Behavior Research Methods (2019) 51:2477–2497
https://doi.org/10.3758/s13428-018-1091-y

Analysis of single-case experimental count data using the linear mixed effects model: A simulation study

Lies Declercq[1,2] · Laleh Jamshidi[1,2] · Belén Fernández-Castilla[1,2] · S. Natasha Beretvas[3] · Mariola Moeyaert[4] · John M. Ferron[5] · Wim Van den Noortgate[1,2]

Published online: 13 August 2018

Abstract

When (meta-)analyzing single-case experimental design (SCED) studies by means of hierarchical or multilevel modeling, applied researchers almost exclusively rely on the linear mixed model (LMM). This type of model assumes that the residuals are normally distributed. However, very often SCED studies consider outcomes of a discrete rather than a continuous nature, like counts, percentages or rates. In those cases the normality assumption does not hold. The LMM can be extended into a generalized linear mixed model (GLMM), which can account for the discrete nature of SCED count data. In this simulation study, we look at the effects of misspecifying an LMM for SCED count data simulated according to a GLMM. We compare the performance of a misspecified LMM and of a GLMM in terms of goodness of fit, fixed effect parameter recovery, type I error rate, and power. Because the LMM and the GLMM do not estimate identical fixed effects, we provide a transformation to compare the fixed effect parameter recovery. The results show that, compared to the GLMM, the LMM has worse performance in terms of goodness of fit and power. Performance in terms of fixed effect parameter recovery is equally good for both models, and in terms of type I error rate the LMM performs better than the GLMM. Finally, we provide some guidelines for applied researchers about aspects to consider when using an LMM for analyzing SCED count data.

Keywords Generalized linear mixed model · Linear mixed model · Single-case experimental design · Monte Carlo simulation

Introduction

A single-case experimental design (SCED) is an experimental design where one subject, participant, or case is observed repeatedly over time, resulting in a time series. During this time series, one or more dependent variables are measured under different levels in order to assess the effect of the particular treatment or intervention (Onghena & Edgington, 2005). Often the time series includes at least one baseline phase and one treatment phase. Studies using an SCED design frequently report results of a small number of multiple cases. When generalizing the results of several SCED studies in a meta-analysis, the data of interest is then of a hierarchical nature: measurements are nested within cases, which in turn are nested within studies. This hierarchical nesting of the data can be taken into account elegantly by using hierarchical or multilevel modeling for statistical analysis (Van den Noortgate & Onghena 2003a, b, 2008).

In the basic multilevel model for meta-analysis of SCED data as proposed in previous research (Raudenbush & Bryk 2002; Moeyaert et al., 2014; Shadish et al., 2008, 2013; Van den Noortgate & Onghena 2007), the observed scores for each case are assumed to be normally distributed around their expected value. However, Shadish and Sullivan (2011) have reported that the outcome variables measured in SCED

This research is funded by the Institute of Education Sciences, U.S. Department of Education, grant number R305D150007. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

✉ Lies Declercq
lies.declercq@kuleuven.be

[1] Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

[2] ITEC imec, Kapeldreef 75, Leuven, Belgium

[3] Department of Educational Psychology, University of Texas, Austin, TX 78712, USA

[4] Department of Educational Psychology and Methodology, State University of New York, New York City, NY, USA

[5] Department of Educational Measurement and Research, University of South Florida, Tampa, FL 33620, USA

studies are very often of a discrete rather than continuous nature, and for these discrete outcomes the assumption of conditional normality does not hold. To account for both the hierarchical and the count nature of SCED data, two frameworks can be combined: linear mixed modeling (LMM) (Hox, 2010; Gelman & Hill, 2009; Snijders & Bosker, 2012) and generalized linear modeling (GLM) (Gill, 2001; McCullagh & Nelder, 1999). Both frameworks have proven to provide very flexible tools. From their most basic forms, they expand into more specialized models in a clear and simple manner. Combining both frameworks results in a generalized linear mixed model (GLMM) (Hox, 2010; Gelman & Hill, 2009; Snijders & Bosker, 2012; Jiang, 2007), which is specified by (1) a distribution for the random effects, (2) a linear combination of predictor variables, (3) a function linking this linear predictor to the expected value of the response variable conditional on the random effects, and (4) a distribution for the response variable around this expected value. GLMMs can be very well customized to the particular type of data at hand, i.e., count data in SCED meta-analyses (Shadish et al., 2013).

One downside of the GLMM framework is that it is relatively complex to understand. Customizing a generalized linear mixed model requires a more general mathematical understanding of both the GLM and the LMM framework. Even though efficient estimation methods are available in many popular software packages (Zhou et al., 1999; Bates et al., 2015; Molenberghs et al., 2002) and even though these models have proven their robustness and their power (Abad et al., 2010; Capanu et al., 2013; Yau & Kuk, 2002), they might be somewhat intimidating for social scientists to apply. Another difficult aspect of the GLMM framework is that the more sophisticated the model, the more information is needed to make sure the GLMM estimation converges (Li & Redden, 2015; Abad et al., 2010). However, in SCED contexts typically a relatively small number of data points is available (Shadish & Sullivan, 2011) and this might result in less reliable GLMM estimates (Nelson & Leroux, 2008).

For an assessment of the current use of GLMMs in SCED contexts, we have access to data collected for a recent review conducted by the same team of authors of this simulation study (Jamshidi et al., 2017). This systematic review includes 178 systematic reviews and meta-analyses of SCED studies from the last three decades and includes a description of their study characteristics. Of the included studies, only 22 (12%) used hierarchical or mixed modeling and 19 of those were published after 2010. Only about half of these studies reported the type of measurement scale of the dependent variable, but those that did reported almost exclusively rates, percentages, or counts. Yet all of these 22 studies used an LMM rather than a GLMM. Together with the aforementioned issues of the complexity of the GLMM, this observation encourages us to look deeper into the consequences of misspecifying SCED count data with an LMM (which assumes normally distributed outcomes).

To this end, a simulation study is conducted in which count data with a hierarchical structure are generated according to a two-level GLMM, assuming a Poisson distribution of scores within the phases. The simulated datasets are analyzed by fitting the GLMM used for data generation, as well as by fitting a two-level LMM that assumes normality of the scores within phases. The main aim of this study is to investigate whether the GLMM, as the theoretically correctly specified model, outperforms the LMM across all conditions, and, if not, in which conditions the LMM performs well enough (or better).

