time series, the autocorrelation parameter is part of the variance expression for the time series, and because the estimation of autocorrelation is biased by an amount that depends on the unknown autocorrelation parameter (Marriott & Pope, 1954), it is challenging to correct for bias. However, in the context of a mixed linear model, autocorrelation is estimated with less bias (Petit-Bois et al., 2016), and more importantly, the autocorrelation parameter is not part of the variance expression used for standardization. Second, the approximate bias correction for small sample size that we use has been shown to work relatively well in related estimators with similar sample sizes (Pustejovsky et al., 2014; Ugille et al., 2014). However, when a more complex error structure is estimated for a mixed linear model (e.g., a first-order autoregressive covariance structure at level-1 as opposed to $\sigma^2_{ek}\boldsymbol{I}$), the RMSE would be expected to be larger. In addition, the process of standardizing the data creates an additional source of error in the estimator, and as a result, we would anticipate the precision of $\hat{\delta}^{W}_{IPD}$ would be less than that of $\hat{\psi}^{W}_{IPD}$, but further work is needed to index the size of these differences.

Between-Series Approach to Estimating δ_t Using Individual Participant Data ($\hat{\delta}^{B}_{IPD}$)

Alternatively, the standardized effect size estimate δ_t can be obtained using the between-series comparison approach by utilizing the following three step process: (1) specify and estimate a single-level between-series model for each of the K studies; (2) for each of the K studies, standardize the observations from the subexperiments; and (3) specify and estimate a two-level mixed linear model for the standardized individual participant data, where the first level is consistent with the model used for standardizing and the second is based on researcher assumptions about variability across studies.

To formally define $\hat{\delta}^{B}_{IPD}$, suppose the single-level between-series model in Equation 5 is estimated for each of the K studies using REML. Then, the standardized observations, $Y^{B(s)}_{ijk}$, are obtained:

$$Y^{B(s)}_{ijk} = \frac{Y_{ijk}}{\sqrt{\hat{\sigma}^2_{eAk}}} J(x), \tag{10}$$

where $\hat{\sigma}^2_{eAk}$ is the estimated baseline phase error variance from the between-series estimator (which is a combination of within- and between-case variation in the baseline phase) for study k. Also, $J(x) = 1 - \frac{3}{4x-1}$ is a bias correction factor, where we estimated the degrees of freedom as $x = N_{Bk} + N_{Ak} - (L+1)$, where N_{Bk} is the number of treatment phase observations in the L subexperiments of study k, and N_{Ak} is the number of baseline observations at that same point in time in the L subexperiments of study k. Similar to the within-series approach for estimating δ_t, the standardized observations, $Y^{B(s)}_{ijk}$, are then used to estimate the two-level between-series model as shown in Equations 5 and 6. The effect size estimate at the time of interest is

$$\hat{\delta}^{B}_{IPD} = \hat{\beta}_{(LM+M)0}. \tag{11}$$

This between-series estimator for the standardized data at the study-level has not been previously defined or investigated. Because the variance estimates are expected to have small sample bias, even when participants are randomly assigned to baselines and the homogeneity of error variances and random effect variances across studies hold, an approximate bias correction is used. However, further work is needed to assess the degree to which the meta-analytic estimator $\hat{\delta}^{B}_{IPD}$ is biased, as well as whether it is more biased than $\hat{\delta}^{W}_{IPD}$. In addition, for the same reasons that $\hat{\psi}^{B}_{IPD}$ is expected to be less precisely estimated than $\hat{\psi}^{W}_{IPD}$, we expect $\hat{\delta}^{B}_{IPD}$ to be less precisely estimated than $\hat{\delta}^{W}_{IPD}$. Furthermore, it is anticipated that $\hat{\delta}^{B}_{IPD}$ would be estimated with less precision than $\hat{\psi}^{B}_{IPD}$, but again

additional work is needed to index the loss of precision that comes from standardizing.

Within-Series Approach to Estimating ψ_t Using Aggregate Data ($\hat{\psi}^{W}_{AD}$)

An alternative within-series approach to estimating the unstandardized effect size, ψ_t, is to estimate the effect size for each of the K studies and then meta-analyze the effect size estimates using the following two step process: (1) specify and estimate the two-level mixed linear model in Equations 1 and 2 for each of the K studies and obtain the study-specific treatment effect estimates from the fixed effects (i.e., $\hat{\gamma}_{10k}$ from Equation 2) and their corresponding standard errors, and (2) specify and estimate a model for the treatment effects using the inverse sampling variance as weights.

More specifically, let $\hat{\psi}^{W}_{k}$ be the estimated within-series effect size for study k, which is estimated through REML from the within-series model in Equations 1 and 2. Then, the effect size for K studies can be estimated with the meta-analytic model:

$$\hat{\psi}^{W}_{k} = \beta^{W}_{0} + u^{W}_{k} + e^{W}_{k}, \tag{12}$$

where u^{W}_{k}, the deviation of the true effect size for study k from the mean true effect size, is assumed distributed as $N(0, \sigma^{2}_{Wuk})$, and the residual, e^{W}_{k}, is assumed distributed as $N(0, \sigma^{2}_{Wuk})$. If one assumed the true effect size did not vary across studies, u^{W}_{k} could be removed from Equation 12 yielding a fixed effect meta-analytic model. The effect size estimate from the aggregate data within series approach is

$$\hat{\psi}^{W}_{AD} = \hat{\beta}^{W}_{0}. \tag{13}$$

The subscript *AD* in Equation 13 indicates that the effect size estimate is derived from aggregated data.

Note that the meta-analytic within-series estimator for aggregated data $\hat{\psi}^{W}_{AD}$ is estimated with a two-step estimation procedure. In the first step, the two-level mixed linear model in Equations 1 and 2 is used to obtain study specific estimates of $\hat{\psi}^{W}_{k}$, which are EBLUE. In the second step, the meta-analytic mean (or weighted combination) of these independent estimates is obtained, and thus $\hat{\psi}^{W}_{AD}$ is expected to be an approximately unbiased estimator as long as the model in the first step was correctly specified. In addition, because the treatment variable is a level-1 variable in the multilevel models estimated in the first step, we expect the asymptotic order of magnitude of the error variance of the study specific effect estimates to be inversely related to the number of observations collected for the study. Furthermore, the error variance for $\hat{\psi}^{W}_{AD}$ would be expected to depend on the number of studies being aggregated in the second

step, as well as the error variances, and thus similar to $\hat{\psi}^{W}_{IPD}$, we expect that $\hat{\psi}^{W}_{AD}$ will become more precise with longer series lengths, more participants within a study, more studies, and less variability in the true effects between studies.

