Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Multilevel Meta-Analysis of Individual Participant Data of Single-Case Experimental Designs: One-Stage versus Two-Stage Methods

Lies Declercq , Laleh Jamshidi , Belén Fernández Castilla , Mariola Moeyaert , S. Natasha Beretvas , John M. Ferron & Wim Van den Noortgate

To cite this article: Lies Declercq , Laleh Jamshidi , Belén Fernández Castilla , Mariola Moeyaert , S. Natasha Beretvas , John M. Ferron & Wim Van den Noortgate (2020): Multilevel Meta-Analysis of Individual Participant Data of Single-Case Experimental Designs: One-Stage versus Two-Stage Methods, Multivariate Behavioral Research, DOI: 10.1080/00273171.2020.1822148

To link to this article: https://doi.org/10.1080/00273171.2020.1822148

Published online: 30 Sep 2020.

MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2020.1822148

Multilevel Meta-Analysis of Individual Participant Data of Single-Case Experimental Designs: One-Stage versus Two-Stage Methods

Lies Declercq[a,b], Laleh Jamshidi[a,b], Belén Fernández Castilla[a,b], Mariola Moeyaert[c], S. Natasha Beretvas[d], John M. Ferron[e], and Wim Van den Noortgate[a,b]

[a]Faculty of Psychology and Educational Sciences, KU Leuven; [b]ITEC, imec research group, KU Leuven; [c]Department of Educational Psychology and Methodology, University at Albany, Albany, NY; [d]Department of Educational Psychology, University of Texas; [e]Department of Educational Measurement and Research, University of South Florida

ABSTRACT

To conduct a multilevel meta-analysis of multiple single-case experimental design (SCED) studies, the individual participant data (IPD) can be analyzed in one or two stages. In the one-stage approach, a multilevel model is estimated based on the raw data. In the two-stage approach, an effect size is calculated for each participant and these effect sizes and their sampling variances are subsequently combined to estimate a meta-analytic multilevel model. The multilevel model in the two-stage approach has fewer parameters to estimate, in exchange for the reduction of information of the raw data to effect sizes. In this paper we explore how the one-stage and two-stage IPD approaches can be applied in the context of meta-analysis of single-case designs. Both approaches are compared for several single-case designs of increasing complexity. Through a simulation study we show that the two-stage approach obtains better convergence rates for more complex models, but that model estimation does not necessarily converge at a faster speed. The point estimates of the fixed effects are unbiased for both approaches across all models, as such confirming results from methodological research on IPD meta-analysis of group-comparison designs. In light of these results, we discuss the implementation of both methods in R.

KEYWORDS

Single-case experimental design; effect size; multilevel modeling; meta-analysis; individual participant data

Introduction

In a single-case experimental design (SCED), an outcome is repeatedly measured within one subject, case or entity under multiple conditions or phases. These conditions differ due to an intervention or treatment which is introduced (and in some designs also withdrawn again) by the experimenter. For example, in a single-case design by Fiala and Sheridan (2003), curriculum-based measurement probes were collected for children with below-average reading skills during two phases. The first phase is the so-called baseline phase, which is followed by a treatment phase in which the children do paired reading sessions with their parents four times a week. Typically, reports on SCED studies, like the paper by Fiala and Sheridan (2003), include a time series plot for each of the cases (Figure 1). This practice results in an important advantage for secondary analysis or meta-analysis: by means of software, one can fairly easily retrieve the individual participant data (IPD) from these graphs and directly analyze the raw data. This is a substantial asset, especially when the goal is to combine data from multiple participants from different SCED studies in order to synthesize individual findings in a meta-analysis.

In a traditional so-called aggregated data (AD) meta-analysis, the unit of analysis is the summary statistics reported in the primary studies (Cooper & Patall, 2009). Over the years, and especially since the digital revolution has enabled sharing of (individual participant) data on a larger scale, the AD meta-analytic approach has been scrutinized and consensus came to be that IPD meta-analysis is to be preferred instead (Cooper & Patall, 2009; Stewart & Tierney, 2002; Tudur Smith et al., 2016). Given that the individual participant data correspond exactly to the raw data used to calculate the aggregated data, IPD meta-analysis can not only exactly reproduce the corresponding AD meta-analytic results, but also offers several additional advantages. When conducting an IPD meta-analysis, researchers are able to check the raw data and standardize it across studies. In some cases, primary studies might report summary

CONTACT Lies Declercq lies.declercq@kuleuven.be Faculty of Psychology and Educational Sciences, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium.

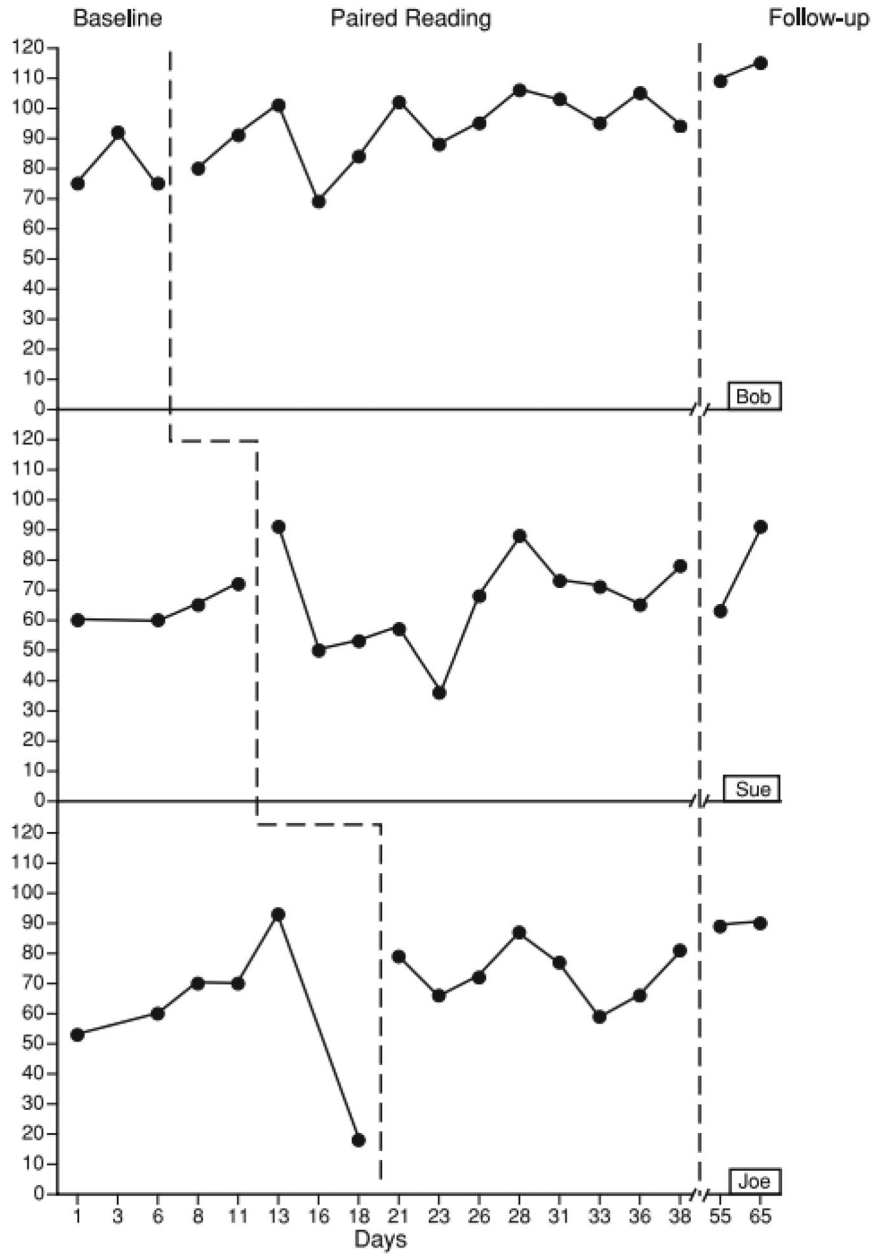

Figure 1. Illustration of raw SCED data available as a time series graph. Reprinted from the multiple-baseline study by Fiala and Sheridan (2003) assessing the effect of paired reading with a parent on reading skills in children.

statistics only on a selection of measured outcomes. By obtaining the IPD, all measured outcomes are available for meta-analysis and outcome reporting bias can be reduced. With the raw data, researchers can perform more detailed subgroup analyses with covariates at the participant level. This leads to higher powered analyses with less risk of ecological bias and also offers the possibility to add moderators on both between- and within-study level (Berlin et al., 2002; Cooper & Patall, 2009; Debray et al., 2015).

Despite the theoretical advantages of IPD meta-analysis, in practice AD meta-analysis has not yet been widely displaced by IPD meta-analysis. (Cooper & Patall, 2009). This is due to the fact that AD is more commonly available in published studies and that the cost of retrieving IPD is usually much higher in terms of time and resources. This is especially true in social sciences, where it is much less likely that researchers are able to retrieve the full set of IPD underlying group-level statistics in published reports (Cooper & Patall, 2009). In the context of single-case research however, these concerns are much less of an issue. The highly adaptable single-case design is associated with many very diverse effect sizes (Hedges

et al., 2013; Manolov & Moeyaert, 2017), and because it is not trivial to standardize them and to make them comparable across studies, combining them in a meaningful AD meta-analysis is equally non-trivial. Moreover, an important drawback of IPD meta-analysis in group-comparison contexts, namely the high time and resource costs of retrieving the raw data, is much less of an issue within single-case studies. With the raw data commonly available as time series graphs in the primary studies, there is no need to contact authors of primary studies and to rely on their goodwill in sharing their data sets. A review of SCED meta-analyses by Jamshidi et al. (2020) confirms that in 69% of the reviewed studies, data were retrieved from graphs in the primary studies, indicating indeed that SCED meta-analysis is often conducted on IPD rather than on summary statistics or effect sizes.

Two approaches are common when conducting an IPD meta-analysis (Burke et al., 2017; Debray et al., 2015). In the two-stage approach, the IPD are aggregated within studies and a study-specific estimate of the treatment effect is obtained. Then, meta-analytic fixed or random effects methods are used to obtain an overall treatment effect estimate by calculating a weighted average of the study-specific estimates. As opposed to an AD meta-analysis based on effect sizes originally reported in the primary studies, a two-stage IPD meta-analysis has the advantage that it allows for choosing and calculating one particular effect size measure across all participants and studies, as such avoiding the need for transformation or standardization of different effect size measures across studies. In a one-stage IPD meta-analysis on the other hand, all individual observations are analyzed in a hierarchical or multilevel model, which accounts for within-study dependencies.

Cooper and Patall (2009) have drawn the parallel between IPD meta-analysis and SCED meta-analysis. In literature on SCED meta-analysis (and in the field of social sciences in general), the IPD terminology is not commonly used. In line with Cooper and Patall (2009), we choose to adopt the terms 'individual participant data' or 'IPD' (with 'participant' rather than 'patient') as well as 'one-stage' and 'two-stage' meta-analysis as used in biomedical sciences, in order to promote cross-disciplinary communication by using consistent terminology across disciplines. Our aim with this article is to explore how the one-stage and two-stage approaches can be applied to SCED meta-analysis and whether similar issues, as described extensively in methodological literature on the one-stage versus two-stage approach in the field of medicine, arise. In the following sections we explain in more detail how the one-stage and two-stage IPD approach correspond to practices in SCED meta-analysis.