As to the conditions in which the LMM leads to acceptable performance, we have two hypotheses. First, if the expected count responses in the baseline and/or treatment phase are relatively high, the LMM might perform relatively better than when the expected counts are small due to better normal approximations of Poisson distributions with larger expected values (Stroup, 2013). Second, if the sample size is small, the LMM might perform relatively better than the GLMM due to the GLMM being a too complex model to estimate when information is sparse (Hembry et al., 2015).

Various simulation conditions are taken into account. These conditions differ in the number of cases, the number of measurements within cases, the average baseline response, the average effect size and the true variance component values. To analyze the performance of the model fits, we look at common goodness of fit criteria, fixed effect parameter recovery, the Type I error rate and the power. The goal is to provide applied researchers with recommendations on the required criteria (e.g., the required sample size or the required average count in the baseline and/or treatment phase) for reliable analysis of count data with simpler LMMs.

Methodology

For simplicity, the simulation in this study will only take into account two levels (measurements nested within cases). The model used to simulate the SCED count data is a GLMM with an underlying Poisson distribution and a log link function:

$$
\begin{aligned}
&Y_{ij} \sim \text{Poisson}\left(\lambda_{ij}\right)\\
&\log\left(\lambda_{ij}\right) = \beta_{0j} + \beta_{1j} D_{ij}\\
&\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} + u_{1j} \end{cases}\\
&\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim \text{MVN}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0} & \sigma_{u01} \\ \sigma_{u01} & \sigma^2_{u1} \end{pmatrix} \right],
\end{aligned}
\tag{1}
$$

where $i = 1, \ldots, I$ indicates the measurement occasion and $j = 1, \ldots, J$ the case. The variable D_{ij} is a dummy variable indicating the phase of the experiment: D_{ij} equals 0 if the measurement was taken during the baseline phase, while D_{ij} equals 1 if the measurement was taken during the treatment phase. The random effects u_{0j} and u_{1j} have respective variances σ^2_{u0} and σ^2_{u1}, and their covariance is σ_{u01}. In this GLMM, γ_{00} represents the average of the logarithm of the baseline level, and γ_{00} represents the logarithm of the treatment effect across the J cases.

Two models are used to analyze the simulated data: one GLMM identical to the one used to generate the data in Eq. 1, and one two-level LMM as defined below:

$$
\begin{aligned}
&Y_{ij} \sim \mathrm{N}\left(\mu_{ij}, \sigma_e\right)\\
&\mu_{ij} = \beta^*_{0j} + \beta^*_{1j} D_{ij}\\
&\begin{cases} \beta^*_{0j} = \gamma^*_{00} + u^*_{0j} \\ \beta^*_{1j} = \gamma^*_{10} + u^*_{1j} \end{cases}\\
&\begin{pmatrix} u^*_{0j} \\ u^*_{1j} \end{pmatrix} \sim \mathrm{MVN}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \left(\sigma^*_{u0}\right)^2 & \sigma^*_{u01} \\ \sigma^*_{u01} & \left(\sigma^*_{u1}\right)^2 \end{pmatrix} \right].
\end{aligned} \tag{2}
$$

Simulation conditions

Design parameters

We refer to I, the number of measurements per case, and to J the number of cases, as 'design parameters' because they influence the single-case experimental design implemented in the simulation. A common practice in SCED research is to vary the length of the baseline phase (Shadish & Sullivan, 2011), so the time point on which the treatment or intervention is introduced is different over cases. This is a so-called multiple baseline design and it is the design implemented in this simulation study. In an SCED context, I will typically be quite small and often J will be even smaller (Shadish & Sullivan, 2011). This might have a significant influence on the fit of the LMM and especially of the GLMM, since the latter, more complex model can be more difficult to estimate if the number of data points is small (Nelson & Leroux, 2008). For many measurements and cases, the fit to the simulated data is expected to be good. The number of measurements I will be defined as either 8, 12, or 20. These values were deliberately chosen to be somewhat smaller than common numbers of measurement occasions, as reported by Moeyaert and et al. (2013) and based on Ferron et al. (2010) and Shadish and Sullivan (2011) and Swanson and Sachse-Lee (2000). This was done in order to test the hypothesis on better relative performance of the LMM with small sample sizes. The number of cases, J, will be defined as either 4, 8, or 10. These values are also close to the values for J chosen in Moeyaert and et al. (2013), which were based on recommendations of Barlow and Hersen (1984) and Kazdin and Kopel (1975) and on the review by Shadish and Sullivan (2011), but the values in this study were chosen to be more spread apart. This was done to have a slightly larger range in levels when considering J as a factor in the analysis of the simulation results. For all combinations of I and J, a list of starting point values (i.e., the first measurement that is part of the treatment phase) is defined. This list has length J and contains the starting point $i \in [1, I]$ for every case j. These starting points were chosen so that they were evenly distributed among different cases and so that both the baseline and the treatment phase contained a substantial number of measurements. Table 1 provides a summary of the design parameter combinations and their corresponding lists of starting point values.

Model parameters

In the GLMM (1) used for generating data, the raw data points Y_{ij} are generated by random sampling from a Poisson $\left(\lambda_{ij}\right)$ distribution. For sufficiently large values of λ_{ij}, however, the normal distribution with mean λ_{ij} and variance λ_{ij} is a good approximation to the Poisson distribution (Johnson et al., 2005). This leads to a hypothesis stating that for the GLMM generated data with sufficiently large λ_{ij}s, the LMM (2) might result in a relatively better fit. To verify this hypothesis, the simulation conditions need to distinguish between generated data that are 'highly discrete' in nature (smaller λ_{ij} values) and generated data that have a more 'continuous' nature (larger λ_{ij} values) due to good approximations by the normal distribution. With two phases (baseline and treatment) and two characterizations (highly discrete or approximately continuous in nature), we obtain four conditions based on the the responses (Table 2). Without loss of generality, this study only includes one

Table 1 Timing of intervention for simulated cases

I	J	Starting point values
8	4	(3, 4, 5, 6)
12	4	(3, 6, 6, 9)
20	4	(5, 10, 10, 15)
8	8	(2, 2, 3, 3, 5, 5, 6, 6)
12	8	(3, 3, 5, 5, 7, 7, 9, 9)
20	8	(5, 5, 8, 8, 12, 12, 15, 15)
8	10	(2, 2, 3, 3, 4, 4, 5, 5, 6, 6)
12	10	(3, 3, 5, 5, 6, 6, 7, 7, 9, 9)
20	10	(5, 5, 8, 8, 10, 10, 12, 15, 15)

Note. I indicates the number of measurements, J indicates the number of cases. The starting point values indicate the first measurement that is part of the treatment phase

Table 2 Categorization of the average baseline response and treatment response

Average Baseline Response	Average Treatment Response
Highly discrete (HD)	Highly discrete (HD)
Highly discrete (HD)	Approximately continuous (AC)
Approximately continuous (AC)	Approximately continuous (AC)

combination of a phase with a highly discrete average response and a phase with an approximately continuous average response, i.e., the second combination listed in Table 2.