Between-Series Approach to Estimating ψ_t Using Aggregate Data ($\hat{\psi}^{B}_{AD}$)

Similarly, an alternative between-series approach to estimating the unstandardized effect size, ψ_t, involves a two-step process: (1) specify and estimate the between-series model of Equation 5 for each of the K studies and obtain the study specific treatment effect estimates from the fixed effects (i.e., $\gamma_{(LM+M)k}$ from Equation 5), and (2) specify and estimate a model for the treatment effects using the inverse standard errors as weights.

More specifically, let $\hat{\psi}^{B}_{k}$ be the estimated between-series effect size for study k from the between-series model in Equation 5. Then, the meta-analytic model is

$$\hat{\psi}^{B}_{k} = \beta^{B}_{0} + u^{B}_{k} + e^{B}_{k}, \tag{14}$$

where u^{B}_{k}, the deviation of the true effect size for study k from the mean true effect size, is assumed distributed as $N(0, \sigma^{2}_{Buk})$, and the residual, e^{B}_{k}, is assumed distributed as $N(0, \sigma^{2}_{Bek})$. If one assumed the true effect size did not vary across studies, u^{B}_{k} could be removed from Equation 14 yielding a fixed effect meta-analytic model. The meta-analytic mean effect size for the K studies can be estimated from the meta-analytic model by using weighted least square (WLS) methods in which the study effect sizes are weighted by their precision. The effect size estimate can be obtained as

$$\hat{\psi}^{B}_{AD} = \hat{\beta}^{B}_{0}. \tag{15}$$

Similar to the within-series approach, the meta-analytic between-series estimator for aggregated data $\hat{\psi}^{B}_{AD}$ is estimated with a two-step approach (i.e., the models described in Equations 5 and 14). The study specific estimator $\hat{\psi}^{B}_{k}$ has been previously studied though simulations, in which all bias estimates were close to 0 (Ferron et al., 2014). Because the between-series estimator $\hat{\psi}^{B}_{AD}$ is a meta-analytic mean of independent approximately unbiased estimates, $\hat{\psi}^{B}_{k}$, we would expect that $\hat{\psi}^{B}_{AD}$ is an approximately unbiased estimator. Furthermore, the asymptotic order of magnitude of the estimation variance of a study specific effect is expected to be inversely related to the number of subexperiments and participants in the study. The variance of $\hat{\psi}^{B}_{AD}$ would be expected to depend on the number of studies being aggregated, the study specific effect size variances, and the variance in true effect sizes between studies, and thus similar to $\hat{\psi}^{B}_{IPD}$, we expect that $\hat{\psi}^{B}_{AD}$ will become more precise with less variance in true effect sizes

between studies, more participants within a subexperiment, more subexperiments, and more studies, but not with longer series.

Within-Series Approach to Estimating δ_t Using Aggregate Data ($\hat{\delta}^{W}_{AD}$)

An alternative within-series approach to estimating standardized effect size, δ_t, is to estimate the standardized effect size for each of the K studies and then meta-analyze the standardized effect size estimates using the following two step process: (1) specify and estimate the two-level mixed linear model in Equations 1 and 2 for each of the K studies and compute the study specific standardized treatment effect estimates, δ_k, and (2) specify and estimate a meta-analytic model for the treatment effects.

For study k, the standardized treatment effect when assuming the model from Equations 1 and 2 is

$$\delta_k^W = \frac{\gamma_{10k}}{\sqrt{\sigma_{ek}^2 + \sigma_{u0k}^2}}, \tag{16}$$

where γ_{10k} is the unstandardized treatment effect at time t of interest, σ_{ek}^2 is the within-participant variation for study k, and σ_{u0k}^2 is the between-participant variation in baseline levels estimated for study k. The study specific effect is estimated with the bias correction factor:

$$\hat{\delta}_k^W = \frac{\hat{\gamma}_{10k}}{\sqrt{\hat{\sigma}_{ek}^2 + \hat{\sigma}_{u0k}^2}} J(x). \tag{17}$$

The meta-analytic model for the treatment effects is

$$\hat{\delta}_k^W = \beta_0^W + u_k^W + e_k^W, \tag{18}$$

where u_k^W, the deviation of the true effect size for study k from the mean true effect size, is assumed distributed as $N(0, \sigma_{Wuk}^2)$, and the residual, e_k^W, is assumed distributed as $N(0, \sigma_{Wek}^2)$. A fixed effect meta-analytic model could be obtained by removing u_k^W from Equation 18. The effect size estimate at the time of interest is

$$\hat{\delta}_{AD}^W = \hat{\beta}_0^W. \tag{19}$$

The meta-analytic within-series model for standardized aggregated data also involves a two-step estimation procedure (i.e., the model described in Equations 1, 2, 17, and 18). Pustejovsky et al. (2014) investigated a study specific standardized treatment effect size estimator that differed from the first step in of our estimator only in the method of estimating the degrees of freedom for the bias correction and found the bias was relatively small across a variety of models and simulation conditions (absolute bias less than 3% when there was at least four

participants). Because our estimator $\hat{\delta}_{AD}^{W}$ is a meta-analytic mean of treatment effect estimates that are expected to be slightly biased, we would expect it too to be slightly biased.

Between-Series Approach to Estimating δ_t Using Aggregate Data ($\hat{\delta}_{AD}^{B}$)

An alternative between-series approach to estimating standardized effect size, δ_t, is to estimate the effect size for each of the K studies and then meta-analyze the effect size estimates using the following two step process: (1) specify and estimate the between-series model in Equation 5 for each of the K studies and compute the study specific treatment effect estimates δ_k, and (2) specify and estimate a meta-analytic model for the treatment effects using the precision as weights.