One- and two-stage multilevel meta-analysis of SCED data

The distinction between one-stage and two-stage IPD meta-analysis stem from medical sciences, where the randomized controlled trial (RCT) is considered as the golden standard. When combining RCTs across separate but similar clinical studies, the structure of the data is a two-level hierarchy of participants nested within trials or studies. This structure is similar to that of the data from one SCED study, because—despite the nomenclature—in single-case research it is common practice to replicate the design across a small number of participants (Shadish & Sullivan, 2011). However, since participants are measured at several time points, the first level is that of the observations which are nested in participants at the second level. When synthesizing several SCED studies with multiple participants, an additional level of dependency is introduced: measurements are nested within participants which are nested within studies. To account for this nested structure, Van den Noortgate & Onghena (2003b, 2003a, 2003c, 2007, 2008) proposed a hierarchical linear model with three levels to synthesize the raw SCED data from multiple cases. Since all raw data are combined into a single model, this approach can be interpreted as a one-stage IPD method.

However, the review by Jamshidi et al. (2020) suggests that in practice, SCED meta-analysis is not conducted in a one-stage fashion. Although retrieval of the IPD through reverse-engineering of the reported times series graphs is common practice, only a minority of the reviewed meta-analysis (12%) directly used the retrieved raw IPD to synthesize the primary SCED studies. Instead, a vast majority of the reviewed meta-analyses (90%) used effect sizes (e.g., percentages of non-overlapping data, improvement rate differences or standardized mean differences). This implies that researchers most commonly apply a two-stage IPD approach, where they first retrieve raw data from time-series graphs in the primary studies, subsequently calculate effect sizes and finally combine those effect sizes in a meta-analytic model. Van den Noortgate and Onghena (2008) illustrate an alternative to the one-stage approach which allows for statistically combining effect sizes from SCED studies. They propose to use a standardized mean difference (obtained

from a regression coefficient) as an effect size measure expressing the effect of the treatment for a particular case. These effect sizes are then combined in a three-level meta-analytical model and as such this procedure comes down to a two-stage IPD approach. The standardized mean differences proposed by Van den Noortgate and Onghena (2008) are particularly useful when the scale of the outcome is not the same across the cases and studies being aggregated. In situations where the outcome is on the same scale, however, not standardizing is preferable due to the fact that the error introduced through estimation of the standardization quantity is avoided. The three-level meta-analytical model as applied by Van den Noortgate and Onghena (2008) can therefore also be an option for alternative (unstandardized) effect sizes for SCED data, such as other mean difference indices, regression coefficients or descriptive metrics indexing immediate changes in level or changes in slope.

Objectives

In this article we systematically compare a one- and a two-stage approach for a multilevel IPD meta-analysis of SCED data. For three models of increasing complexity, we simulate datasets and apply both approaches. For more complex models, the three-level models involve more regression coefficients and therefore more parameters to estimate. This is particularly true for the variance components, since the dimensions of the covariance matrices at the higher level(s) increase quickly; if a model has n coefficients, adding one coefficient results in one additional variance and $2n$ additional covariances (i.e., $2n+1$ additional parameters) to estimate. The two-stage approach has an important potential benefit over the one-stage approach when the underlying model is more complex. The multilevel model estimated based on the effect sizes is reduced, so there are less parameters to estimate. This might result in faster estimation procedures and better convergence rates compared to the one-stage approach.

However, a drawback of the two-stage approach is the loss of information by reducing the rich raw data to effect sizes. In a three-level design with I observations nested in J cases nested in K studies, the two-stage approach uses $I \times J \times K$ raw data points, while the one-stage approach only uses $J \times K$ effect sizes. It is not clear if the reduction in data combined with the smaller model in the one-stage approach will result in better or worse performance compared to the two-stage approach. Possibly a less complex model performs actually better than a complex model when the available information is sparse. This might also be the case for the two-stage approach. To investigate this further, we compare the performance of both approaches by assessing the statistical properties of the estimations (the bias, MSE, confidence interval coverage, Type I error rate and size of the standard errors), the convergence rate and convergence speed. Below, we start by describing in more detail the one-stage and two-stage approach for a basic SCED meta-analysis.

Three-level IPD analysis of SCED data

One-stage approach

Suppose SCED measurements Y have been obtained from a two-phased design. First, the outcome Y is measured in an initial baseline phase without any intervention. After a number of measurement occasions, the treatment or intervention is implemented and the outcome continues to be measured under the treatment condition. Suppose we have measurements Y from multiple two-phased designs replicated across cases and across studies. If we denote the measurement occasion by i, the participant as j and the study as k, we can model the measurements Y_{ijk} at the highest level as

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} D_{ijk} + e_{ijk}, \tag{1}$$

where the regression coefficients β_{0jk} and β_{1jk} are specific to case j from study k. The regressor D_{ijk} is a dummy variable which equals 0 in the baseline phase and 1 after the treatment has been implemented. The residuals e_{ijk} are independent and identically normally distributed with mean 0 and variance σ_e^2, assuming they are not autocorrelated (i.e., there is no dependency between errors due to similarity between consecutive observations). The case-specific regression coefficients β_{0jk} and β_{1jk} can be decomposed into a fixed effect $\gamma_{...}$, a case-specific random effect $u_{.jk}$ and a study-specific random effect $v_{..k}$:

$$\begin{cases} \beta_{0jk} = \gamma_{000} + u_{0jk} + v_{00k} \\ \beta_{1jk} = \gamma_{100} + u_{1jk} + v_{10k} \end{cases}$$

$$\begin{pmatrix} u_{0jk} \\ u_{1jk} \end{pmatrix} \sim \text{MVN}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u0}^2 & \\ \sigma_{u01} & \sigma_{u1}^2 \end{pmatrix} \right], \tag{2}$$

$$\begin{pmatrix} v_{00k} \\ v_{10k} \end{pmatrix} \sim \text{MVN}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{v0}^2 & \\ \sigma_{v01} & \sigma_{v1}^2 \end{pmatrix} \right].$$

Often in SCED analysis, special interest goes out to the estimation of the fixed effect parameter γ_{100}, which expresses the overall average treatment effect on the intercept. Furthermore, SCED practitioners might be interested in estimating the between-subject variance σ^2_{u1} and within-study variance σ^2_{v1} of this treatment effect.

Two-stage approach

In the first stage of the two-stage approach, effect sizes need to be calculated from the raw IPD data. Van den Noortgate and Onghena (2008) illustrate how to obtain effect size measures from SCED data, based on a simple linear regression model per case and using the ordinary least squares (OLS) method. For the two-phase SCED data as described in the previous paragraph, the simple linear regression model is identical to the first-level equation of the one-stage multilevel model (Equation 1):

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} D_{ijk} + e_{ijk} \tag{3}$$

We denote the residual variance per case as $\sigma^2_{e_{jk}}$. In the baseline phase, when $D_{ijk} = 0$, the expected score for case jk is β_{0jk}. In the treatment phase, when $D_{ijk} = 1$, the expected score for case jk changes to $\beta_{0jk} + \beta_{1jk}$. As such β_{1jk} can be interpreted as an effect size of case jk. From the fitted model, we obtain an estimation b_{1jk} of the effect size β_{1jk}. The effect sizes b_{1jk} can be used in an alternative three-level meta-analytical model. We assume that the effect sizes b_{1jk} vary around β_{1jk}, the 'true' effect for case jk with some random error r_{jk}:

$$b_{1jk} = \beta_{1jk} + r_{jk}. \tag{4}$$

The within-case sampling variances are considered to be known because they have been estimated as $\hat{\sigma}^2_{b_{1jk}}$ in the first step of the two-stage approach (Equation 3). Note that in the one-stage approach, the residuals at the first level are assumed to have a common variance σ^2_e across cases and studies. In the two-stage approach, the within-case variances (i.e., the diagonal elements of $\hat{\sigma}^2_{b_{1jk}}$) are not assumed to be identical. The precision of these estimations will largely depend on the number of measurements I. The individual 'true' effect sizes β_{1jk} can be decomposed into an overall fixed effect γ_{100}, a case-specific random effect u_{1jk} and a study-specific random effect v_{10k} :

$$\beta_{1jk} = \gamma_{100} + u_{1jk} + v_{10k}. \tag{5}$$

The random effects u_{1jk} and v_{10k} are assumed to be univariate normally distributed with means 0 and variances σ^2_{u1} and σ^2_{v1} respectively.

An important difference between the three-level model in the one-stage approach (Equations 1 and 2) and this three-level model is that there are fewer model parameters to estimate: only one fixed effect (γ_{100}) instead of two (γ_{000} and γ_{100}), and only two variance components (σ^2_{u1} and σ^2_{v1}) instead of seven ($\sigma^2_e, \sigma^2_{u0}, \sigma_{u01}, \sigma^2_{u1}, \sigma^2_{v0}, \sigma_{v01}$ and σ^2_{v1}). Since multilevel models can take long to estimate, especially when the model is complex and involves many parameters, the smaller three-level model might significantly reduce the estimation time of the two-stage approach compared to the one-stage approach. A downside of the two-stage approach is that the effect sizes are aggregated data: the three-level model in the two-stage approach is based on only $J \times K$ observations b_{1jk}, whereas the three-level model in the one-stage approach is based on $I \times J \times K$ observations Y_{ijk}. It is unclear whether this loss of information will affect the statistical properties of the estimations of the two-stage approach.

In the following section, we illustrate two more complicated SCED multilevel models and how the two-stage approach can be adapted accordingly. Note that although standardization will in practice often be required due to studies measuring the dependent variable in different ways, we focus on unstandardized raw data and effect sizes in this simulation study. Standardization for both raw data and effect sizes as illustrated by Van den Noortgate and Onghena (2008) happens before the application of the meta-analytical three-level models and with an identical standardization quantity (i.e., the root mean squared error from the individual OLS regression estimations per case) and would therefore not have an impact on the -comparison between the one-stage and two-stage approach.