The aim of this section is to define values for the nominal fixed effects parameters (γ_{00} and γ_{10}) and variance components (σ^2_{u0}, σ^2_{u1} and σ_{u01}) in such a way that they cover the three combinations of interest listed in Table 2. Thus, the question is the following: how do the values for the model parameters γ_{00}, γ_{10}, σ^2_{u0}, σ^2_{u1} and σ_{u01} affect the model's average response, i.e., $\mathrm{E}\left(\lambda_{ij}\right)$? From the linear expression for λ_{ij} in the GLMM (1) it follows that

$$\begin{aligned}\lambda_{ij} &= \exp\left(\beta_{0j} + \beta_{1j} D_{ij}\right)\\ &= \exp\left(\beta_{0j}\right)\exp\left(\beta_{1j} D_{ij}\right).\end{aligned}$$

Thus, in the baseline phase $\mathrm{E}\left(\lambda_{ij}\right)$ equals $\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]$ and in the treatment phase $\mathrm{E}\left(\lambda_{ij}\right)$ equals $\mathrm{E}\left[\exp\left(\beta_{0j}\right)\exp\left(\beta_{1j}\right)\right]$. An expansion of these expected values of exponentials of β_{0j} and β_{1j} can be obtained based on properties of the multivariate lognormal distribution. Since $\left(u_{0j}, u_{1j}\right)^{\mathsf{T}}$ is sampled from a multivariate normal distribution, $\left(\beta_{0j}, \beta_{1j}\right)^{\mathsf{T}}$ also follows a multivariate normal distribution. Therefore, $\exp\left(\boldsymbol{\beta}\right) = \left[\exp\left(\beta_{0j}\right), \exp\left(\beta_{1j}\right)\right]^{\mathsf{T}}$ follows a multivariate lognormal distribution, of which the elements k of the mean vector of $\exp\left(\boldsymbol{\beta}\right)$ equal

$$\mathrm{E}\left[\exp\left(\boldsymbol{\beta}\right)\right]_k = \exp\left(\mu_k + \frac{1}{2}\Sigma_{kk}\right) \tag{3}$$

and the elements kl of the covariance matrix of $\exp\left(\boldsymbol{\beta}\right)$ equal

$$\begin{aligned}\mathrm{Var}\left[\exp\left(\boldsymbol{\beta}\right)\right]_{kl} &= \exp\left[\mu_k + \mu_l + \frac{1}{2}\left(\Sigma_{kk} + \Sigma_{ll}\right)\right]\\ &\quad\times\left[\exp\left(\Sigma_{kl}\right) - 1\right].\end{aligned} \tag{4}$$

So, if $\boldsymbol{\beta} = \left(\beta_{0j}, \beta_{1j}\right)^{\mathsf{T}} \sim \mathrm{MVN}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ with $\boldsymbol{\mu} = \left(\gamma_{00}, \gamma_{10}\right)^{\mathsf{T}}$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix}\sigma^2_{u0} & \sigma_{u01}\\ \sigma_{u01} & \sigma^2_{u1}\end{pmatrix},$$

we have that

$$\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right] = \exp\left(\gamma_{00} + \frac{\sigma^2_{u0}}{2}\right) \tag{5}$$

$$\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right] = \exp\left(\gamma_{10} + \frac{\sigma^2_{u1}}{2}\right) \tag{6}$$

$$\begin{aligned}\mathrm{Var}\left[\exp\left(\beta_{0j}\right)\right] &= \mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]^2\\ &\quad\times\left[\exp\left(\sigma_{u0}\right) - 1\right]\end{aligned} \tag{7}$$

$$\begin{aligned}\mathrm{Var}\left[\exp\left(\beta_{1j}\right)\right] &= \mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2\\ &\quad\times\left[\exp\left(\sigma_{u1}\right) - 1\right]\end{aligned} \tag{8}$$

$$\begin{aligned}\mathrm{Cov}\left[\exp\left(\beta_{0j}\right), \exp\left(\beta_{1j}\right)\right] &= \mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]\\ &\quad\times\left[\exp\left(\sigma_{u01}\right) - 1\right]\end{aligned} \tag{9}$$

Equation 5 describes the average baseline response. By combining Eq. 5, Eq. 6 and the formula for the expected value of the product of two dependent variables, an expression for the average treatment response can be derived:

$$\begin{aligned}\mathrm{E}\left[\exp\left(\beta_{0j}\right)\exp\left(\beta_{1j}\right)\right] &= \mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]\\ &\quad+\mathrm{Cov}\left[\exp\left(\beta_{0j}\right), \exp\left(\beta_{1j}\right)\right]\\ &= \mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]\\ &\quad\times\left[1 + \exp\left(\sigma_{u01}\right) - 1\right]\\ &= \mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\\ &\quad\times\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]\exp\left(\sigma_{u01}\right)\end{aligned} \tag{10}$$

An important point to notice here is that the expected treatment response $\mathrm{E}\left[\exp\left(\beta_{0j}\right)\exp\left(\beta_{1j}\right)\right]$ is not merely equal to the expected baseline response $\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]$ times $\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]$. Equation 10 shows the influence of the σ_{u01} parameter.

These derivations illustrate how the average baseline and treatment responses depend on the model parameters in a not very straightforward way. The average baseline response depends in a non-linear way on not only γ_{00} but also σ^2_{u0}. The average treatment response depends in a non-linear way on all five model parameters γ_{00}, γ_{10}, σ^2_{u0}, σ^2_{u1} and σ_{u01} together. Therefore, in this simulation study, nominal values for Eqs. 5 and 6 are chosen rather than values for the γ_{00}, γ_{10}, σ^2_{u0}, σ^2_{u1} model parameters directly. This makes managing the categorization of conditions in the Table 2 categories easier. Summarizing the choice of values for $\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]$, $\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]$ and σ_{u01}, the conditions and their categorizations are listed as 'highly discrete' or 'approximately continuous' in Table 3.