For study k, the standardized treatment effect is

$$\delta_k^B = \frac{\gamma_{(LM+M)k}}{\sqrt{\sigma_{eAk}^2}}, \tag{20}$$

where $\gamma_{(LM+M)k}$ is the unstandardized treatment effect at time t of interest after the start of intervention and σ_{eAk}^2 is the combined within- and between-participant variation for study k during baseline (as defined by Equation 5). The estimated effect for study k is

$$\hat{\delta}_k^B = \frac{\hat{\gamma}_{(LM+M)k}}{\sqrt{\hat{\sigma}_{eAk}^2}} J(x). \tag{21}$$

The meta-analytic model for the treatment effects is

$$\hat{\delta}_k^B = \beta_0^B + u_k^B + e_k^B, \tag{22}$$

where the residuals u_k^B and e_k^B are assumed normally distributed, with variances of σ_{Buk}^2 and σ_{Bek}^2, respectively. A fixed effect meta-analytic model could be obtained by removing u_k^B from Equation 22. The effect size estimate at the time of interest is

$$\hat{\delta}_{AD}^B = \hat{\beta}_0^B. \tag{23}$$

Neither the standardized study-specific effect size estimator using the between-series model ($\hat{\delta}_k^B$) nor the meta-analytic estimator $\hat{\delta}_{AD}^B$ has been previously defined or examined. Because these estimators make use of variance components that are estimated through REML with small sample sizes, some bias is expected, and thus an approximate correction for small sample size bias was proposed. However, simulation work is needed to index the amount of bias associated with the estimator $\hat{\delta}_{AD}^B$. In addition, we expect $\hat{\delta}_{AD}^B$ to be less precisely estimated than $\hat{\delta}_{AD}^W$ for the same sample size reasons the other between-series

estimators are expected to be less precisely estimated than their within-series counterparts. Furthermore, it is anticipated that $\hat{\delta}^{B}_{AD}$ would be estimated with less precision than $\hat{\psi}^{B}_{AD}$, but again additional work is needed to index the loss of precision that comes from standardizing.

Purpose of the Study

Because REML estimation of the meta-analytic effect in each of the eight methods is based on large sample theory, two simulation studies were conducted to empirically compare the proposed methods for estimating ψ_t and δ_t under various small sample size conditions. In the first simulation study, we examined the accuracy of the eight methods for estimating the meta-analytic effect sizes under the conditions where each of the models was correctly specified. It is important to note, however, that the assumptions are not the same for each method. The within-series estimators rely on the assumption that the functional form of the relationship between time and the outcome is correctly specified, whereas the between-series estimators rely on the assumption that participants are randomly assigned to baseline lengths. In addition, the IPD approaches rely on assumptions of homogeneity of error variance and random effect variances across studies, which are relaxed in the AD models. In a second simulation study, we examined the accuracy of eight methods under the conditions where some of the model assumptions are not satisfied.

Simulation Study 1

The purpose of the first simulation study was to compare the eight mixed linear modeling approaches to meta-analyzing MB studies. Of particular interest was determining to what degree there is bias in the standardized effect estimators ($\hat{\delta}^{W}_{IPD}$, $\hat{\delta}^{B}_{IPD}$, $\hat{\delta}^{W}_{AD}$, and $\hat{\delta}^{B}_{AD}$). Also of interest was the degree to which the standardized estimators are less precise than their unstandardized counterparts (e.g., error variance in $\hat{\delta}^{W}_{IPD}$ versus $\hat{\psi}^{W}_{IPD}$) and to what degree the between-series estimators are less precise than the within-series estimators (e.g., error variance of $\hat{\psi}^{W}_{IPD}$ versus $\hat{\psi}^{B}_{IPD}$).

Factors that were manipulated in this simulation study included the series length, number of participants per study, number of studies, and level of model complexity. The series length was varied using two levels, 20 and 40, and the number of participants per study was 4 or 8, which is commonly observed in MB studies (e.g., Botella et al., 2000; Rantz et al., 2009). The same number of participants was assumed across studies. When there were four participants, each participant entered into the treatment phase at different time points, resulting in three occasions of temporal staggering. When the series length was 20, the baseline lengths were 5, 8, 11, and 14; whereas when the series length was 40, the

baseline lengths were 10, 16, 22, and 28. With eight participants, two participants entered into the treatment phase at the same time. The baseline lengths for 20 and 40 observations were the same for eight participant designs as they were for four participant designs. The number of studies was varied using 10 or 30, which represents a small or medium number of studies included in the meta-analyses of single-case design, respectively (e.g., Ganz et al., 2012; Wang et al., 2011). Model complexities considered in this study included no trend in any phase, trend in the treatment phase, and trend in the treatment phase plus autocorrelation, which is commonly observed in MB studies.

Data were generated based on the three-level model in Equations 1–3. That is, the level-1 error term, e_{ijk} , was generated from a first-order autoregressive process with a mean of 0, the variance of .67, and an autocorrelation (ρ) varying at 0 or .20. The level-2 error terms, u_{0jk} and u_{1jk}, were independently generated from a normal distribution with means of 0 and variances of .33. Also, u_{2jk} and u_{3jk} were assumed to be 0, indicating the time trends in baseline and treatment are the same for all cases. The level-3 error terms, ν_{00k} and ν_{10k}, were independent and normally distributed with means of 0 and variances of .50 and 0, respectively, and v_{20k} and v_{30k} were fixed at 0, implying that the time trends in baseline and treatment are the same for all studies. When no trend in the baseline phase was generated, the treatment effect, β_{100}, equaled 1, and the other fixed effect parameters (β_{000}, β_{200}, and β_{300}) were fixed at 0. In the second condition, in which a linear trend in the treatment phase was generated, we set $\beta_{100} = .50$, and $\beta_{300} = .25$. On top of this linear trend, a first-order autoregressive model with ρ of .20 was generated for the third model-generating condition. Series were generated using the ARMASIM function in SAS (SAS Institute, 2014) and thus adding autocorrelation was accompanied by a decrease in the white noise variance so that the within-series variance stayed at .67 across conditions. We chose those population values for the fixed and variance component parameters to make the population standardized as well as unstandardized effect parameters equal to 1.