More complex three-level models for SCED data

SCED data with a linear time trend

The simple models in Equations 1 and 3 only have an intercept and one predictor D. They assume a horizontal average trajectory in both the baseline and the treatment phase. Suppose we want to model a linear time trend in both phases. The one-stage model is extended by adding a covariate T_{ijk} to the first level. The time can be expressed by real time (e.g., days) or as a time indication (e.g., session number). Modeling the baseline trajectory as a straight line with intercept γ_{000} and a slope γ_{100}, the treatment can have an effect on either of them. Therefore the second model includes not only the treatment dummy D_{ijk} as a

covariate, but also $D_{ijk}T_{ijk}$:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}T_{ijk} + \beta_{2jk}D_{ijk} + \beta_{3jk}D_{ijk}T_{ijk} + e_{ijk} \tag{6}$$

where e_{ijk} again has a $N(0, \sigma_e^2)$ distribution. Assuming the time variable T_{ijk} is centered around the first observation in the treatment phase, β_{2jk} can be interpreted as the case-specific average effect on the intercept and β_{3jk} as the case-specific average effect on the slope. Again all covariates are decomposed into a fixed effect, a random effect at the case level and a random effect at the study level:

$$\begin{aligned} \boldsymbol{\beta}_{jk} &= \boldsymbol{\gamma} + \mathbf{u}_{jk} + \mathbf{v}_{0k} \\ \mathbf{u}_{jk} &\sim \text{MVN}(0, \boldsymbol{\Sigma}_{\mathbf{u}}), \\ \mathbf{v}_{0k} &\sim \text{MVN}(0, \boldsymbol{\Sigma}_{\mathbf{v}}) \end{aligned} \tag{7}$$

The bold symbols represent multidimensional vectors: $\boldsymbol{\beta}_{jk}$ equals $(\beta_{0jk}, \beta_{1jk}, \beta_{2jk}, \beta_{3jk})^\top$, $\boldsymbol{\gamma}$ equals $(\gamma_{000}, \gamma_{100}, \gamma_{200}, \gamma_{300})^\top$, $\mathbf{u}_{jk}$ equals $(u_{0jk}, u_{1jk}, u_{2jk}, u_{3jk})^\top$, $\mathbf{v}_{0k}$ equals $(v_{00k}, v_{10k}, v_{20k}, v_{30k})^\top$, Σ_u equals

$$\begin{pmatrix} \sigma^2_{u0} & & & \\ \sigma_{u01} & \sigma^2_{u1} & & \\ \sigma_{u02} & \sigma_{u12} & \sigma^2_{u2} & \\ \sigma_{u03} & \sigma_{u13} & \sigma_{u23} & \sigma^2_{u3} \end{pmatrix}$$

and Σ_v equals

$$\begin{pmatrix} \sigma^2_{v0} & & & \\ \sigma_{v01} & \sigma^2_{v1} & & \\ \sigma_{v02} & \sigma_{v12} & \sigma^2_{v2} & \\ \sigma_{v03} & \sigma_{v13} & \sigma_{v23} & \sigma^2_{v3} \end{pmatrix}$$

Estimating this model directly based on the raw data involves the estimation of four fixed effects (the elements of $\boldsymbol{\gamma}$) and 21 variance components (σ_e^2 and all the elements of Σ_u and Σ_v). The fixed effects of interest are those related to the effect of treatment: γ_{200}, which expresses the overall average effect of the treatment on the intercept, and γ_{300}, which expresses the overall average effect of the treatment on the slope. The difference with the previous intercept-only model is that for this linear time trend model, the effect of the treatment can no longer be expressed in one single effect size (i.e., γ_{100} for the intercept-only model). Per case, we will have two effect sizes: γ_{200} and γ_{300}.

Analogously to the two-stage approach for the intercept-only model, we can calculate per case effect sizes by fitting a multiple OLS linear regression model for each case *jk*:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}T_{ijk} + \beta_{2jk}D_{ijk} + \beta_{3jk}D_{ijk}T_{ijk} + e_{ijk} \tag{8}$$

This model yields estimates b_{2jk} and b_{3jk} and these estimates will serve as effect sizes for the subsequent three-level meta-analytic model. Because we now have two effect sizes per case, the multilevel model will be a multivariate model assuming a joint two-dimensional distribution of the b_{2jk}'s and b_{3jk}'s:

$$\begin{aligned} &\begin{cases} b_{2jk} = \gamma_{200} + u_{2jk} + v_{20k} + r_{2jk} \\ b_{3jk} = \gamma_{300} + u_{3jk} + v_{30k} + r_{3jk} \end{cases} \\ &\begin{pmatrix} r_{2jk} \\ r_{3jk} \end{pmatrix} \sim \text{MVN}\big(0, \hat{\boldsymbol{\sigma}}^2(\mathbf{b})\big), \\ &\begin{pmatrix} u_{2jk} \\ u_{3jk} \end{pmatrix} \sim \text{MVN}(0, \boldsymbol{\Sigma}_{\mathbf{u}}), \\ &\begin{pmatrix} v_{20k} \\ v_{30k} \end{pmatrix} \sim \text{MVN}(0, \boldsymbol{\Sigma}_{\mathbf{v}}) \end{aligned} \tag{9}$$

The covariance matrix of the sampling errors r_{2jk} and r_{3jk} is estimated in the first step of the two-stage approach (Equation 8) and is denoted as $\hat{\boldsymbol{\sigma}}^2(\mathbf{b})$. The three covariance matrices $\hat{\boldsymbol{\sigma}}^2(\mathbf{b}), \boldsymbol{\Sigma}_{\mathbf{u}}$ and $\boldsymbol{\Sigma}_{\mathbf{v}}$ are elements of $\mathbb{R}^{2\times2}$. The random effects u_{2jk} and u_{3jk} express the deviation of participant *jk* from the mean of study *k*. Finally, the random effects v_{20k} and v_{30k} express the deviation of the mean of study *k* from the overall mean effect sizes γ_{200} and γ_{300}. Estimating the meta-analytic multilevel model in Equation 9 involves estimating two fixed effects (γ_{200} and γ_{300}) and six (co)variance components (the elements of $\boldsymbol{\Sigma}_{\mathbf{u}}$ and $\boldsymbol{\Sigma}_{\mathbf{v}}$), which is substantially less than the number of parameters to be estimated when applying the one-stage approach on the SCED raw data.

SCED data with a quadratic time trend

An extension of the linear time trend model (Equations 6 and 7) is the quadratic time trend model, in which the trajectory is allowed to be parabolic (as well as linear). On top of the intercept and the first-order time covariate, a second order time covariate (plus its interaction with the treatment dummy variable) is included at the first level. The full three-level quadratic time trend model is

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_{1jk}T_{ijk} + \beta_{2jk}T^2_{ijk} + \beta_{3jk}D_{ijk} + \beta_{4jk}D_{ijk}T_{ijk} \\ &\quad + \beta_{5jk}D_{ijk}T^2_{ijk} + e_{ijk} \\ \boldsymbol{\beta}_{jk} &= \boldsymbol{\gamma} + \mathbf{u}_{jk} + \mathbf{v}_{0k} \\ \mathbf{u}_{jk} &\sim \text{MVN}(0, \boldsymbol{\Sigma}_{\mathbf{u}}), \\ \mathbf{v}_{0k} &\sim \text{MVN}(0, \boldsymbol{\Sigma}_{\mathbf{v}}) \end{aligned} \tag{10}$$

Note that in this model the parameters $\boldsymbol{\beta}_{jk}, \boldsymbol{\gamma}, \mathbf{u}_{jk}$ and $\mathbf{v}_k$ belong to $\mathbb{R}^6$, while $\boldsymbol{\Sigma}_{\mathbf{u}}$ and $\boldsymbol{\Sigma}_{\mathbf{v}}$ are 6×6 covariance matrices. The fixed effect parameters of interest are those related to the treatment effect on the baseline's intercept (γ_{300}) and its first and second-

Table 1. Three-level estimation model specifications.The index *i* denotes the measurement occasion, *j* the case and *k* the study.

Model	Specification[a]	Covariance structure[b]	# Fixed effects		# (Co)variance parameters	
			One-stage	Two-stage	One-stage	Two-stage
1	$Y_{ijk} = \gamma_{000} + u_{0jk} + v_{00k} + (\gamma_{100} + u_{1jk} + v_{10k})D_{ijk} + e_{ijk}$	$\begin{pmatrix} \sigma^2_{.0} & \\ \sigma_{.01} & \sigma^2_{.1} \end{pmatrix}$	2	1	7	2
2	$Y_{ijk} = \gamma_{000} + u_{0jk} + v_{00k} + (\gamma_{100} + u_{1jk} + v_{10k})T_{ijk} + (\gamma_{200} + u_{2jk} + v_{20k})D_{ijk} + (\gamma_{300} + u_{3jk} + v_{30k})D_{ijk}T_{ijk} + e_{ijk}$	$\begin{pmatrix} \sigma^2_{.0} & & & \\ \sigma_{.01} & \sigma^2_{.1} & & \\ \sigma_{.02} & \sigma_{.12} & \sigma^2_{.2} & \\ \sigma_{.03} & \sigma_{.13} & \sigma_{.23} & \sigma^2_{.3} \end{pmatrix}$	4	2	21	6
3	$Y_{ijk} = \gamma_{000} + u_{0jk} + v_{00k} + (\gamma_{100} + u_{1jk} + v_{10k})T_{ijk} + (\gamma_{200} + u_{2jk} + v_{20k})T^2_{ijk} + (\gamma_{300} + u_{3jk} + v_{30k})D_{ijk} + (\gamma_{400} + u_{4jk} + v_{40k})D_{ijk}T_{ijk} + (\gamma_{500} + u_{5jk} + v_{50k})D_{ijk}T^2_{ijk} + e_{ijk}$	$\begin{pmatrix} \sigma^2_{.0} & & & & & \\ \sigma_{.01} & \sigma^2_{.1} & & & & \\ \sigma_{.02} & \sigma_{.12} & \sigma^2_{.2} & & & \\ \sigma_{.03} & \sigma_{.13} & \sigma_{.23} & \sigma^2_{.3} & & \\ \sigma_{.04} & \sigma_{.14} & \sigma_{.24} & \sigma_{.34} & \sigma^2_{.4} & \\ \sigma_{.05} & \sigma_{.15} & \sigma_{.25} & \sigma_{.35} & \sigma_{.45} & \sigma^2_{.5} \end{pmatrix}$	6	3	43	12

[a]For simulating the raw data sets, all nominal fixed effects not related to the treatment were set to 1 and all nominal fixed effects related to the treatment were identical and equal to $\gamma = 0$ or $\gamma = 2$.
[b]The covariance structure is identical at the case and the study level. The residual variance σ_e was fixed to 1, the case- and study-level variances were equal to $\sigma^2 = 1$ or $\sigma^2 = 4$ and the case- and study-level covariances are equal to $\sigma^2\rho$ with $\rho = 0$ or $\rho = 0.5$.

order time coefficients (γ_{400} and γ_{500}). Estimating this model for the raw data involves the estimation of six fixed effects (the elements of $\boldsymbol{\gamma}$) and 43 variance components (the elements of $\boldsymbol{\Sigma}_{\mathbf{u}}$ and $\boldsymbol{\Sigma}_{\mathbf{v}}$ and the residual variance σ_e^2).