After having defined values for $\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]$ and $\mathrm{E}\left[\exp\left(\beta_{0j}\right)\exp\left(\beta_{1j}\right)\right]$ in Table 3, the choice of σ^2_{u0} and σ^2_{u1} values will uniquely determine the corresponding values for γ_{00} and γ_{10} as shown in Eqs. 5 and 6. Since there are no particular restrictions for values of γ_{00} and γ_{10}, the focus will now be on well defining values for the variance com-

Table 3 Simulation conditions based on categorization of the average baseline response, treatment response, and effect size

$E[\exp(\beta_{0j})]$ Average baseline response	$E[\exp(\beta_{1j})]$	$E[\exp(\beta_{0j})\exp(\beta_{1j})]$ Average treatment response	Baseline category	Treatment category	Multiplicative effect category
(a) $\sigma_{u01} = 0$					
4	1	4	Highly discrete (HD)	Highly discrete (HD)	Zero
4	1.5	6	Highly discrete (HD)	Highly discrete (HD)	Small
2	3.5	7	Highly discrete (HD)	Highly discrete (HD)	Large
4	3.5	14	Highly discrete (HD)	Approximately continuous	Large
20	1	20	Approximately continuous (AC)	Approximately continuous (AC)	Zero
30	1.6	48	Approximately continuous (AC)	Approximately continuous (AC)	Small
20	3.5	70	Approximately continuous (AC)	Approximately continuous (AC)	Large
(b) $\sigma_{u01} = \log(1.05)$					
4	1	4.2	Highly discrete (HD)	Highly discrete (HD)	Zero
4	1.5	6.3	Highly discrete (HD)	Highly discrete (HD)	Small
2	3.5	7.35	Highly discrete (HD)	Highly discrete (HD)	Large
4	3.5	14.7	Highly discrete (HD)	Approximately continuous (AC)	Large
20	1	21	Approximately continuous (AC)	Approximately continuous (AC)	Zero
30	1.6	50.4	Approximately continuous (AC)	Approximately continuous (AC)	Small
20	3.5	73.5	Approximately continuous (AC)	Approximately continuous (AC)	Large

Note. σ_{u01} is the covariance between the random effects

ponents σ^2_{u0} and σ^2_{u1}. These variance components have an influence on the variance of $\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]$ and $\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]$ as shown in Eqs. 7 and 8. Note that in Table 3, deliberate choices of values were made for these expected values because they should cover all categories. If the variance of $\exp\left(\beta_{0j}\right)$ and $\exp\left(\beta_{0j}\right)\exp\left(\beta_{1j}\right)$ is large, however, a relatively high amount of generated β's will yield values of $\exp\left(\beta_{0j}\right)$ and $\exp\left(\beta_{0j}\right)\exp\left(\beta_{1j}\right)$, which do not fall into the foreseen categories from Table 3. This is due to positive skewness of the lognormal distribution when the underlying normal variance (i.e., σ^2_{u0} and σ^2_{u1}) is larger. Therefore, two values are defined for both σ^2_{u0} and σ^2_{u1}: $\left[\log\left(1.35\right)\right]^2$ and $\left[\log\left(1.50\right)\right]^2$. According to Eqs. 7 and 8, the corresponding variances for $\sigma^2_{u0} = \sigma^2_{u1} = \left[\log\left(1.50\right)\right]^2$ equal:

$$\begin{aligned}
\mathrm{Var}\left[\exp\left(\beta_{0j}\right)\right] &= \mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\left[\exp\left(\sigma_{u0}\right) - 1\right] \\
&= \mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\left[\exp\left(\log\left(1.50\right)\right) - 1\right] \\
&= \frac{1}{2}\cdot\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right] \\
\mathrm{Var}\left[\exp\left(\beta_{1j}\right)\right] &= \mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]\left[\exp\left(\sigma_{u1}\right) - 1\right] \\
&= \frac{1}{2}\cdot\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right].
\end{aligned}$$

So the variances will equal 35% ($\sigma^2_{u0} = \sigma^2_{u1} = \left[\log\left(1.35\right)\right]^2$) or 50% ($\sigma^2_{u0} = \sigma^2_{u1} = \left[\log\left(1.50\right)\right]^2$) of the expected values.

A final condition to check is whether the choices of values for the variance components yield a positive semi-definite covariance matrix. This is equivalent to making sure that the correlation between β_{0j} and β_{1j} is between -1 and 1, or that $|\sigma_{u01}| \leq |\sigma_{u0}\sigma_{u1}|$. Checking this restriction for the largest value of σ_{u01} (i.e., $\sigma_{u01} = \log(1.05)$) and the smallest values of σ_{u0} and σ_{u1} (i.e., $\sigma_{u0} = \sigma_{u1} = \log(1.35)$), one can verify that this condition is indeed met:

$$|\sigma_{u01}| \leq |\sigma_{u0}\sigma_{u1}| \Leftrightarrow \log(1.05) \leq \left[\log\left(1.35\right)\right]^2.$$

Analysis

Goodness of fit

To assess the goodness of fit of the GLMM and the LMM, the Akaike information criterion (AIC, Akaike (1998)) and the Bayesian information criterion (BIC, Schwarz (1978) and Claeskens and Jansen (2015)) are used. In every iteration of the simulation, the AIC and the BIC of the GLMM and the LMM fits are computed. Next, a relative AIC and BIC score is calculated by taking the relative difference of the LMM and GLMM goodness of fit criteria (resp. denoted as $\mathrm{AIC_L}$ or $\mathrm{BIC_L}$ for LMM and $\mathrm{AIC_G}$ or $\mathrm{BIC_G}$ for GLMM):

$$S_{\mathrm{AIC}} = \frac{\mathrm{AIC_L} - \mathrm{AIC_G}}{\mathrm{AIC_G}} \tag{11}$$

$$S_{\mathrm{BIC}} = \frac{\mathrm{BIC_L} - \mathrm{BIC_G}}{\mathrm{BIC_G}} \tag{12}$$

The motivation behind these scores is that they provide a comparison between the LMM and GLMM in one score, and that these scores in turn are comparable across conditions. This facilitates representation of the goodness of fit results in a clear and compact figure later in the analysis. When $S_{\mathrm{AIC}} < 0$ or $S_{\mathrm{BIC}} < 0$, the LMM fit results in a lower AIC or BIC and this would lead to the conclusion that the LMM provides a better fit than the GLMM. The reverse finding, i.e., $S_{\mathrm{AIC}} > 0$ or $S_{\mathrm{BIC}} > 0$, would lead to the conclusion that the GLMM provides a better fit than the LMM. Per condition, the mean $\overline{S}_{\mathrm{AIC}}$ and $\overline{S}_{\mathrm{BIC}}$ are each calculated over all iterations.