For each of the 24 conditions ($2 \times 2 \times 2 \times 3$), 3,000 data sets were simulated. For each data set, the treatment effect was estimated using each of the eight estimators introduced earlier, which varied depending on whether researchers choose to use within-series or between-series models, analyze individual participant data or study effect sizes, and to standardize or not. For analyzing effect sizes, we used a fixed effects meta-analytic model. All mixed linear models were estimated using REML through the Mixed Procedure in SAS and all WLSs regressions were estimated using the Regression Procedure in SAS. To make the results comparable between within-series and between-series estimation, all treatment effects were estimated at a time three observations into the treatment phase. That is, for all the estimators that are based on the between-series models, the treatment effect at the third point into the treatment phase was estimated

(i.e., $M = 3$ in Equations 5 and 6), and for all estimators based on the within-series models time was centered per participant so that when a trend was estimated, the treatment effect corresponded to the time of the third treatment observation. In addition, all models estimated were specified to match the model used in data generation. When the data generation model did not include time trends, time effects were not included in the model estimated, and when the data were generated with an autocorrelation parameter of 0, no autocorrelation parameter was included in the model estimated. Thus, the complexity of the models estimated increased with the complexity of the data generated. The SAS codes for estimating effect sizes with the within- and between-series estimators are available in the Appendix in the online version of the journal.

Simulation outcomes of focal interest included bias in the average treatment effect across studies and the RMSE associated with this effect. Bias and RMSE were computed as

$$\text{Bias} = \frac{\sum_{i=1}^{3,000} \hat{\theta}_i - \theta_i}{3,000}, \tag{24}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{3,000} (\hat{\theta}_i - \theta_i)^2}{3,000}}, \tag{25}$$

where $\hat{\theta}_i$ is the estimated parameter of interest, and θ_i is the true parameter for the ith replicated data.

Study 1 Results

Table 2 shows the bias of the eight approaches for the effect size estimates. As shown in Table 2, minimal bias across the eight approaches was found. Less than 3% relative bias was observed across simulation conditions. Maximum bias, 2.5% of the population value, was observed when the between-series model was used to analyze the standardized effect sizes with the smallest sample size condition ($I = 20$, $J = 4$, and $K = 10$). Overall, when the meta-analysis was performed for the standardized effect size, more bias was observed than the unstandardized effect size. For example, for the condition where the within-series model was used to analyze IPD when the sample size was relatively small ($I = 20$, $J = 4$, and $K = 10$) and only the level effect was present, the bias of the standardized data was 1.5% as opposed to 0% for the unstandardized data. This pattern is consistent across simulation conditions. As expected, the unstandardized effect estimators showed no notable bias (the relative bias is less than 1% across all conditions). For the standardized effect estimators, there was only a small amount of bias, and this bias decreased with increasing sample size.

TABLE 2.
Percentage of Bias of Eight Approaches for Effect Size Estimates

				$K = 10$				$K = 30$			
				$J = 4$		$J = 8$		$J = 4$		$J = 8$	
Model Complexity	Scale	Data	Fitted Model	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$
Level only	Raw	IPD	Within	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
			Between	0.2	0.0	−0.2	0.4	0.2	−0.4	−0.3	−0.4
		AD	Within	0.0	0.1	0.0	0.0	−0.1	0.0	0.0	0.0
			Between	0.2	0.0	−0.2	0.4	0.2	−0.4	−0.3	−0.4
	Std.	IPD	Within	1.5	1.6	0.9	0.8	1.1	1.3	0.7	0.6
			Between	2.5	2.4	0.9	1.5	2.0	1.8	0.8	0.7
		AD	Within	0.9	0.3	0.7	0.3	0.3	−0.1	0.5	0.1
			Between	2.5	2.4	0.9	1.5	2.0	1.8	0.8	0.7
Level and slope	Raw	IPD	Within	0.0	−0.2	−0.1	0.1	0.3	0.0	0.1	−0.1
			Between	−0.4	−0.6	0.0	0.2	−0.1	0.5	−0.1	−0.1
		AD	Within	−0.1	−0.1	−0.1	0.0	0.2	0.0	0.1	−0.2
			Between	−0.3	−0.8	−0.1	0.0	−0.4	0.2	−0.2	−0.1
	Std.	IPD	Within	1.7	1.3	0.8	0.9	1.8	1.4	1.1	0.6
			Between	0.5	0.0	−0.2	−0.1	−0.1	0.6	−0.6	−0.5
		AD	Within	0.9	0.3	0.4	0.3	0.8	0.2	0.6	0.0
			Between	−0.1	−0.7	−0.6	−0.5	−0.6	0.0	−0.9	−0.8

(continued)

TABLE 2. *(continued)*

				$K = 10$				$K = 30$			
				$J = 4$		$J = 8$		$J = 4$		$J = 8$	
Model Complexity	Scale	Data	Fitted Model	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$
Level, slope, and AR(1)	Raw	IPD	Within	−0.2	0.3	0.1	0.0	−0.1	0.0	0.0	−0.1
			Between	−0.5	0.2	0.0	−0.1	0.1	0.1	0.1	−0.2
		AD	Within	−0.5	0.2	0.1	0.2	0.0	0.0	0.1	−0.1
			Between	−0.1	0.3	0.0	0.0	0.0	0.1	0.1	−0.1
	Std.	IPD	Within	1.6	1.9	1.2	0.9	1.4	1.4	1.0	0.7
			Between	0.3	0.5	−0.7	−0.8	0.0	0.0	−1.0	−1.3
		AD	Within	0.7	0.9	0.8	0.5	0.5	0.2	0.6	0.1
			Between	−0.3	−0.1	−1.1	−1.1	−0.6	−0.5	−1.2	−1.5

Note. Bias were multiplied by 100. K = number of studies; J = number of participants; I = number of measurement occasions; Std. = standardized; IPD = individual participant data; AD = aggregated data; AR(1) = first-order autoregressive; within = within-series model; between = between-series model.