To apply the two-stage approach while modeling a quadratic time trend, effect sizes can again be calculated per case jk with a multiple OLS linear regression model:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}T_{ijk} + \beta_{2jk}T_{ijk}^2 + \beta_{3jk}D_{ijk} + \beta_{4jk}D_{ijk}T_{ijk} + \beta_{5jk}D_{ijk}T_{ijk}^2 + e_{ijk} \tag{11}$$

where jk does not vary. From these per-case model fits, we obtain three effect sizes b_{3jk}, b_{4jk} and b_{5jk}, and a sampling covariance matrix $\hat{\boldsymbol{\sigma}}^2(\mathbf{b})$ per case, which we combine again in a multivariate, three-level meta-analytic model:

$$\begin{cases} b_{3jk} = \gamma_{300} + u_{3jk} + v_{30k} + r_{3jk} \\ b_{4jk} = \gamma_{400} + u_{4jk} + v_{40k} + r_{4jk} \\ b_{5jk} = \gamma_{500} + u_{5jk} + v_{50k} + r_{5jk} \end{cases}$$
$$\begin{pmatrix} r_{3jk} \\ r_{4jk} \\ r_{5jk} \end{pmatrix} \sim \text{MVN}\big(0, \hat{\boldsymbol{\sigma}}^2(\mathbf{b})\big),$$
$$\begin{pmatrix} u_{3jk} \\ u_{4jk} \\ u_{5jk} \end{pmatrix} \sim \text{MVN}(0, \boldsymbol{\Sigma}_{\mathbf{u}}),$$
$$\begin{pmatrix} v_{30k} \\ v_{40k} \\ v_{50k} \end{pmatrix} \sim \text{MVN}(0, \boldsymbol{\Sigma}_{\mathbf{v}}) \tag{12}$$

The covariance matrices $\hat{\boldsymbol{\sigma}}^2(\mathbf{b}), \boldsymbol{\Sigma}_{\mathbf{u}}$ and $\boldsymbol{\Sigma}_{\mathbf{v}}$ are now elements of $\mathbb{R}^{3\times3}$. Estimating the meta-analytic multilevel model in Equation 12 involves estimating three fixed effects (γ_{300}, γ_{400} and γ_{500}) and 12 variance components (the elements of $\boldsymbol{\Sigma}_{\mathbf{u}}$ and $\boldsymbol{\Sigma}_{\mathbf{v}}$).

Methodology

All three described one-stage models (Equations 1 and 2 for Model 1, Equations 6 and 7 for Model 2 and Equation 10 for Model 3) are summarized in Table 1. These were the models used to estimate the data. For simulating raw data for each of these models, the coefficients of the covariates not related to the treatment were fixed to 1 (i.e., γ_{000} for Model 1, γ_{000} and γ_{100} for Model 2, and γ_{000}, γ_{100} and γ_{200} for Model 3 in Table 1). This means that in the baseline phase, all coefficients in all models equal 1. The effect sizes, i.e., the coefficients of the covariates related to the treatment (i.e., γ_{100} for Model 1, γ_{200} and γ_{300} for Model 2, and γ_{300}, γ_{400} and γ_{500} for Model 3 in Table 1), were set simultaneously to either 2 or 0. This means that either the treatment has effect on all baseline covariates, or on none of them. For Σ_u and Σ_v, the covariance matrices on the second and third level, we chose a compound symmetry structure to generate the raw data. The variances (i.e., the diagonal elements in the matrices in Table 1) were set to 1 or 4 at both levels, and the correlations to 0 or 0.5 at both levels (thus the off-diagonal elements in the matrices in Table 1 equaled 0, 0.5 or 2). The residual variance at the first level σ_e^2 was fixed at 1. For Models 2 and 3 the time variable was coded to be centered around the first treatment observation. We varied the size of the simulated datasets by varying the number of measurement occasion I ($I = 20$, 28 or 40), the number of cases J ($J = 3$, 5 or 10) and the number of studies K ($K = 5$, 7 or 10). The values for I and J were chosen based on the review of SCED studies by Shadish and Sullivan (2011). The number of studies K was kept fairly small compared to values reported in a recent review of SCED meta-analyses (Jamshidi et al., 2020), so that the combination of smaller datasets and more complex models provided a challenge for both the one-stage and the two-stage approach in terms of convergence and estimation speed. The datasets within studies were designed to have an identical number of baseline and treatment phase observations, so the intervention was set to take place after the first half of I observations within each case. In total these parameter variations lead to $2 \times 2 \times 2 \times 3 \times 3 \times 3 = 216$ conditions. For each condition, we simulated 1000 datasets with raw data which we analyzed with the one-stage approach. For each dataset, we then calculated effect sizes and analyzed those using the two-stage approach.

The simulation study was implemented in R. To generate the raw data and fit the three-level model for the one-stage approach, we used the `lmer` function from the `lme4` package (Bates et al., 2015). Single parameter hypothesis testing for the one-stage approach was done with the contrast testing function `contest` from the `lmerTest` package. For the two-stage approach, we calculated effect sizes by fitting a multiple linear regression model (using OLS) based on the first level of the three-level model used to generate the raw data. These effect sizes were combined in a three-level model with the function `rma.mv` from the `metafor` package (Viechtbauer, 2010). We chose not to use the `lmer` function here. The reason for this is that in the three-level meta-analytic model for the effect sizes, the sampling variances at the first level are assumed to be known and need

to be fixed at these known values in the model. However, the `lmer` function does not allow specification of a known variance-covariance matrix for the sampling errors. It only allows fixing the sampling variances up to a proportionality constant via the `weights` argument (Viechtbauer, 2016). Therefore we opted to use the `rma.mv` function from the `metafor` package instead, because `rma.mv` allows for specifying known sampling variances as weights. We made sure to implement both functions with identical options: both functions used REML estimation and the BOBYQA algorithm (Powell, 2009) with a maximum of 10.000 function evaluations as the estimation optimization method. The simulation code can be requested from the last author of this paper.

To evaluate and compare the one-stage and the two-stage approach, we calculated the bias, the MSE, the coverage proportion of the 95% confidence interval, the Type I error rate and the bias in the standard errors of the relevant fixed effects of interest (i.e., γ_{100} for Model 1, γ_{200} and γ_{300} for Model 2 and γ_{400}, γ_{500} and γ_{600} for Model 3). For the associated variance components (i.e., the within- and between-study variances and covariances related to the relevant fixed effect coefficients), we looked at the bias and the mean squared error only.

To calculate *p*-values and confidence intervals, `lmer` and `rma.mv` do not offer the same options in terms of the type of hypothesis test. The `lme4` package by default offers likelihood ratio tests, profile confidence intervals and parametric bootstrap confidence intervals, but it does not offer *p*-values or Wald-type confidence intervals (Bates, 2006; Bates et al., 2015; Bolker, 2018). For Wald *t*-tests, users can resort to the `lmerTest` package (Kuznetsova et al., 2017), which offers *t*-tests with the Satterthwaite or Kenward-Roger adjustment for the degrees of freedom (Kenward & Roger, 1997; Satterthwaite, 1941). The `metafor` package by default offers likelihood ratio tests and Wald-type *z*-tests. It also includes the option to adjust the standard errors of the estimated coefficients by mimicking the Knapp and Hartung (2003) method and performing a t-test instead (note that unlike in the actual Knapp and Hartung (2003) method, the covariance matrix of the fixed effects is not adjusted). With some manual coding effort, parametric and non-parametric bootstrap confidence intervals can also be obtained with `metafor` (Viechtbauer, 2018).

Because of the differences in functionality and options regarding inference in `lme4` and `metafor`, we tried to streamline the inference procedure for the one-stage and two-stage approach as much as possible. As a first step, we conducted a simple Wald *z*-test based on the point estimates and the corresponding standard errors. Although the *z*-test's normality assumption for the null distribution does not hold for finite samples, this was the only way to make a fair comparison between the one-stage and two-stage approach. Secondly, we conducted a *t*-test for both approaches: for the one-stage approach we used the extended `lmer` function and the contrast testing function `contest` from the package `lmerTest` (Kuznetsova et al., 2017) to conduct single parameter hypothesis testing based on a Wald-type *t*-test with Satterthwaite degrees of freedom (Satterthwaite, 1941) and to calculate the associated confidence interval limits. For the two-stage approach we used the option `test = "t"` of the `rma.mv` function, which returns a *t*-test with degrees of freedom adjusted by mimicking the Knapp and Hartung (2003) method, namely $KJ - p$, where p is the total number of model coefficients including the intercept if it is present (Viechtbauer, 2010).

Furthermore we evaluated and compared the one-stage and two-stage approach in terms of speed and convergence. Because both approaches were implemented with functions from different packages, comparisons between them in terms of speed and convergence rely greatly on the implementation of the specific functions, i.e., `lmer` for the one-stage approach and `rma.mv` for the two-stage approach. Speed was measured by taking a time stamp just before and after the three-level model function call in `R` and by taking the difference between both. Both the `lmer` and `rma.mv` function calls included the option to conduct the *t*-test. The *z*-test was calculated afterwards based on the estimates and standard errors obtained from the `lmer` and `rma.mv` objects. Thus although the *z*-test was technically not included in the speed calculation, its attribution to the speed results of the approaches would only be marginal due to the simple and fast calculation. Convergence behavior was slightly different for the one-stage and two-stage approach due to inherent difference between the handling of convergence issues by `lmer` and `rma.mv`. The `lme4` package documentation lists some issues regarding testing convergence due to the difficulty of evaluating the gradient and the Hessian (Bates et al., 2018). As a consequence, `lmer` throws convergence warnings rather than errors, and the package authors provide some suggestions on how to troubleshoot these warnings. Such steps are beyond the scope of this simulation study and we considered `lmer` model fits which resulted in warnings as non-convergent. We did this after we confirmed that although

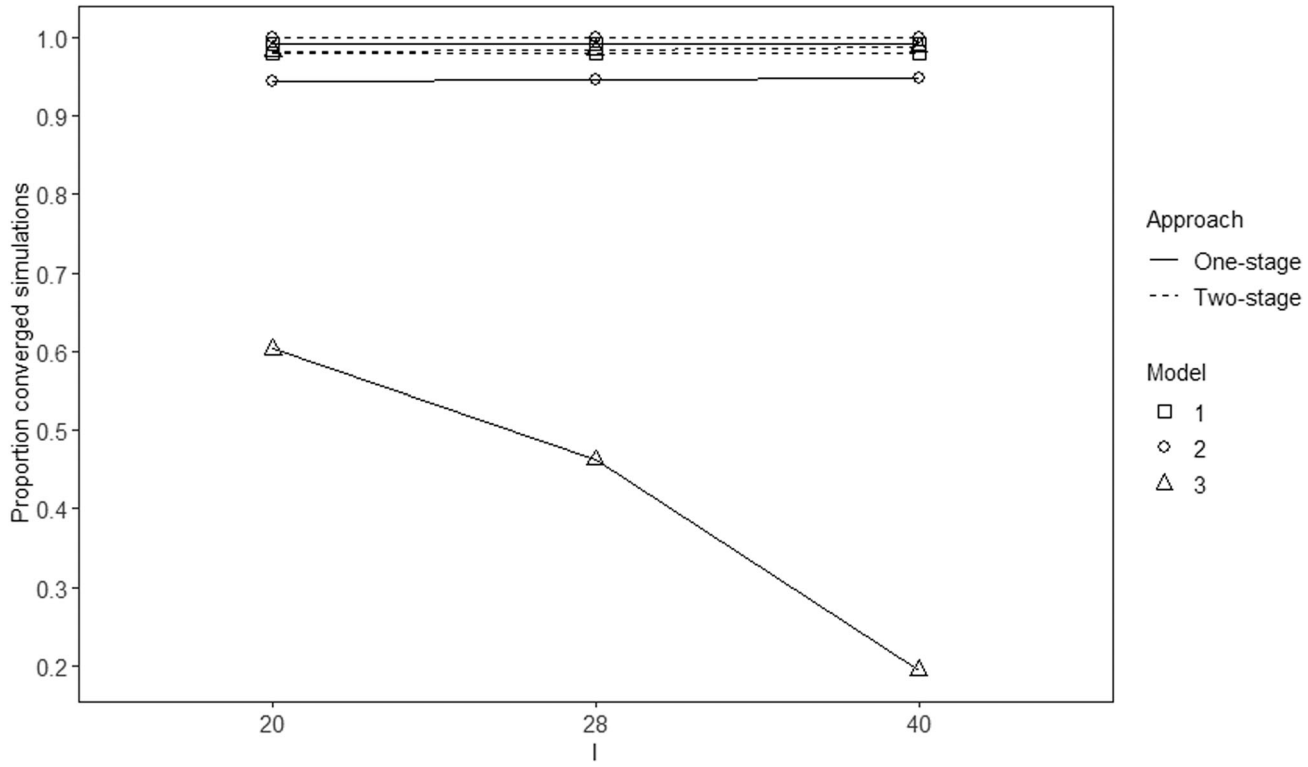

Figure 2. Convergence rates in function of number of measurements *I* for different approaches and models.

estimation results are still given when warnings occur, the point estimations proved to be less reliable (in terms of bias and MSE) than those from `lmer` model fits without warnings. The behavior of `rma.mv` is more straightforward: when convergence issues occur, an error is thrown and no results are shown. In such cases the model estimation was considered as not converged.