Fixed effect parameter recovery

In SCED research, the main interest is usually in the treatment effect and its size (Van den Noortgate & Onghena, 2008). When analyzing SCED data with the classical continuous linear mixed model as expressed in Eq. 2, the corresponding parameter of interest is γ^*_{10}. This parameter expresses the average increase or decrease in baseline response across cases after the treatment or intervention. Note that this is an additive change: the average baseline response changes from

$$\mathrm{E}\left(\mu_{ij}|D_{ij} = 0\right) = \gamma^*_{00}$$

to

$$\mathrm{E}\left(\mu_{ij}|D_{ij} = 1\right) = \gamma^*_{00} + \gamma^*_{10}$$

in the treatment phase. Thus, the fixed effect γ^*_{10} expresses the average difference between the expected baseline response and the expected treatment response. However, the GLMM fixed effect parameter γ_{10} cannot be interpreted in the same way. Indeed, interpretation of γ_{10} is not as straightforward. Equations 5, 6 and 10 show how the expected treatment response does not even merely equal the expected baseline response times $\exp\left(\gamma_{10}\right)$ because of the influence of the variance components.

This observation leads to the following complication in this simulation study. Data are generated from the GLM model as defined in Eq. 1, with a nominal value for γ_{10}. Afterwards, the LMM as defined in Eq. 2 is fit, which yields an estimate $\hat{\gamma}^*_{10}$. However, this $\hat{\gamma}^*_{10}$ will not be comparable with the nominal γ_{10}, since γ_{10} and γ^*_{10} are two different parameters and they do not express the same concept.

To address this complication, two approaches are proposed. Both approaches provide a transformation of the

parameters of one of the models into a new parameter. This new parameter is comparable to the fixed effect parameter of the other model and therefore a fixed parameter recovery assessment can be conducted based on the new parameter estimate from the first model and the fixed effect parameter estimate from the second model. Note that a general investigation of transformations of effect sizes based on the LMM to effect sizes based on the GLMM and vice versa is not within the scope of this paper, though this might be interesting for future research.

The first approach consists of a transformation of the GLMM parameters into a new parameter Δ_G, which expresses an effect size comparable to the fixed effect γ_{10}^* of the LMM. By comparing the estimate for Δ_G from the GLMM and the estimate for γ_{10}^* from the LMM we can assess the fixed effect parameter recovery. The second approach is analogous, but uses the LMM as a starting point instead. Based on a transformation the LMM parameters, it introduces a new fixed effect parameter Γ_L and this Γ_L is subsequently compared to γ_{10} to assess fixed parameter recovery.

The first metric Δ is defined as the additive effect of the treatment. This additive effect should express the difference of the average baseline response and the average treatment response:

$$\Delta = \mathrm{E}\,(\mathrm{Tx}) - \mathrm{E}\,(\mathrm{B}) \tag{13}$$

For the GLMM, Eqs. 5 and 10 can be used to define a Δ_G parameter:

$$\begin{aligned}\Delta_G &= \mathrm{E}\left[\exp\left(\beta_{0j}\right)\exp\left(\beta_{1j}\right)\exp\left(\sigma_{u01}\right)\right] - \mathrm{E}\left[\exp\left(\beta_{0j}\right)\right] \\ &= \exp\left(\gamma_{00} + \frac{\sigma_{u0}^2}{2}\right)\left[\exp\left(\gamma_{10} + \frac{\sigma_{u1}^2}{2} + \sigma_{u01}\right) - 1\right]\end{aligned} \tag{14}$$

The parameter Δ_G can be computed for every condition using the parameters used in data generation and substituting them into Eq. 14. The estimator $\hat{\Delta}_G$ can be estimated for each simulated dataset by substituting the estimated parameters into Eq. 14:

$$\hat{\Delta}_G = \exp\left(\hat{\gamma}_{00} + \frac{\hat{\sigma}_{u0}^2}{2}\right)\left[\exp\left(\hat{\gamma}_{10} + \frac{\hat{\sigma}_{u1}^2}{2} + \hat{\sigma}_{u01}\right) - 1\right] \tag{15}$$

For the LMM, a Δ_L parameter is defined analogously:

$$\begin{aligned}\Delta_L &= \mathrm{E}\left(\beta_{0j} + \beta_{1j}\right) - \mathrm{E}\left(\beta_{0j}\right) \\ &= \mathrm{E}\left(\beta_{0j}\right) + \mathrm{E}\left(\beta_{1j}\right) - \mathrm{E}\left(\beta_{0j}\right) \\ &= \mathrm{E}\left(\beta_{1j}\right) \\ &= \gamma_{10}^*\end{aligned} \tag{16}$$

The parameter Δ_L can be computed for every condition using the parameters used in data generation and substituting them into Eq. 14. The estimator $\hat{\Delta}_L$ can be estimated for each simulated dataset by $\hat{\gamma}_{10}^*$:

$$\hat{\Delta}_L = \hat{\gamma}_{10}^* \tag{17}$$

The second metric Γ is defined by the following expression based on the expected baseline and treatment responses and on the variance in the baseline and in the treatment:

$$\Gamma = \log\left[\left(\frac{\mathrm{E}\,(\mathrm{Tx})}{\mathrm{E}\,(\mathrm{B})}\right)^2 \sqrt{\frac{\mathrm{E}\,(\mathrm{B})^2 + \mathrm{Var}\,(\mathrm{B})}{\mathrm{E}\,(\mathrm{Tx})^2 + \mathrm{Var}\,(\mathrm{Tx})}}\right] \tag{18}$$

For the GLMM, it can be shown that the above expression equals γ_{10}. These calculations are provided in Appendix A. Thus a Γ_G parameter is defined as:

$$\Gamma_G = \gamma_{10} \tag{19}$$

The parameter Γ_G can be computed for every condition using the parameters used in data generation and substituting them into Eq. 19. The estimator $\hat{\Gamma}_G$ can be estimated for each simulated dataset by $\hat{\gamma}_{10}$:

$$\hat{\Gamma}_G = \hat{\gamma}_{10} \tag{20}$$

For the LMM, according to Eq. 2 we have that

$$\begin{aligned}\mathrm{E}\,(\mathrm{B}) &= \gamma_{00}^* \\ \mathrm{Var}\,(\mathrm{B}) &= \left(\sigma_{u0}^*\right)^2 \\ \mathrm{E}\,(\mathrm{Tx}) &= \gamma_{00}^* + \gamma_{10}^* \\ \mathrm{Var}\,(\mathrm{Tx}) &= \left(\sigma_{u0}^*\right)^2 + \left(\sigma_{u1}^*\right)^2 + 2\sigma_{u01}^*.\end{aligned}$$