TABLE 3.
Root Mean Squared Error of Eight Approaches for Effect Size Estimates

				$K = 10$				$K = 30$			
				$J = 4$		$J = 8$		$J = 4$		$J = 8$	
Model Complexity	Scale	Data	Fitted Model	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$
Level only	Raw	IPD	Within	.063	.045	.045	.032	.032	.032	.032	.000
			Between	.230	.228	.161	.164	.134	.130	.095	.095
		AD	Within	.071	.045	.045	.032	.045	.032	.032	.000
			Between	.230	.228	.161	.164	.134	.130	.095	.095
	Std.	IPD	Within	.084	.063	.055	.045	.045	.045	.032	.032
			Between	.249	.249	.170	.173	.148	.141	.100	.100
		AD	Within	.089	.071	.055	.045	.045	.045	.032	.032
			Between	.249	.249	.170	.173	.148	.141	.100	.100
Level and slope	Raw	IPD	Within	.148	.122	.105	.084	.084	.071	.055	.045
			Between	.249	.257	.179	.179	.148	.145	.100	.105
		AD	Within	.167	.138	.110	.095	.095	.084	.063	.055
			Between	.308	.321	.192	.192	.184	.184	.114	.114
	Std.	IPD	Within	.161	.130	.110	.089	.089	.077	.063	.055
			Between	.297	.305	.187	.187	.173	.173	.110	.110
		AD	Within	.170	.141	.110	.095	.095	.084	.063	.055
			Between	.302	.311	.190	.190	.179	.176	.110	.110

(continued)

TABLE 3. *(continued)*

				$K = 10$				$K = 30$			
				$J = 4$		$J = 8$		$J = 4$		$J = 8$	
Model Complexity	Scale	Data	Fitted Model	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$
Level, slope, and AR(1)	Raw	IPD	Within	.155	.126	.110	.084	.084	.071	.063	.045
			Between	.255	.253	.176	.179	.148	.145	.105	.105
		AD	Within	.179	.148	.118	.095	.100	.077	.063	.055
			Between	.329	.321	.195	.195	.192	.187	.114	.118
	Std.	IPD	Within	.167	.141	.118	.089	.095	.071	.063	.055
			Between	.311	.303	.184	.187	.182	.176	.110	.114
		AD	Within	.179	.152	.122	.095	.100	.077	.063	.055
			Between	.321	.311	.187	.187	.184	.179	.110	.114

Note. K = number of studies; J = number of participants; I = number of measurement occasions; Std. = standardized; IPD = individual participant data; AD = aggregated data; AR(1) = first-order autoregressive; within = within-series model; between = between-series model.

Table 3 shows the RMSE of the eight approaches for the effect size estimates. As could be expected from the bias results, analyzing standardized effect sizes produced slightly larger RMSE values than raw effect sizes across simulation conditions (e.g., marginal RMSE for raw effect size was .126 as opposed to .135 for standardized effect size).

Consistent with expectations and as shown in Table 3, the RMSE was substantially higher for the between-series estimators than the within-series estimators. Marginal RMSE for the within-series estimators was .080 as opposed to .181 for the between-series estimators, and the differences were more pronounced when the series length was 40 than when they were 20. For series lengths of 40 the marginal RMSEs were .072 and .181 for the within- and between-series estimators, respectively; and for series, lengths of 20 the marginal RMSEs were .089 and .181 for the within- and between-series estimators, respectively. This pattern was expected because the within-series estimators use all generated data from a study and thus have a sample size that doubles when the series length doubles (e.g., the sample size per study when $J = 4$ is 80 when $I = 20$, and 160 when $I = 40$), whereas the between-series estimators had the same number of individuals in each subexperiment regardless of the series length, and thus the sample size for those estimators did vary with series length (e.g., the sample size per study when $J = 4$ is 9 both when $I = 20$ and when $I = 40$).

In addition, meta-analyzing IPD yielded slightly lower RMSE values than using AD as the marginal RMSE for IPD was .126 as opposed to .135 for AD. The same pattern was observed in the standardized effect size result (e.g., for the within-series model, RMSE for standardized IPD was .063 as opposed to .071 for standardized AD, when $K = 10$, $J = 4$, and $I = 40$). This result implies that intermediate standardization of effect size estimates in both within- and between-series models increased the RMSE (i.e., standard deviation) under the conditions studied. However, it is important to note that IPD approaches have different assumptions than the AD approaches. The IPD approaches assumed the variance was homogeneous across studies, which allowed more data to be used in estimating the standardizer. Because the data generation was consistent with this assumption, it could be expected that the IPD approach would provide a more stable estimator of the standardizer.

Lastly, as models became more complex to estimate, higher RMSE values across the eight approaches were observed. Including a nonzero slope parameter in the data generation and estimation models increased, the RMSE of the estimates across simulation conditions and introducing autocorrelation to the data generation and estimation models further increased the RMSE. The marginal RMSE values for the model based on a level effect; level and slope effect; and level, slope, and autocorrelation effect conditions were .102, .144, and .147, respectively. The effect of increasing model complexity was similar across the within-series estimators (marginal RMSEs of .043, .097, and .101) and the

between-series estimators (marginal RMSEs of .161, .190, and .193). However, the increment of RMSEs decreased as the numbers of studies and participants increased ($J = 8$ and $K = 30$).

Simulation Study 2

The second simulation study was conducted to extend the comparison of the eight mixed linear modeling approaches to meta-analyzing MB studies to conditions where (a) the effects varied randomly across studies, (b) participants were not randomly assigned to baseline lengths, and (c) there were unknown events that impacted the times series. For each of these extensions, the simulation methods paralleled those used in the initial simulation. We examined data conditions that varied in series length (10 and 40), number of cases (4 and 8), and number of studies (10 and 30), and for each of the new conditions, we started with the simplest data generation model from the initial simulations (i.e., the one with no trends or autocorrelation). To simulate variance in the treatment effect across studies, the level-3 error terms, ν_{00k} and ν_{10k}, were independently sampled from normal distributions with means of 0 and variances of .25 and .25 (as opposed to .50 and 0 in the initial simulations).

To simulate conditions without random assignment of participants to baseline lengths and to mimic the sometimes used practice of assigning those with the most problematic levels of baseline behavior to the shortest baselines, the participants were ordered and assigned to baseline lengths based on the value of their level-2 intercept error (u_{0jk}). Specifically, when there were four participants, the participant with the lowest value of u_{0jk} was assigned to the shortest baseline, the participant with the next lowest value of u_{0jk} was assigned to the next shortest baseline, and so on. For the conditions with eight cases, the two participants with lowest values of u_{0jk} were assigned to the shortest baseline, the two participants with the next lowest values of u_{0jk} were assigned to the next shortest baseline, and so on. We expected that this change would lead to negative bias in each of the between-series estimators, because in each subexperiment, those assigned to treatment are those with more problematic behavior and those left in baseline had less problematic behavior.