For all of the aggregated results on the statistical properties of the estimations and the speed and convergence of the model fits, we used ANOVA's to determine if and how the results vary across the simulation conditions. This was only used as a preliminary analysis in order to select those factors with a significant effect on a specific result, because likely the ANOVA assumptions of normality and homogeneity are not met. The factors we included in these ANOVAs were the approach used (one-stage or two-stage), the underlying model (Model 1, 2 or 3), the number of measurements, the number of cases, the number of studies, the value of the treatment effect(s), the value of the between- and within study variance and the value of the between- and within study correlation. We looked at the statistical significance of main and interaction effects and we calculated η^2-values to evaluate the size of the effects. Although the main effects for approach and model were not always large (i.e., $\eta^2 > .26$) according to the classical standards by Cohen (1988), we still included them in all results below because they are crucial to the research question which motivated this simulation study (i.e., is one or the other approach more suitable for simple versus more complex models?).

Results

Convergence

The simulation results showed that the convergence rates for the two-stage approach remain fairly stable when the model complexity increases: with 98% convergence rate for Model 1, 99.9% for Model 2, and 98.5% for Model 3, the convergence rate did not drop below 98%. The rates for the one-stage approach however decreased substantially for more complex models: for Model 1 99.2% of the simulations converged, whereas the convergence rate dropped to 94.6% for Model 2 and to 42% for Model 3. The ANOVA analysis indicated that the underlying model ($\eta^2 = .17$) and the approach ($\eta^2 = .10$) used have the biggest impact on the convergence, as well as their interaction ($\eta^2 = .17$). Furthermore, small interaction effects with the number of measurements *I* became clear ($\eta^2 = .02$ for the interaction of model and *I* and the interaction of approach, model and *I*). Figure 2 shows how the convergence rate for the one-stage approach is slightly lower (but not affected by *I*) for Model 2. The low

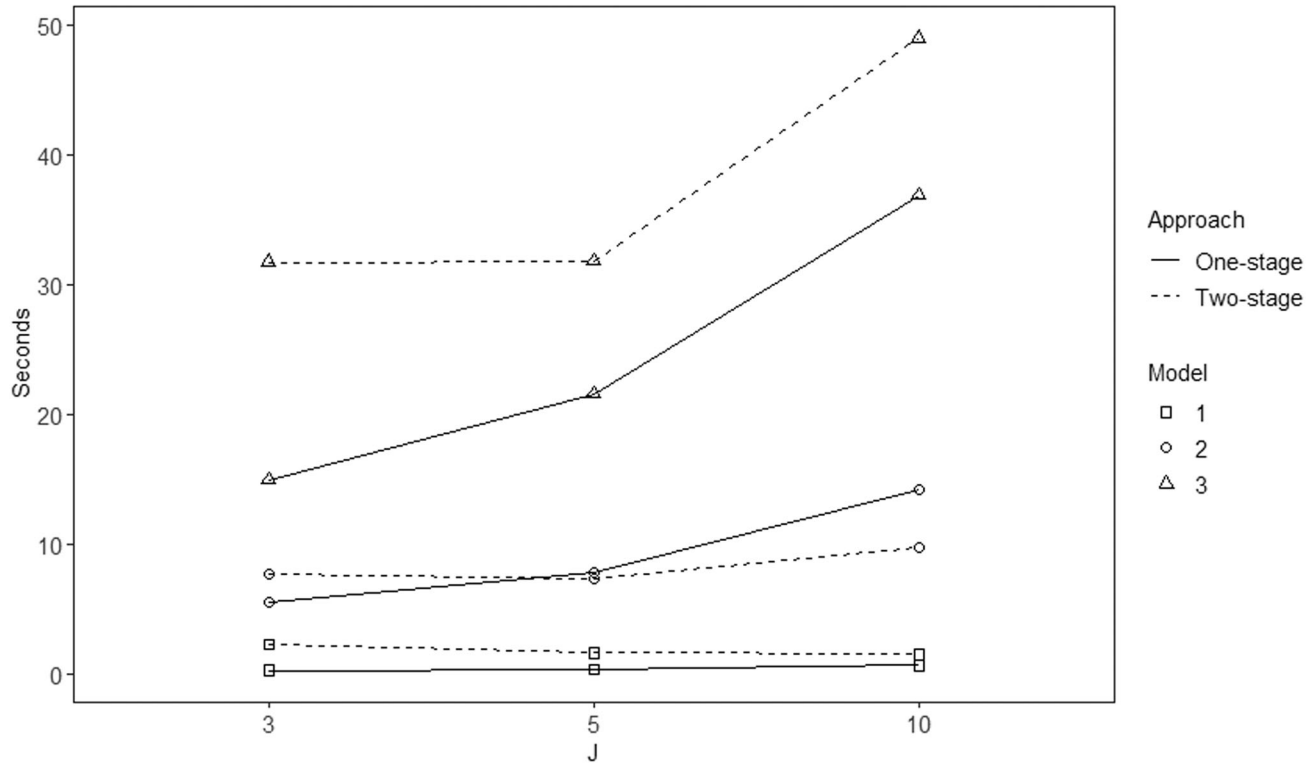

Figure 3. Speed of model fit in function of number of cases J for different approaches and models.

Table 2. Mean squared error for the overall treatment effect on the intercept.

		Model 1		Model 2		Model 3	
σ^2	K	One-stage	Two-stage	One-stage	Two-stage	One-stage	Two-stage
1	5	0.25	0.25	0.27	0.27	0.32	0.31
	7	0.18	0.18	0.19	0.19	0.22	0.22
	10	0.12	0.12	0.13	0.14	0.16	0.15
4	5	0.98	0.97	0.99	0.99	1.05	1.02
	7	0.69	0.69	0.71	0.71	0.79	0.74
	10	0.48	0.48	0.50	0.50	0.55	0.52

convergence rate for the one-stage approach when using Model 3 is clearly distinguishable and Figure 2 shows how it becomes worse when the number of measurements I increases.

Speed

Speed was mostly affected by the underlying model ($\eta^2 = .51$) and to a minor extent by the interaction of the approach and the number of cases J with the model ($\eta^2 = .03$ for J, $\eta^2 = .04$ for the interaction of model and J and $\eta^2 = .02$ for the interaction of approach and model). Figure 3 shows how again for models 1 and 2, both approaches reach convergence similarly fast. For Model 3, the model with the most parameters to estimate, the one-stage approach is faster, especially when the number of cases J is small. As J becomes larger, the one-stage and two-stage approach become slower and the difference in speed between both becomes smaller to 37 (one-stage approach) and 49 (two-stage approach) seconds on average for $J = 10$. In general both the `lmer()` and `rma.mv()` analyses converged in under a minute. The longest fit to complete took 6.08 minutes for Model 3 with the two-stage approach, in a condition with $\gamma = 0$, $\sigma^2 = 1$, $\rho = 0$, $I = 40$, $J = 10$ and $K = 10$.

Table 3. Bias for the overall treatment effect on the intercept.

	One-stage	Two-stage
Model 1	0.0012	0.0011
Model 2	0.0012	0.0006
Model 3	−0.0018	0.0004

Statistical properties of fixed effect estimations

To compare fixed effect estimates, we first verified the correlation between the one-stage and the two-stage approach estimates. The Pearson correlation coefficient between the deviations of the one-stage and two-stage estimates from the true nominal fixed effect was $r = .9994$. This confirms that the one-stage and two-stage fixed effect estimates are indeed consistent. We compared them in terms of MSE, bias, confidence interval coverage, Type I error rate and bias of the associated standard errors below. Because the conclusions are similar across different coefficients

Table 4. Empirical coverage probability of 95% confidence intervals for the overall treatment effect on the intercept.

	Model 1		Model 2		Model 3	
Wald-type CI	One-stage	Two-stage	One-stage	Two-stage	One-stage	Two-stage
Normal	.9039	.9067	.9152	.9072	.9256	.9133
Student's t[a]	.9509	.9167	.9579	.9124	.9644	.9168

[a]Using Satterthwaite (1941) degrees of freedom for the one-stage approach and degrees of freedom based on the Knapp and Hartung (2003) method for the two-stage approach.

(treatment effect on the intercept, the linear slope and the quadratic slope), we include only results for the treatment effect on the intercept, since this is the only fixed effect parameter included in all three models. Results for the other fixed effect parameters are available upon request from the last author.

MSE

The ANOVA analysis did not reveal any substantial difference in MSE between approaches ($\eta^2 < .001$) or for different models ($\eta^2 < .001$). Overall, the MSE is larger when $\sigma^2 = 4$ and smaller when $\sigma^2 = 1$ ($\eta^2 = .10$). Furthermore, the MSE increases when the number of studies K decreases ($\eta^2 = .02$). The impact of the number of units at the highest level (rather than the lower levels) of a multilevel model on the efficiency of the overall fixed effects estimations has been observed and described in previous simulation studies (Moeyaert et al., 2013a, 2013b). MSE values are shown in Table 2. Despite the small effects of model and approach as indicated by the ANOVA, we see from Table 2 that the MSE increases with model complexity and that for the most complex Model 3 the MSE is very slightly higher when applying the one-stage approach.

Bias

The fixed effect estimates were unbiased across all simulation conditions, independent from the underlying model or approach used (Table 3). The highest bias reported in Table 3 (for Model 3 with the one-stage approach) corresponds to a relative bias of ${}^{-0.0018}/_{2} = -.09\%$, which is negligible. This is consistent with results from previous simulation studies on multilevel modeling of SCED data (Ferron et al., 2009, 2010; Moeyaert et al., 2013a, 2013b; Owens & Ferron, 2012; Van den Noortgate & Onghena, 2003a, 2003b, 2008).