Thus a Γ_L parameter is defined as follows:

$$\Gamma_L = \log\left[\left(\frac{\gamma_{00}^* + \gamma_{10}^*}{\gamma_{00}^*}\right)^2 \times \sqrt{\frac{\left(\gamma_{00}^*\right)^2 + \left(\sigma_{u0}^*\right)^2}{\left(\gamma_{00}^* + \gamma_{10}^*\right)^2 + \left(\sigma_{u0}^*\right)^2 + \left(\sigma_{u1}^*\right)^2 + 2\left(\sigma_{u01}^*\right)}}\right] \tag{21}$$

The parameter Γ_L can be computed for every condition using the parameters used in data generation and substituting them into Eq. 19. The estimator $\hat{\Gamma}_L$ can be estimated

for each simulated dataset by substituting the estimated parameters into Eq. (21):

$$\hat{\Gamma}_L = \log\left[\left(\frac{\hat{\gamma}^*_{00}+\hat{\gamma}^*_{10}}{\hat{\gamma}^*_{00}}\right)^2 \times \sqrt{\frac{\left(\hat{\gamma}^*_{00}\right)^2+\left(\hat{\sigma}^*_{u0}\right)^2}{\left(\hat{\gamma}^*_{00}+\hat{\gamma}^*_{10}\right)^2+\left(\hat{\sigma}^*_{u0}\right)^2+\left(\hat{\sigma}^*_{u1}\right)^2+2\left(\hat{\sigma}^*_{u01}\right)}}\right] \tag{22}$$

For each of these parameters Δ and Γ, the relative bias (RB) and the mean squared error (MSE) are calculated.

Inference

In SCED meta-analysis, researchers are interested in finding out if there is an effect of a treatment or intervention. This is expressed in an effect size: a metric indicating the direction and the size of the effect. In multilevel modeling of SCED meta-analytical data, typically the fixed effects are chosen as effect sizes (i.e., γ_{10} in a GLMM (1) and γ^*_{10} in a LMM (2)). Because the data in this simulation study are simulated according to the GLMM in Eq. 1, the parameter of interest here is γ_{10}. The binary hypotheses on which inference in the GLMM setting is based, are:

$$H_0 : \gamma_{10} = \Gamma_G = 0$$
$$H_\alpha : \gamma_{10} = \Gamma_G \neq 0$$

We calculate the proportion of rejections of the null hypothesis per condition, i.e., the proportion of GLMMs estimated yielding a p value smaller than the significance level α for the $\hat{\gamma}_{10}$ estimate. In conditions where the nominal γ_{10} equals 0, this proportion equals the type I error rate. In conditions where the nominal γ_{10} does not equal 0, this proportion equals the power.

For the LMM however, p values are calculated based on a different set of hypotheses:

$$H_0 : \gamma^*_{10} = \Delta_L = 0$$
$$H_\alpha : \gamma^*_{10} = \Delta_L \neq 0$$

Again, we calculate the proportion of null hypothesis rejections per condition. We have to interpret this proportion based on the nominal Δ_G value (14), since Δ_L should estimate the same additive treatment effect. In conditions where the nominal Δ_G equals 0, the proportion of rejections

Table 4 Simulation condition factors summary

Parameter	Value	Motivation
γ_{00}	$\log(2) - \frac{\sigma^2_{u0}}{2}$	Average baseline response highly discrete: $\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right] = 2$
	$\log(4) - \frac{\sigma^2_{u0}}{2}$	Average baseline response highly discrete: $\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right] = 4$
	$\log(20) - \frac{\sigma^2_{u0}}{2}$	Average baseline response approximately normal: $\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right] = 20$
γ_{10}	0	To test $H_0 : \gamma_{10} = 0$
	$-\frac{\sigma^2_{u1}}{2}$	To test $H_0 : \gamma^*_{10} = 0 \Leftrightarrow \left[\left(\gamma_{10} = -\frac{\sigma^2_{u1}}{2}\right) \wedge (\sigma_{u01} = 0)\right]$
	$\log(3.5) - \frac{\sigma^2_{u1}}{2}$	Larger average multiplicative effect: $\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right] = 3.5$
σ^2_{u0}	$\left[\log(1.35)\right]^2$	$\mathrm{Var}\left[\exp\left(\beta_{0j}\right)\right] = 35\% \cdot \mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]$
	$\left[\log(1.50)\right]^2$	$\mathrm{Var}\left[\exp\left(\beta_{0j}\right)\right] = 50\% \cdot \mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]$
σ^2_{u1}	$\left[\log(1.35)\right]^2$	$\mathrm{Var}\left[\exp\left(\beta_{1j}\right)\right] = 35\% \cdot \mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]$
	$\left[\log(1.50)\right]^2$	$\mathrm{Var}\left[\exp\left(\beta_{1j}\right)\right] = 50\% \cdot \mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]$
σ_{u01}	0	To test $H_0 : \gamma^*_{10} = 0 \Leftrightarrow \left[\left(\gamma_{10} = -\frac{\sigma^2_{u1}}{2}\right) \wedge (\sigma_{u01} = 0)\right]$
	$\log(1.05)$	Small influence on multiplicative effect: $\exp\left(\sigma_{u01}\right) = 1.05$
I	8	
	12	Common SCED values
	20	
J	4	
	8	Common SCED values
	10	

per condition equals the type I error rate. In conditions where the nominal Δ_G does not equal 0, this proportion equals the power. Note that according to Eq. 14 we have that

$$\Delta_G = 0 \Leftrightarrow \left[\left(\gamma_{10} = -\frac{\sigma_{u1}^2}{2}\right) \wedge (\sigma_{u01} = 0)\right].$$

This expression will be the motivation for the choice of values for γ_{10} and σ_{u01}.

The p values for the LMM are computed based on the approximate Wald F-test with Satterthwaite denominator degrees of freedom (Gumedze & Dunne, 2011; Satterthwaite, 1946). The underlying p values for the GLMM are computed based on an approximate Wald Z-test. The choice of Z-test for inference based on the GLMM was due to practical constraints with `lme4` (Bates et al., 2015), the package we used for simulation in R (R Core Team, 2017). We elaborate on this further in Appendix B. The significance level α is set to .05.