To simulate conditions with unknown event effects (e.g., a participant changing medicines during the study or a parent trying a different behavioral management technique), a time point for the beginning of the event was randomly selected from among all the time points in the participant's series. A value of .20 was then added to the outcome at this time point and all subsequent time points. If the event coincided with the first observation, the series mean would increase by .20 and we would expect no bias in any of the treatment effect estimators, whereas if the event was selected to coincide with the treatment, we would expect the within-series treatment effect estimators to be biased by .20, and the between-series effect estimators to be unbiased. Because the time of

the event is selected randomly for each participant, we expect there would be a bias between 0 and .20 for the raw score within-series estimators, but no bias in the raw score between series estimators. For the standardized estimators, defining bias is problematic because the variance of the series is impacted by an amount that depends on which time points were randomly chosen for the different participants. We will still present the difference between the average estimated effect and a standardized effect of 1 and anticipate larger discrepancies for the within-series estimators than the between-series estimators.

Study 2 Results

The bias and RMSE for the eight approaches under each of these alternative data conditions are shown in Tables 4 and 5, respectively. When random study effects were added, the bias remained minimal across sample size conditions for all estimators, with the exception that the standardized between-series IPD estimator showed bias ranging from 9.6% to 11.4% for conditions that had four participants per study.

The RMSE results also showed the similar pattern as the bias results, and the standardized between-series IPD estimator showed the highest RMSE of .515 for the condition where 10 studies, four participants per study, and 20 measurement occasions were considered. Consistent with the RMSE results from the first simulation study, as the numbers of studies and participants increased, RMSE decreased for the between-series estimators, and as the numbers of studies, participants, and measurement occasions increased and RMSE decreased for the within-series estimators.

When participants were assigned to baseline lengths systematically, with those having more problematic behavior being assigned to shorter baselines, the between-series estimators were substantially negatively biased (ranging from -49% to -69% across estimators and data conditions); whereas the within-series estimators showed small levels of negative bias (ranging from -1% to -4% across estimators and data conditions). Moreover, RMSE of the within-series estimators substantially decreased as the numbers of studies, participants, and measurement occasions increased, whereas RMSE of the between-series estimators consistently showed the relatively high values ranging from .551 to .711. For the conditions where event effects were randomly added to the time series the raw score, within-series estimators consistently showed a 10% positive bias, which corresponds to half the value of the event effect, whereas the raw score between-series estimators showed no appreciable bias (-1% to $+1\%$).

Discussion

In the present study, we proposed various meta-analytic approaches for MB studies in single-case research. The approaches we described here were methods of estimating ψ_t and δ_t using both within- and between-series models. We also

TABLE 4.
Percentage of Bias of Eight Approaches for Effect Size Estimates Under Alternative Data Conditions

				$K = 10$				$K = 30$			
				$J = 4$		$J = 8$		$J = 4$		$J = 8$	
Data Condition	Scale	Data	Fitted Model	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$
Random study effects	Raw	IPD	Within	−0.5	0.2	0.1	−0.1	−0.1	0.1	−0.2	−0.1
			Between	−0.3	0.0	0.3	0.3	0.0	0.2	−0.4	−0.2
		AD	Within	−0.7	0.1	−0.3	−0.3	−0.1	−0.3	−0.1	−0.1
			Between	1.1	−0.2	0.4	0.4	−0.3	1.1	−0.2	−0.2
	Std.	IPD	Within	2.0	2.5	1.1	1.1	2.2	2.4	0.9	1.1
			Between	11.0	9.9	0.3	1.0	9.6	11.4	0.1	0.4
		AD	Within	−0.8	−0.2	−0.8	−0.4	−0.8	−1.1	−0.8	−0.6
			Between	3.6	2.4	−3.8	−3.4	1.7	3.2	−4.6	−4.5
Systematic assignment of cases	Raw	IPD	Within	−3.1	−1.7	−3.5	−2.1	−3.2	−1.7	−3.6	−2.1
			Between	−60.8	−61.3	−66.5	−67.0	−60.9	−60.6	−66.9	−67.1
		AD	Within	−2.5	−1.7	−3.3	−2.0	−2.8	−1.8	−3.5	−2.1
			Between	−59.8	−61.3	−67.9	−68.5	−61.9	−60.7	−68.7	−68.6
	Std.	IPD	Within	−2.0	−0.5	−2.9	−1.5	−2.2	−0.7	−3.2	−1.6
			Between	−49.3	−50.0	−63.4	−63.8	−51.4	−50.2	−64.0	−64.1
		AD	Within	−3.1	−2.0	−3.6	−2.2	−3.7	−2.6	−4.0	−2.5
			Between	−53.0	−54.3	−65.3	−65.7	−55.4	−54.1	−66.0	−65.9

(continued)

TABLE 4. *(continued)*

				$K = 10$				$K = 30$			
				$J = 4$		$J = 8$		$J = 4$		$J = 8$	
Data Condition	Scale	Data	Fitted Model	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$
Random event effects	Raw	IPD	Within	9.9	10.2	9.9	9.9	9.9	9.9	10.0	9.8
			Between	0.4	0.0	0.1	1.0	0.2	0.2	−0.2	0.0
		AD	Within	9.9	10.4	10.0	9.9	9.9	10.0	9.9	9.8
			Between	−0.4	−0.8	0.0	1.2	0.2	0.3	−0.3	0.0
	Std.	IPD	Within	11.1	11.8	10.3	10.5	11.2	11.5	10.2	10.2
			Between	8.1	8.7	−1.2	−0.4	7.7	7.0	−2.2	−1.9
		AD	Within	9.0	9.8	9.2	9.2	8.7	8.8	8.9	8.8
			Between	1.6	2.0	−4.1	−3.2	1.2	0.8	−5.2	−4.9

Note. Bias were multiplied by 100. K = number of studies; J = number of participants; I = number of measurement occasions; Std. = standardized; IPD = individual participant data; AD = aggregated data; within = within-series model; between = between-series model.