95% confidence interval coverage

The ANOVA revealed that there was no substantial difference in empirical coverage probability of the Wald z-type 95% confidence intervals across simulation conditions, models or approaches. Table 4 shows how for all three models and both approaches, only 90% to 92% of the obtained 95% confidence intervals contain the true nominal parameter value. The one-stage approach performs slightly better for models 2 and 3. In an attempt to improve the confidence interval coverage, we recalculated the confidence intervals by using a t-distribution with Satterthwaite degrees of freedom for the one-stage approach and the Knapp and Hartung (2003) like degrees of freedom as provided by `rma.mv` for the two-stage approach. This drastically improved the coverage probability when using the one-stage approach, but only slightly improved the coverage for the two-stage approach so that for models 1 and 2 the 91% coverage probability obtained for Model 3 was now also reached. The improved coverage probability for the one-stage approach is in line with the results from Ferron et al. (2009).

Type I error rate

Just as for the confidence intervals, we evaluated Type I error rates based on two types of hypothesis tests: a z-test to make a fair comparison between the one-stage and the two-stage approach, and a t-test where both approaches again using the two different types of degrees of freedom. Based on the ANOVA's, we also include the effect of the number of studies K ($\eta^2 = .001$) in Table 5 next to the underlying model ($\eta^2 < .001$) and approach used ($\eta^2 < .0001$). When calculating p-values based on the Z-statistic, both the one-stage and two-stage approach obtain equally bad Type I error rates (as expected), especially for small K and the less complex Model 1. The Type I error rates again improve substantially for the one-stage approach when using the t-test with Satterthwaite degrees of freedom. Using a t-distribution (with Knapp-Hartung degrees of freedom) did not substantially improve the results of the two-stage approach: the Type I error rates based on the t-test remain too high compared to the nominal $\alpha = .05$ level.

Relative bias of standard errors

The relative bias in the standard errors is calculated by comparing the mean standard error within a condition with the sample standard deviation of the

Table 5. Type I error rates for the overall treatment effect on the intercept based on a nominal $\alpha = .05$ significance level.

		Model 1		Model 2		Model 3	
Hypothesis test	*K*	One-stage	Two-stage	One-stage	Two-stage	One-stage	Two-stage
Normal	5	.11	.11	.10	.11	.08	.10
	7	.09	.09	.08	.09	.08	.09
	10	.08	.08	.08	.08	.07	.08
Student's *t* [a]	5	.05	.09	.04	.10	.03	.09
	7	.05	.08	.04	.09	.04	.09
	10	.05	.07	.05	.08	.04	.08

[a]Using Satterthwaite (1941) degrees of freedom for the one-stage approach and degrees of freedom based on the Knapp and Hartung (2003) method for the two-stage approach.

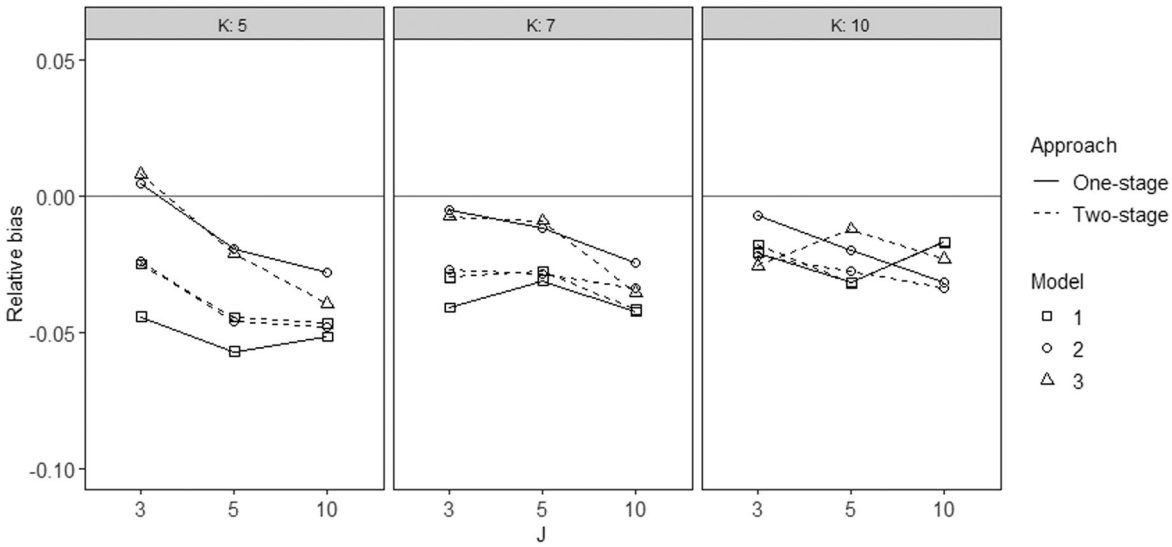

Figure 4. Relative bias in the standard errors of the fixed effect parameter estimate $\hat{\gamma}$ in function of the number of cases *J* and the number of studies *K*.

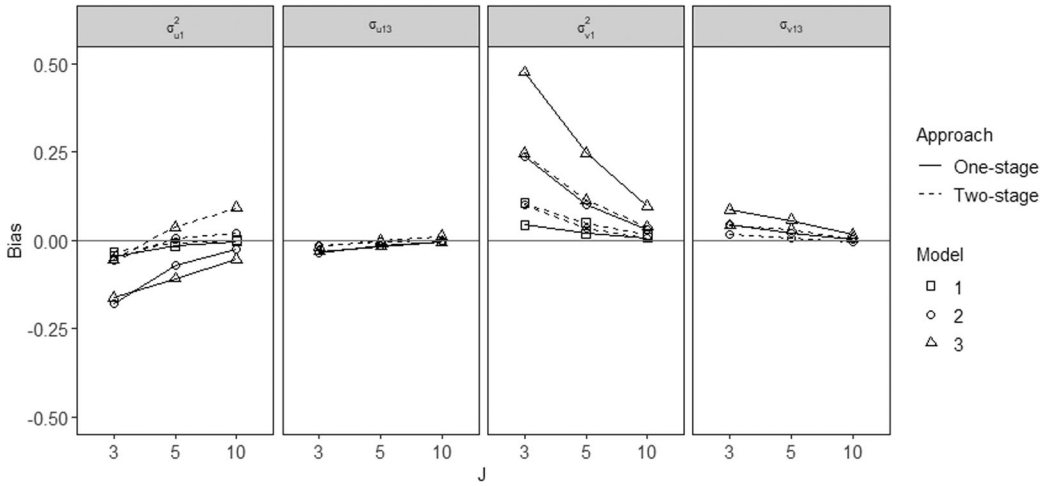

Figure 5. Bias in the variance component estimates in function of the number of cases *J*.

estimates within that condition. For conditions in which many or all of the 1000 simulations converged, this yields reliable results. However, in some conditions very few or no simulations converged. We therefore decided to only include conditions in which at least half of the simulations converged, in order to obtain a mean standard error and a sample standard deviation of the estimates based on at least $n = 500$ observations. As such, 216 of $216 \times 3 \times 2 = 1296$ or about 16% of the conditions were left out, including

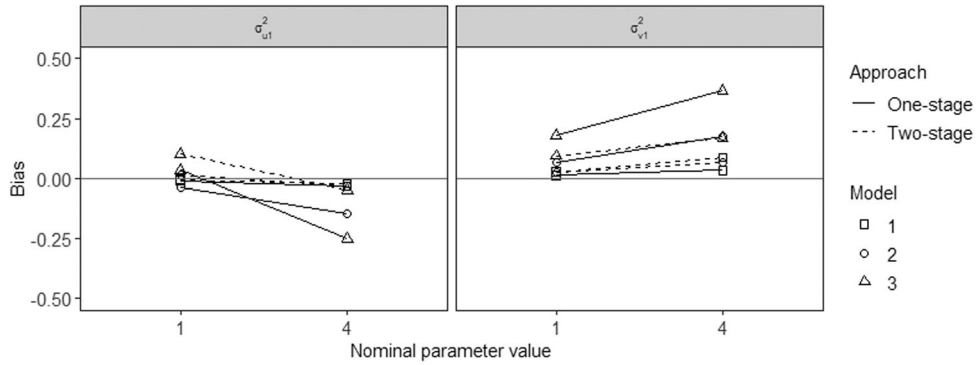

Figure 6. Bias in the variance estimates in function of the true value σ^2 for the within- and between-case variance.

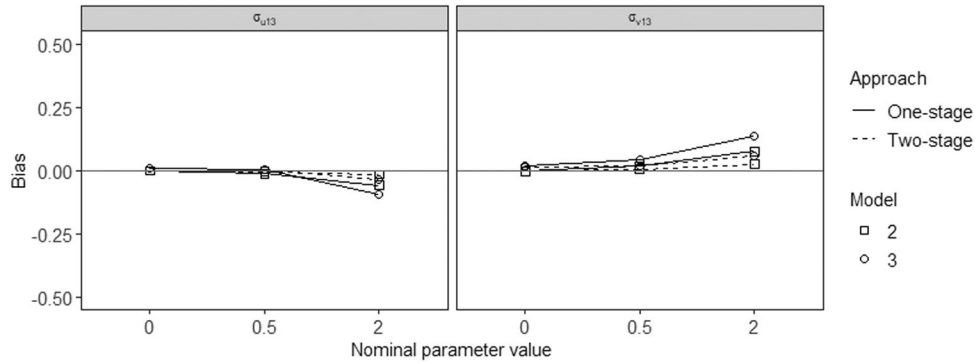

Figure 7. Bias in the covariance estimates in function of the true value $\sigma^2\rho$ for the within- and between-case covariance.

all conditions using the one-stage approach with Model 3. Based on the ANOVA the effect of the number of studies K ($\eta^2 = .017$) and the number of cases J ($\eta^2 = .040$) was incorporated in Figure 4. The underlying model had the largest effect on the relative bias of the standard errors ($\eta^2 = .052$), while the effect of the approach used was smaller ($\eta^2 = .006$). The small effects of these factors on the relative bias of the standard errors is shown in Figure 4: the relative bias is small to negligible almost everywhere. Only for $K = 5$, $J \geq 5$ and Model 1 with the one-stage approach, the relative bias exceeds 5%.

Statistical properties of (co)variance estimations

To compare (co)variance estimations, we again verified the correlation between the deviations of the one-stage and the two-stage approach estimates from the true nominal (co)variance values. The Pearson correlation coefficients equaled $r = .9895$ for the variances and $r = .9941$ for the covariances. This confirms that the one-stage and two-stage (co)variance estimates are indeed consistent and we compare them in terms of bias and MSE below. Note that we can only compare (co)variance parameters estimated in both approaches, i.e., those related to the treatment coefficients (see Table 1). Across models this means there are $2 + 6 + 12 = 20$ different (co)variances to compare. In the results below we only report the results for the within- and between-case variance in treatment effect on the intercept ($\hat{\sigma}^2_{u1}$ and $\hat{\sigma}^2_{v1}$, respectively), which are estimated for all three models, and the within and between-case covariances between the treatment effect on the intercept and the treatment effect on the time trend ($\hat{\sigma}_{u13}$ and $\hat{\sigma}_{v13}$, respectively), which are estimated in Model 2 and Model 3. Results for the other (co)variance parameters are available upon request from the last author.