Simulation conditions

All design and model parameters and their choice of values have been summarized in Table 4, together with a motivation based on the calculations and analyses described in the previous paragraphs. The total number of conditions equals $3 \times 3 \times 2 \times 2 \times 2 \times 3 \times 3 = 648$. For Δ_G (Eq. 14), which depends on all five model parameters (γ_{00}, γ_{10}, σ_{u0}^2, σ_{u1}^2 and σ_{u01}), the particular choice of values for these parameters (see Table 4) resulted in 22 unique nominal parameter values. For Γ_G (Eq. 19), which depends on γ_{10} and on σ_{u1}^2, this resulted in $(2 \times 2) + 1 = 5$ unique nominal parameter values. The 22 Δ_G nominal parameter values range from 0 to 53.5. The (rounded) Γ_G nominal parameter values equal $-.0822$, $-.0450$, 0, 1.1706, and 1.2077. To keep a balance between feasibility and precision with as many as 648 conditions, we generate $N = 2000$ datasets per condition. With this number of simulated datasets, a condition with a true type I error rate of .05 would have an estimated type I error rate with a standard error of $\sqrt{\frac{.05 \times .95}{2000}} = .0049$, and because we will analyze the results across multiple conditions rather than within individual conditions, the analyses will be based on multiples of 2000 datasets.

Recall that for fixed parameter recovery, the relative bias and the MSE will be analyzed. Two careful considerations have to be made in order to obtain meaningful results for the relative bias and the MSE. First, since Γ_G can take on negative nominal parameter values, the sign of the relative bias will be influenced when dividing by these nominal parameter values. Therefore we opt to calculate a modified relative bias by dividing the bias by the absolute value of the nominal parameter value:

$$\frac{\overline{\hat{\theta}_i} - \theta}{|\theta|}$$

Second, the MSE is relative with respect to the nominal parameter value, which makes MSEs difficult to compare when the range of nominal parameter values is large (as it is for Δ_G). Therefore we opt to calculate a relative MSE by dividing the MSE by the squared nominal parameter value:

$$\frac{\text{MSE}\left(\hat{\theta}_i\right)}{\theta^2}$$

Summary of results

To address the previously stated research objectives, we will compare the LMM and GLMM results and discuss their performance in terms of goodness of fit (S_{AIC} (Eq. 11) and S_{BIC} (Eq. 12)), fixed effect parameter recovery (quantified by the MSEs and relative bias of the Δ and Γ estimators), type I error rate and power. We studied the effect of the following design factors: baseline-treatment category (as defined in Table 3), effect size category (as defined in

Table 5 Eta-squared values (η^2) for association of design factors with outcomes

	S_{AIC}	S_{BIC}	MSE Δ	MSE Γ	RB Δ	RB Γ	Type I error rate	Power
Model			.0004	.0002	.0853	.0833	**.3909**	.0325
I	.0230	.0141	.0099	.0177	.0045	.0442	.0296	.0003
J	.0572	.0396	.0976	.0809	.0080	.0050	.0478	.0137
Baseline-treatment category	**.6188**	**.6510**	.0304	.0553	.0011	.0388	**.2487**	.0022
Effect size	.0645	.0589	**.2453**	**.3288**	.0438	**.2464**		**.8835**
Model:I			.0001	.0001	.0004	.0007	.0034	.0000
Model:J			.0005	.0001	.0136	.0005	.0658	.0121
Model:(Baseline-treatment category)			.0001	.0002	.0021	.0039	.0421	.0000
Model:(Effect size)			.0002	.0001	.0306	.0910		.0072
	.7635	.7636	.3844	.4833	.1892	.5139	.8283	.9515

Table 3), number of measurements I and number of cases J. We choose to study the effect of the baseline treatment and the effect size categories rather than the effect of the individual model parameters γ_{00}, γ_{10}, σ^2_{u0}, σ^2_{u1} and σ_{u01}, because (1) they are more easily interpretable, (2) they relate directly to the research questions stated in this study, and (3) simulation conditions were generated using these categories rather than the individual parameters. To assess which impact these factors have on the performance outcomes, we conduct an ANOVA analysis and calculate η^2 values. The results are shown in Table 5. To avoid discussing trivial effects, we will only discuss factors that explain at least 14% of the variance in the outcome variables (results shown in bold in Table 5). The cutoff value of 14% is based on the rule of thumb suggested by Cohen (1988). However, we choose to include the factor `Model` in all of our results because of our explicit interest in assessing the performance of the LMM by using the GLMM's performance results as a benchmark.

For graphical purposes, the baseline-treatment categories from Table 3 are denoted as follows in the graphical results: a highly discrete average baseline response and a highly discrete average treatment response is denoted as category 'HD-HD' (from 'highly discrete - highly discrete'), a highly discrete average baseline response and an approximately continuous average treatment response is denoted as category 'HD-AC' (from 'highly discrete - approximately continuous') and finally an approximately normal average baseline response and an approximately normal average treatment response is denoted as category 'AC-AC' (from 'approximately continuous - approximately continuous').

Software

We use the open-source R software (R Core Team, 2017) to generate and analyze the SCED count data. The LMM and the GLMM are estimated through the `lmer()` and `glmer()` functions, respectively, both available in the `lme4` package (Bates et al., 2015). Using the default argument settings, the `lmer()` function provides restricted maximum likelihood (REML) estimates for the LMM parameters and the `glmer()` function provides estimates based on a Gauss–Hermite quadrature approximation of the log-likelihood function. In Appendix B, we provide some R code samples and explain how we obtained and analyzed the LMM and GLMM estimates.