TABLE 5.
Root Mean Squared Error of Eight Approaches for Effect Size Estimates Under Alternative Data Conditions

				$K = 10$				$K = 30$			
				$J = 4$		$J = 8$		$J = 4$		$J = 8$	
Data Condition	Scale	Data	Fitted Model	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$
Random study effects	Raw	IPD	Within	.196	.186	.175	.171	.110	.108	.101	.098
			Between	.298	.292	.239	.236	.168	.173	.138	.137
		AD	Within	.239	.246	.203	.202	.138	.145	.117	.119
			Between	.437	.432	.277	.273	.251	.255	.161	.162
	Std.	IPD	Within	.211	.200	.182	.177	.120	.117	.106	.102
			Between	.516	.511	.268	.260	.305	.315	.153	.153
		AD	Within	.242	.250	.202	.203	.138	.146	.118	.118
			Between	.499	.493	.276	.267	.285	.295	.164	.163
Systematic assignment of cases	Raw	IPD	Within	.116	.104	.086	.076	.072	.061	.057	.046
			Between	.653	.654	.684	.689	.623	.621	.675	.677
		AD	Within	.138	.133	.096	.087	.085	.080	.063	.054
			Between	.691	.700	.704	.711	.650	.640	.696	.695
	Std.	IPD	Within	.126	.112	.088	.080	.075	.064	.057	.046
			Between	.640	.638	.662	.666	.563	.551	.650	.651
		AD	Within	.145	.140	.100	.092	.091	.084	.067	.058
			Between	.670	.673	.681	.685	.600	.588	.670	.669

(continued)

TABLE 5. *(continued)*

				$K = 10$				$K = 30$			
				$J = 4$		$J = 8$		$J = 4$		$J = 8$	
Data Condition	Scale	Data	Fitted Model	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$	$I = 20$	$I = 40$
Random event effects	Raw	IPD	Within	.149	.142	.128	.123	.118	.115	.109	.106
			Between	.262	.252	.181	.178	.146	.145	.103	.105
		AD	Within	.168	.167	.136	.130	.128	.127	.112	.110
			Between	.378	.368	.211	.206	.212	.215	.124	.122
	Std.	IPD	Within	.170	.164	.138	.133	.135	.133	.114	.112
			Between	.455	.461	.211	.206	.268	.255	.121	.123
		AD	Within	.174	.174	.137	.130	.124	.123	.106	.103
			Between	.445	.425	.217	.211	.246	.240	.133	.133

Note. K = number of studies; J = number of participants; I = number of measurement occasions; Std. = standardized; IPD = individual participant data; AD = aggregated data; within = within-series model; between = between-series model.

described meta-analytic strategies where individual participant data (unstandardized or standardized) are used to estimate effects, and a traditional strategy where an estimate of ψ_t or δ_t is obtained from each MB study and then these estimates of the effects are aggregated across studies. To evaluate the viability of these approaches, we further conducted an initial simulation study with conditions where the sample sizes (i.e., numbers of measurement occasions, cases, and studies) and the model complexities (i.e., level effect, level and slope effects, and level changes, slope changes, and autocorrelation) were varied, and data were generated consistent with the assumptions of the meta-analytic approaches.

Based on the initial simulation study, we found that the eight meta-analytic approaches we proposed produced minimal bias. Relative bias of the estimated effects was less than 3% of the population values under various sample size and model complexity conditions. It is worthwhile to note that a minimal relative bias of the estimates was obtained even from the condition where sample size is relatively small (e.g., $I = 10$, $J = 4$, and $K = 10$). Given that estimating the standardized effect size from a set of MB studies has been a major concern in meta-analysis contexts due to inaccurate variance estimates (Moeyaert et al., 2015; Ugille et al., 2012), this finding provides a significant contribution in meta-analysis of MB studies. The initial simulation study provided empirical evidence that our alternative approaches have reasonably small bias even for standardized estimates in small sample size conditions. However, in situations where standardization is not needed, we recommend not standardizing, in which case no approximate small sample adjustments are needed, and the effects have smaller RMSE and no bias as opposed to minimal bias.

In addition, the initial simulation study indicates that the within-series approach produced smaller RMSE compared to the between-series approach, which is consistent with previous research that examined study specific unstandardized estimators (Ferron et al., 2014). In principle, parameters of the between-series model are estimated using subexperimental observations across cases, whereas those of the within-series model are estimated using the full set of observations in MB studies. Consequently, the within-series model produces estimates with less sampling error. However, as Ferron et al. (2014) denoted, the within- and between-series approaches are based on different assumptions and thus may be vulnerable to bias under different circumstances.

The between-series estimators rely on the assumption that individuals were randomly assigned to baseline conditions, whereas the within-series model relies on temporal stability assumptions and baseline projections. To further examine this issue, a second simulation study was conducted to examine bias under conditions where assumptions were violated. When instead of randomly assigning participants to baseline conditions, the participants were systematically assigned, such that those with the most problematic baseline levels were assigned to the shortest baselines, the between-series estimates were substantially biased. Thus, between-series estimators should be avoided in circumstances where there

is systematic assignment of cases to baselines. Conversely, when cases were randomly assigned, but event effects were added to randomly selected time points, the within-series model became misspecified and the within-series estimators were biased by half the size of the randomly placed event effects. Based on what we found, if a researcher is confident about the model specification (e.g., confident in the absence of maturation, event, instrumentation, and practice effects), then the within-series approach is recommended for estimating the average effect size across MB studies. If model specification is a primary concern and there was random assignment of cases to baselines, we recommend applied researchers use both within- and between-series approaches for estimating the average effect. By comparing the treatment effect estimates from the between- and within-series models, researchers can potentially detect model misspecification.

We also acknowledge that the simulation studies we have presented have several limitations. First, we focused on estimation of the average treatment effect, and thus our recommendations are limited to that purpose. Additional research is needed to develop methods for estimating individual treatment effects that are standardized to be design comparable, along with methods for exploring potential moderators of such effects. Second, the data generation models assumed no trends or linear trends and a continuous outcome variable. In practice, non-linear trends could occur, such as when the effect of the intervention is delayed or decays with time, and the outcome may be based on counts of behaviors. More complex mixed models with a binomial- or Poisson-based link function or piece-wise mixed models can be adapted for those situations (Hembry et al., 2015; Shadish et al., 2013), but research is needed to extend those approaches to meta-analytic contexts. Third, the variance structure of the data generation model was relatively simple. Homogeneous variances across phases, and cases were considered, and this may not be the case in all MB studies. Although the heterogeneous variance structures have been investigated previously in the contexts of MB studies (Baek & Ferron, 2013; Joo et al., 2019), more precise investigation of the proposed meta-analytic approaches with the heterogeneous variance structure is needed.