Bias

From the ANOVA analyses carried out per (co)variance parameter, the biggest effects on the bias were attributed to the nominal value of the parameter, the number of cases J, the number of studies K and the underlying model used. The effect of the approach used was small. Figures 5–7 show the bias in function of J for the four parameters $\sigma^2_{u1}, \sigma^2_{v1}, \sigma_{u13}$ and σ_{v13} and in function of the nominal value of the parameter, i.e., σ^2 for σ^2_{u1} and σ^2_{v1} and $\sigma^2\rho$ for σ_{u13} and σ_{v13}. From Figure 5 it is clear that the bias decreases as the number of studies J increases. An almost identical pattern was observed when plotting the bias as a function of the number of studies K. The bias appears to be

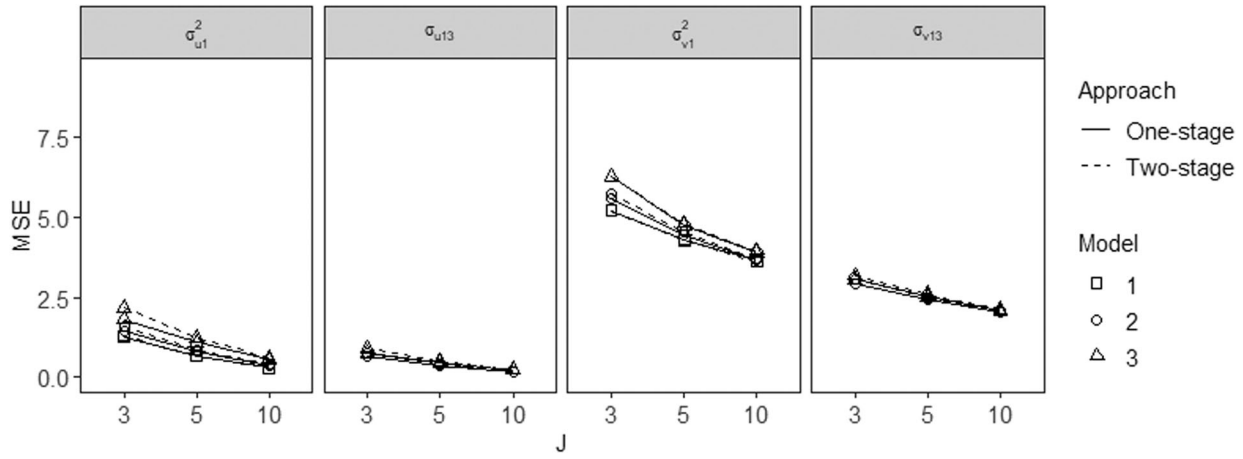

Figure 8. MSE of the variance component estimates in function of the number of cases *J*.

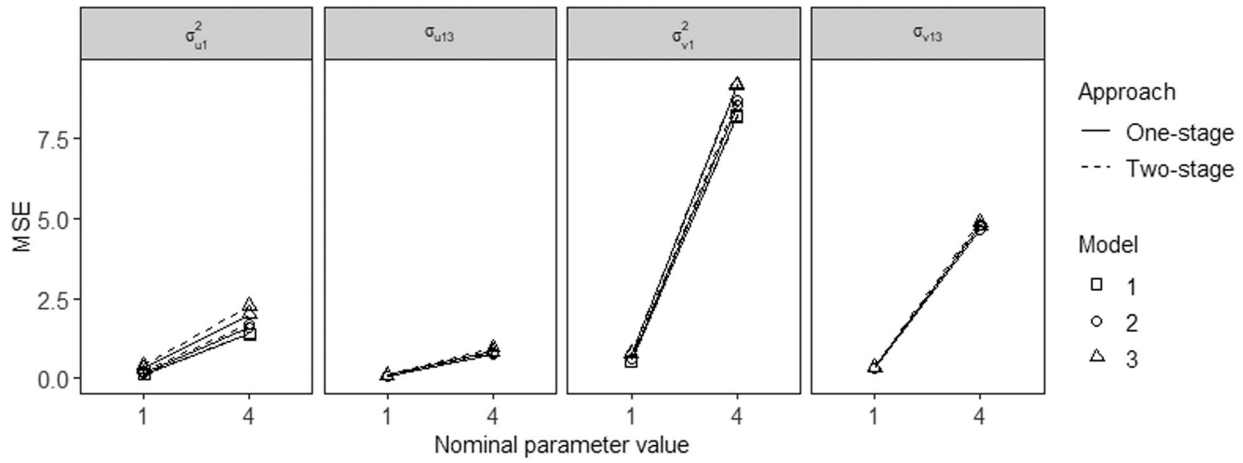

Figure 9. MSE of the variance estimates in function of σ^2, the within- and between-case variance.

larger for the variance parameters compared to the covariance parameters, but this is due to the scale of the *y*-axis. The underlying nominal values of the variances are different from the ones of the covariances and we did not compute the relative bias because this was impossible for conditions in which the covariance equaled 0. From Figures 6 and 7 we see that the bias is larger when the variance or covariance is larger. Figures 5–7 confirm the small effect of the approach. It is however clear that for the variances σ^2_{u1} and σ^2_{v1}, the one-stage approach leads to larger biases when the underlying model is more complex.

MSE

The ANOVA analyses per variance component showed that their MSE is affected by the nominal value of the variance component, the number of cases *J* and the number of studies *K*. Again we only report the results in terms of the number of cases *J*, since the results for the number of studies *K* are very similar. The MSE decreases as the number of cases *J* increases (Figure 8) and is noticeably larger for variance components at the study level (Figures 8 and 9). The MSE's are very similar for both approaches and only slightly larger when the underlying model becomes more complex. For the parameter σ^2_{u1}, the two-stage approach seems to lead to somewhat larger MSE's with more complex underlying models. From Figure 9 finally, we see that the MSE's increase as the nominal value of the variance σ^2 increases.

Discussion

In this simulation study, we wanted to verify how using effect sizes for SCED data in a three-level meta-analytic model impacts the speed, convergence rate and statistical properties of the estimates compared to using the raw data. For several single-case designs, we showed how to calculate effect sizes per case from the individual raw data and how to set up an appropriate three-level model. The advantage of using the effect size approach is that this corresponding three-level

model has fewer parameters to estimate than a three-level model applied to the original raw data. A disadvantage on the other hand was that effect sizes compress the rich information present in the raw data. The question was whether the effect size approach leads to faster and better convergence and how it affects the statistical properties of the estimations.

Based on the results we reported, neither of the approaches outperforms the other on all aspects. The two-stage approach does lead to better convergence rates if the underlying single-case design is more complex, but it was not faster. In terms of MSE and bias of the fixed effect and variance component estimations, the two-stage approach yields similar results as the one-stage approach and the standard errors of its fixed effects are similarly unbiased, but its empirical confidence interval coverage and Type I error rate are consistently worse. The fact that we did not see large differences in efficiency (i.e., bias and MSE) between the one-stage and two-stage estimators is not surprising. Mathew and Nordström (2010) have previously shown that for a fairly general two-level setup (with participants nested within trials), the two-stage approach is asymptotically as efficient as the one-stage approach. Under the assumption that the first-level covariance matrix has been accurately estimated and that the fraction of observations corresponding to any given treatment remains the same across trials, the one-stage and two-stage IPD estimators coincide. Both assumptions hold in our simulation study: the first-level variance estimates are retrieved from the OLS regression on the within-case data, and the fraction of observations corresponding to the baseline and the treatment phase was equal to 0.5 for all simulated datasets.

Much of the simulation results in this study of course depend heavily on the implementation of both approaches in R, using two different packages and several different functions. Although we tried to streamline the code as much as possible, we are aware of the fact that next to comparing the one-stage and the two-stage approach, we are also comparing the `lmer` and the `rma.mv` function in terms of code efficiency and available options. Although the open-source `R` software has the advantage of being more widely available, it would also be interesting to repeat the simulations with e.g., PROC MIXED (Littell et al., 2007) using SAS software. The MIXED procedure allows for fixing the level-one variance as well as obtaining *p*-values and confidence intervals for the fixed effects with different methods of computing the denominator degrees of freedom (*SAS/STAT 15.1 User's Guide, 2018*).

The main difference between `lmer` and `rma.mv` is in their available options for inferential statistics. We conducted a Wald-type *z*-test to make a fair comparison between the two approaches, but since the null distribution can only be asymptotically approximated by a normal distribution, this is usually not recommended. Switching to a *t*-distribution to reduce Type I error inflation requires computation of the degrees of freedom. Unlike its univariate counterpart `rma`, the `rma.mv` function only offers inference based on the normal distribution or on a *t*-distribution with a fairly simple degrees of freedom calculation. As such, the improvement in confidence interval coverage and Type I error rate for the two-stage approach implemented with `rma.mv` (Tables 4 and 5) is minimal. Small sample adjustments for the degrees of freedom, like the Satterthwaite method applied in the one-stage approach with `lmer` (Satterthwaite, 1941), the Kenward-Roger method (Kenward & Roger, 1997), or the Knapp and Hartung between-study variance estimator (Hartung & Knapp, 2001), yield a more substantial improvement. Sánchez-Meca and Marín-Martínez (2008) have illustrated this for the confidence interval coverage in a simulation study in which different confidence interval construction methods and between-study variance estimators are compared, including the Wald-type *z*-test method and the Hartung and Knapp (2001) method. The latter method is currently implemented in the univariate `rma` function in `metafor` and could be implemented for results obtained with the `rma.mv` function by manually adjusting the variance-covariance matrix of the fixed effects.

The speed and convergence of the multilevel model fits are two other aspects which might depend on the implementation of `lmer` and `rma.mv`. Here we attempted to streamline the results by making sure both functions used the same estimation method (REML), optimizer (BOBYQA) and maximum number of function evaluations (10,000). The results of our simulations show that the difference between both model fits is small and that with averages below one minute and a maximum duration of 6 minutes across all simulations, estimation speed for a single model fit is negligible in practice. Non-convergence is also dealt with differently in practice, i.e., when only a single dataset is being analyzed. It is possible to make manual adjustments to the optimization routines with both `lme4` and `metafor` in order to obtain estimations, but in a simulation study with thousands of datasets, this is not a practical option.

Simulation studies on IPD meta-analysis are still relatively rare. The SCED simulation results obtained in this study can be compared with a recent simulation study by Morris et al. (2018), who compared the one-stage and two-stage approach for data obtained from clinical trials. Using a simple linear model with fixed intercept and random treatment effect, Morris et al. (2018) found little to no difference in empirical variance and in coverage of 95% confidence intervals. The first result is in line with our results in Figure 4, where we compare the mean of the standard errors of the treatment effect with the empirical standard deviation of the treatment effect estimates. The 95% confidence interval coverage in Table 4 show a larger difference between the one-stage and the two-stage approach. As discussed earlier this might be mostly due to the implementation of our simulations in `R` with the `metafor` and the `lme4` package, while Morris et al. (2018) have conducted their simulations with SAS and Stata (*Stata multilevel mixed-effects reference manual, 2013*). Legha et al. (2018) have conducted a more extensive simulation study on the IPD one-stage approach, also for a simple linear model similar to Model 1 in this study. One of their key findings is that using the Satterthwaite or Kenward-Roger approach leads to better 95% confidence interval coverage compared to a standard normal approach, although that on occasion the Satterthwaite and Kenward-Roger confidence intervals show some over-coverage. Again this agrees with what we have found in Table 4. The two-stage approach has also been studied in a much earlier simulation study by Stukel and Demidenko (1997), who suggest that the two-stage approach may be more robust to certain forms of model misspecification, in particular when the focus is on a subset of the model coefficients. This is indeed the case for the multilevel models in the current simulation study, where in the second stage (Equation 5) only the treatment effect γ_{100} and its associated variance components are estimated, but not the baseline level γ_{000}.

The models analyzed in this simulation study were still of relatively modest complexity. We did preliminary simulations with more models of increasing complexity, including models with ABAB designs with and without linear or quadratic time trends and with combinations thereof. These took substantially more time to fit using the one-stage as well as the two-stage approach and were finally not included due to the fact that running a full simulation (including 1000 simulations across 216 conditions) for these models was not feasible within a reasonable time frame. Note that although we ran simulations in parallel, it would take 221 days in total to run all simulations in this study sequentially.

Another complexity which might be considered in future simulation studies on SCED raw data and effect sizes, is modeling of discrete or proportional outcomes, which are very common in single-case research (Shadish & Sullivan, 2011). This can be done by means of generalized linear mixed modeling (Declercq et al., 2019). Applying generalized linear mixed models for non-normal data involves integration over the random effects and thus requires more complex estimation techniques (e.g., penalized quasi-likelihood, Gauss-Hermite quadratures or Markov chain Monte Carlo algorithms). This might have a substantial impact on convergence rates and estimation speeds. The simulation study by Declercq et al. (2019) showed that accounting for non-continuous outcomes with generalized linear models makes modeling single-case data considerably more complex, but that using simple linear mixed models as the ones presented in this manuscript would not necessarily lead to inaccurate fixed effect estimates or flawed inferences when applied to count outcomes. Because in this study, we wanted to isolate the complexity of having an increased number of regression coefficients (and how that affects the one-stage and two-stage model), we have chosen to only simulate continuous outcomes.

For all multilevel models presented in this simulation study, random effects were included for every regression coefficient. In practice, researchers conducting SCED meta-analyses will carefully consider for which regression coefficients random effects need to be included, based on the characteristics of the data, the amount of data at hand, and their specific research interests. In a meta-analysis of SCED studies, the conducting researcher is likely to be explicitly interested in how the effect varies across cases and across studies, and as such it is equally likely that the most appropriate model includes multiple random effects.

The results reported in this simulations study are, as for any simulation study, limited to the simulation conditions used: the choices of nominal values for the model parameters and of the number of measurements, cases and studies, as well as the models used to generate and analyze the SCED data (as explained in the previous paragraph). The results of the two-stage approach also depend on the particular choice of SCED effect sizes (i.e., regression coefficients). Instead of regression coefficients, the univariate or multivariate two-stage approach could be applied to other parametric or non-parametric effect sizes (Manolov & Moeyaert, 2017), like nonoverlap indices (e.g., percentage of non-overlapping data or PND),

descriptive indices quantifying changes in level and slope (e.g., mean phase difference or MPD), standardized mean difference indices (e.g., the Hedges et al. (2013) *d* statistic) and other indices based on regression analyses (e.g., the Pustejovsky et al. (2014) *d* statistic).

Conclusion

We simulated raw SCED data with three levels (measurements nested within cases nested within studies) from three designs: an intercept-only model, a model with a linear time trend and a model with a quadratic time trend. On these simulated data, we applied a one-stage IPD approach by fitting a three-level meta-analytic model on the raw data. Then, we calculated effect sizes by fitting individual OLS linear regression models per case and retrieving the regression coefficient(s) related to the treatment. These were subsequently used as the dependent variables in a uni- or multivariate three-level model. As such we applied a two-stage IPD meta-analytic approach. Although reducing the raw data to effect sizes leads to a loss of information, the resulting three-level model has fewer parameters to estimate. For the three models of increasing complexity, we investigated how this impacted the convergence rate and speed of the model estimations as well as the statistical properties of the estimates. Using two different packages and functions in `R`, the results showed that for more complex models, the one-stage approach indeed obtained better convergence rates but that model estimations did not necessarily converge at a faster speed. The precision and the bias of the point estimates was very similar for both approaches and for all models. Inference results were consistently worse for the two-stage approach, although this might be due to the particular implementation and methods used in `R` (i.e., the `rma.mv` function from the `metafor` package). When confronted with convergence issues when estimating a multilevel model from the raw data, one of the options for applied SCED researchers who are conducting a meta-analysis based on raw SCED data could be to turn to the two-stage approach instead, especially if they cannot simplify their model. With the two-stage approach, practitioners might experience less convergence issues with larger, more complex multilevel models (especially if effect sizes are based on larger studies) and they should obtain reliable and valid point estimates. However, they should interpret the corresponding inference results obtained from the multilevel analysis with caution.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant R305D110024 from the Institute of Educational Sciences.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank the two anonymous reviewers for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions or IES is not intended and should not be inferred.

References

Bates, D. (2006). *lmer, p-values and all that.* Retrieved June, 2020, from https://stat.ethz.ch/pipermail/r-help/2006-May/094765.html

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2018). *convergence: Assessing convergence for fitted models.* Retrieved June, 2020, from https://rdrr.io/cran/lme4/man/convergence.html

Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A., & Feldman, H. I. (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Statistics in Medicine*, *21*(3), 371–387. https://doi.org/10.1002/sim.1023

Bolker, B. (2018). *GLMM FAQ.* Retrieved June, 2020, from http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html

Burke, D. L., Ensor, J., & Riley, R. D. (2017). Meta-analysis using individual participant data: One-stage and two-stage approaches, and why they may differ. *Statistics in Medicine*, *36*(5), 855–875. https://doi.org/10.1002/sim.7141

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data

versus aggregated data. *Psychological Methods*, *14*(2), 165–176. https://doi.org/10.1037/a0015565

Debray, T. P. A., Moons, K. G. M., van Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H. H., & Reitsma, J. B. (2015). Get real in individual participant data (IPD) meta-analysis: A review of the methodology. *Research Synthesis Methods*, *6*(4), 293–309. https://doi.org/10.1002/jrsm.1160

Declercq, L., Jamshidi, L., Fernández-Castilla, B., Beretvas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (2019). Analysis of single-case experimental count data using the linear mixed effects model: A simulation study. *Behavior Research Methods*, *51*(6), 2477–2497. https://doi.org/10.3758/s13428-018-1091-y

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, *41*(2), 372–384. https://doi.org/10.3758/BRM.41.2.372

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*(4), 930–943. https://doi.org/10.3758/BRM.42.4.930

Fiala, C. L., & Sheridan, S. M. (2003). Parent involvement and reading: Using curriculum-based measurement to assess the effects of paired reading. *Psychology in the Schools*, *40*(6), 613–626. https://doi.org/10.1002/pits.10128

Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, *20*(12), 1771–1782. https://doi.org/10.1002/sim.791

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, *4*(4), 324–341. https://doi.org/10.1002/jrsm.1086

Jamshidi, L., Heyvaert, M., Declercq, L., Fernández-Castilla, B., Ferron, J. M., Moeyaert, M., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). A systematic review of single-case experimental design meta-analyses: Characteristics of study designs, data, and analyses. *Evidence-based Communication Assessment and Intervention*.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710. https://doi.org/10.1002/sim.1482

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Legha, A., Riley, R. D., Ensor, J., Snell, K. I., Morris, T. P., & Burke, D. L. (2018). Individual participant data meta-analysis of continuous outcomes: A comparison of approaches for specifying and estimating one-stage models. *Statistics in Medicine*, *37*(29), 4404–4420. https://doi.org/10.1002/sim.7930

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2007). *SAS for mixed models* (2nd ed.). SAS institute.

Manolov, R., & Moeyaert, M. (2017). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy*, *48*(1), 97–114. https://doi.org/10.1016/j.beth.2016.04.008

Mathew, T., & Nordström, K. (2010). Comparison of one-step and two-step meta-analysis models using individual patient data. *Biometrical Journal. Biometrische Zeitschrift*, *52*(2), 271–287. https://doi.org/10.1002/bimj.200900143

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2013a). Three-level analysis of single-case experimental data: Empirical validation. *The Journal of Experimental Education*, *82*(1), 1–21. http://www.tandfonline.com/doi/abs/10.1080/00220973.2012.745470 https://doi.org/10.1080/00220973.2012.745470

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2013b). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research*, *48*(5), 719–748. https://doi.org/10.1080/00273171.2013.816621

Morris, T. P., Fisher, D. J., Kenward, M. G., & Carpenter, J. R. (2018). Meta-analysis of Gaussian individual patient data: Two-stage or not two-stage? *Statistics in Medicine*, *37*(9), 1419–1438. https://doi.org/10.1002/sim.7589

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*(3), 795–805. https://doi.org/10.3758/s13428-011-0180-y

Powell, M. J. D. (2009). *The BOBYQA algorithm for bound constrained optimization without derivatives* (Tech. Rep.). Department of Applied Mathematics and Theoretical Physics, Cambridge University.

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*(5), 368–393. https://doi.org/10.3102/1076998614547577

Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods*, *13*(1), 31–48. https://doi.org/10.1037/1082-989X.13.1.31

SAS/STAT 15.1 User's Guide. (2018). SAS Institute Inc. Cary, NC.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*(5), 309–316. https://doi.org/10.1007/BF02288586

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*(4), 971–980. https://doi.org/10.3758/s13428-011-0111-y

Stata multilevel mixed-effects reference manual. (2013). StataCorp LP. College Station, TX.

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions*, *25*(1), 76–97. https://doi.org/10.1177/0163278702025001006

Stukel, T. A., & Demidenko, E. (1997). Two-stage method of estimation for general linear growth curve models. *Biometrics*, *53*(2), 720–728.

Tudur Smith, C., Marcucci, M., Nolan, S. J., Iorio, A., Sudell, M., & Riley, R. (2016). Individual participant data

meta-analyses compared with meta-analyses based on aggregate data. *Cochrane Database of Systematic Reviews*, *9*.doi:10.1002/14651858.MR000007.pub3.

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, *18*(3), 325–346. https://doi.org/10.1521/scpq.18.3.325.22577

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 1–10. https://doi.org/10.3758/bf03195492

Van den Noortgate, W., & Onghena, P. (2003c). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*, *63*(5), 765–790. https://doi.org/10.1177/0013164403251027

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *Behavior Analyst Today*, *8*(2), 52–57.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 142–151. https://doi.org/10.1080/17489530802505362

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W. (2016). *A comparison of the rma() and the lm(), lme(), and lmer() functions*. Retrieved June, 2020, from http://www.metafor-project.org/doku.php/tips:rma_vs_lm_lme_lmer

Viechtbauer, W. (2018). *Bootstrapping with meta-analytic models*. Retrieved June, 2020, from http://www.metafor-project.org/doku.php/tips:bootstrapping_with_ma