Fourth, the AD approaches used a fixed effect model whereas the IPD approaches estimated across study variance in the average treatment effect. It would be helpful if additional research considered random effects models for aggregating study specific effect sizes and random effects models for aggregating case specific effect sizes from within-series models. Last, in the present study, only a single intraclass correlation coefficient (ICC) was considered in the data generation. It would be helpful if additional research examined the degree to which the difference in RMSE values for the between-series versus within-series estimators was impacted by the ICC. In our study, the difference between the RMSE of the within-series estimators and the between-series estimators was greater when the series lengths were longer. This can be explained by

the additional observations from the longer series magnifying the information discrepancy between the observations used by the within-series estimators and the subset of observations used by the between-series estimators. However, how much additional information would be in those additional observations is expected to depend on the ICC. For larger ICCs, additional observations within a case are relatively less informative, which may attenuate the RMSE differences between the within- and between-series estimators.

Nonetheless, the results of this study provide valuable information about how to obtain an average effect estimate using within- and between-series estimators, and study-specific standardized effect sizes, which are being considered for meta-analysis in single-case research. We hope the alternative models and methods developed in this study will be useful to applied researchers and expand research possibilities with the meta-analysis of single-case research.

Authors' Note

The opinions expressed are those of the authors and do not represent views of the institute or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We gratefully acknowledge support from the Institute of Educational Sciences, U.S. Department of Education (through grant no. R305D150007).

References

Baek, E., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across participant variation in autocorrelation and residual variance. *Behavior Research Methods*, *45*, 65–74.

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.

Bond, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*, 406–418.

Botella, C., Baños, R. M., Villa, H., Perpiñá, C., & García-Palacios, A. (2000). Virtual reality in the treatment of claustrophobic fear: A controlled, multiple-baseline design. *Behavior Therapy*, *31*, 583–595.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant versus aggregated data. *Psychological Methods*, *14*, 165–176.

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating casual effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*, *19*, 493–510.

Ganz, J. B., Earles-Vollrath, T. L., Heath, A. K., Parker, R. I., Rispoli, M. J., & Duran, J. B. (2012). A meta-analysis of single case research studies on aided augmentative and alternative communication systems with individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *42*, 60–74.

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

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple for multiple baseline designs across individuals. *Research Synthesis Methods*, *4*, 324–341.

Hembry, I., Bunuan, R., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2015). Estimation of a nonlinear intervention phase trajectory for multiple baseline design data. *Journal of Experimental Education*, *83*, 514–546.

Joo, S. H., & Ferron, J. M. (2019). Application of the within- and between-series estimators to non-normal multiple-baseline data: Maximum likelihood and Bayesian approaches. *Multivariate Behavioral Research*, *54*, 666–689.

Joo, S. H., Ferron, J. M., Moeyaert, M., Beretvas, S. N., & Van den Noortgate, W. (2019). Approaches for specifying the level-1 error structure when synthesizing single-case data. *The Journal of Experimental Education*, *87*, 55–74.

Marriott, F. H. C., & Pope, J. A. (1954). Bias in the estimation of autocorrelations. *Biometrika*, *41*, 390–402.

Moeyaert, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology*, *52*, 191–211.

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). Three-level analysis of single-case experimental data: Empirical validation. *The Journal of Experimental Education*, *82*, 1–21.

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2015). The misspecification of the covariance structures in multilevel models for single-case data. *The Journal of Experimental Education*, *84*, 473–509.

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*, *44*, 795–805.

Petit-Bois, M., Baek, E. K., Van den Noortgate, W., Beretvas, S. N., & Ferron, J. M. (2016). The consequences of modeling autocorrelation when synthesizing single-case studies using a three level model. *Behavior Research Methods*, *48*, 803–812.

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics*, *39*, 368–393.

Rantz, W. G., Dickinson, A. M., Sinclair, G. A., & Van Houten, R. (2009). The effect of feedback on the accuracy of checklist completion during instrument flight training. *Journal of Applied Behavior Analysis*, *42*, 497–509.

Riley, R. D., Lambert, P., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *British Medical Journal*, *340*, 521–525.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, *6*, 15–51.

SAS Institute. (2014). *SAS* (Version 9.4) [Computer Software]. https://www.sas.com

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods*, *18*, 385–405.

Snijders, T. A., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*, 237–259.

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods*, *44*, 1244–1254.

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van den Noortgate, W. (2014). Bias corrections for standardized effect size estimates used with single-subject experimental designs. *The Journal of Experimental Education*, *82*, 358–374.

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 designs. *Evidence-Based Communication Assessment and Intervention*, *2*, 142–151.

Wang, S.-Y., 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.

Authors

SEANG-HWANE JOO is an assistant professor at the University of Kansas, 1122 West Campus Rd., Lawrence, KS 66045, USA; email: sjoo@ku.edu. His research interests are psychometrics and statistical methods in education.

YAN WANG is an assistant professor at the University of Massachusetts Lowell, 220 Pawtucket St., Lowell, MA 01854, USA; email: yan_wang1@uml.edu. Her research interests are structural equation modeling and psychometrics.

JOHN FERRON is a professor at the University of South Florida, 4202 E. Fowler Ave., Tampa, FL 33620, USA; email: ferron@usf.edu. His research interests are single-case research and statistical methods.

S. NATASHA BERETVAS is a professor at the University of Texas, 110 Inner Campus Dr., Austin, TX 78712, USA; email: tberetvas@austin.utexas.edu. Her research interests are meta-analysis and quantitative research methods.

MARIOLA MOEYAERT is an associate professor at the State University of New York, Albany, 1400 Washington Ave, Albany, NY 12222, USA; email: mmoeyaert@albany.edu. Her research interests are single-case research and multilevel modeling.

WIM VAN DEN NOORTGATE is a professor at Katholieke Universiteit Leuven, Belgium, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium; email: wim.vandennoortgate@kuleuven.be. His research interests are meta-analysis and item response theory.

Manuscript received August 22, 2017
First revision received April 23, 2019
Second revision received December 18, 2020
Third revision received May 20, 2021
Accepted July 9, 2021