Results

Goodness of fit criteria

Previously, a relative AIC score S_{AIC} and a relative BIC score S_{BIC} were defined (see Eqs. 11 and 12). Analysis results for both scores are very similar, thus only results for the S_{AIC} scores are reported in this paper. From Table 5, we see that most of the variability in S_{AIC} is associated with the baseline-treatment category. Figure 1 shows the

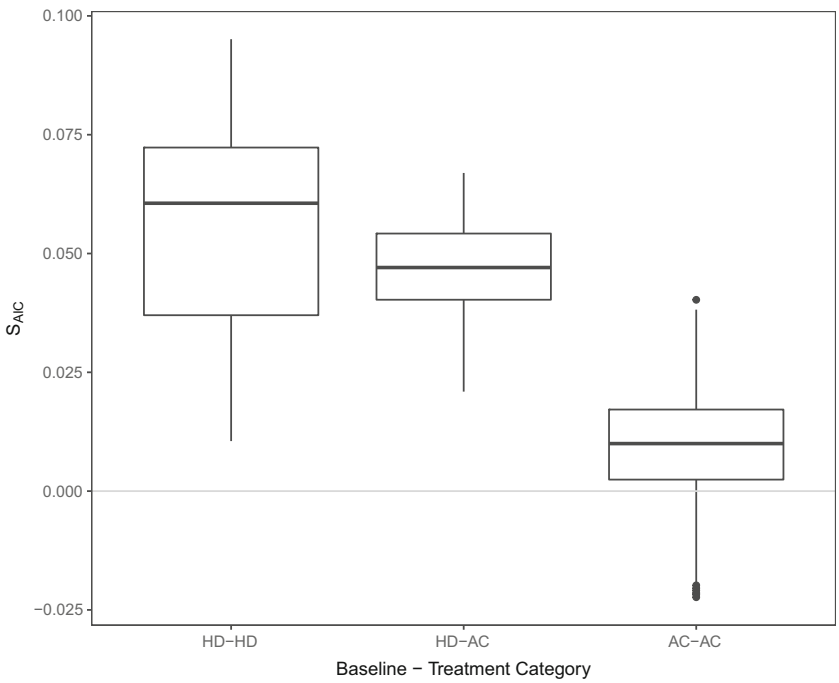

Fig. 1 Mean AIC scores S_{AIC}. Baseline-treatment categories are based on Table 3

distribution of all S_{AIC} scores for each of the baseline-treatment categories. By definition, a negative value of S_{AIC} indicates that the LMM fit results in a lower AIC than the GLMM fit, and thus the LMM performs relatively better. Figure 1 shows that this is almost never the case, except for some observations within the AC-AC category. A closer look at the conditions that yielded negative mean S_{AIC} scores learns that those scores only occur when $J = 4$ or when $I = 8$ and $J = 8$, indicating that only when the data are approximately normal and information is sparse, the GLMM no longer outperforms the LMM in terms of goodness of fit. This is due to both the fact that the GLMM is more complex to estimate and to the fact that count data with higher expected values are better approximated by a normal distribution, which makes the LMM's assumption of normally distributed residuals and therefore a normally distributed dependent variable more plausible.

Fixed effect parameter recovery

For the statistics Δ_G (Eq. 14), Δ_L (Eq. 16), Γ_G (Eq. 19) and Γ_L (Eq. 21), a simple linear regression analysis is conducted to study the relation between the LMM and the GLMM estimators for Δ and Γ. The fitted model predicts the LMM estimate based on the GLMM estimates. A significant regression equation was found for both the Δ and the Γ estimates, with an R^2 of .9986 ($\Delta_L = 0.0225 + 1.0042 \cdot \Delta_G$) and .9963 ($\Gamma_L = -0.0066 + 1.0088 \cdot \Gamma_G$), respectively. This is an important result because it allows for comparison between the GLMM and the LMM based on their parameter estimates. Now that it is clear that there is a way to compare the fixed effect estimations of the GLMM and the LMM, the next step is to assess which model provides the best fixed effect estimator. To assess the quality of the Δ_G, Δ_L, Γ_G and Γ_L as estimators, the relative bias and the relative MSE of all four are analyzed. Note that conditions where $\Delta = 0$ or $\Gamma = 0$ were left out in order to be able to calculate a finite relative bias and relative MSE.

For the relative bias of the Δ estimates, we see from Table 5 that none of the design factors is associated with an η^2 value higher than our cutoff value of 14%. The total η^2 for the relative bias of Δ equals 0.1892. Because we have corrected for all factors on which we defined our simulation conditions in the ANOVA analysis, this low total of η^2 values indicates that most of the variation in relative bias of Δ must be due to sampling error. Across all conditions, the relative bias of Δ ranges from -0.035 to 0.38 with a median of 0.00097, indicating that for many conditions, Δ is unbiased. The factors with higher η^2 values in Table 5 give an indication as to which factors affect biasedness in the Δ estimates. In Fig. 2, the relative bias is shown across different levels of two factors with relatively high η^2 values, i.e., model and effect size. The LMM is the model which estimates Δ directly and its associated estimator Δ_L is less biased than the GLMM's Δ_G estimator. This is especially true when the effect size is small, although even then the relative bias of Δ_G is still reasonably small.

The relative bias of Γ is mostly associated with the size of the effect ($\eta^2 = .2464$ in Table 5) and is shown in Fig. 3. Again, the model which cannot directly estimate the

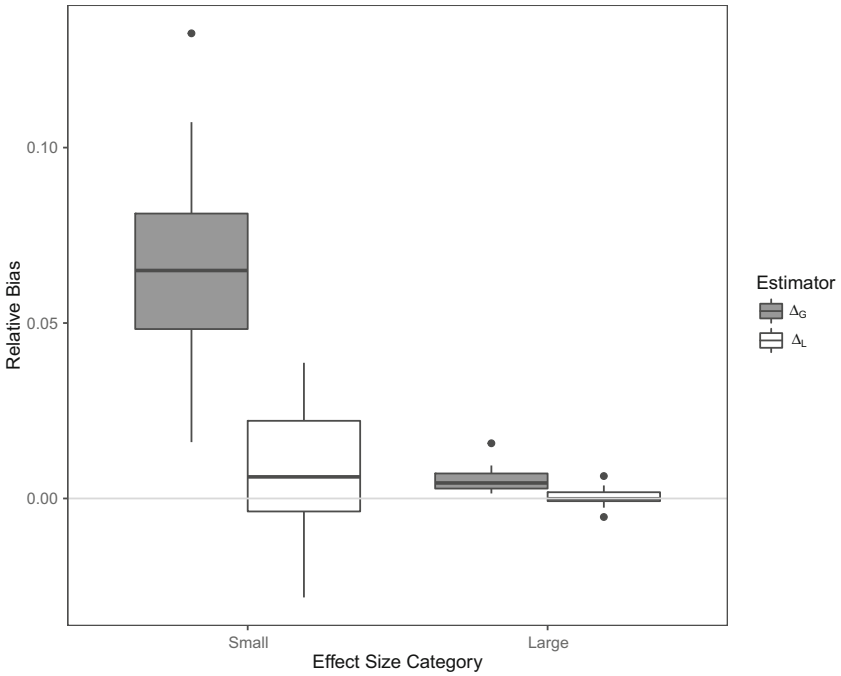

Fig. 2 Relative bias of the Δ_G and Δ_L estimators. Effect size categories are based on Table 3

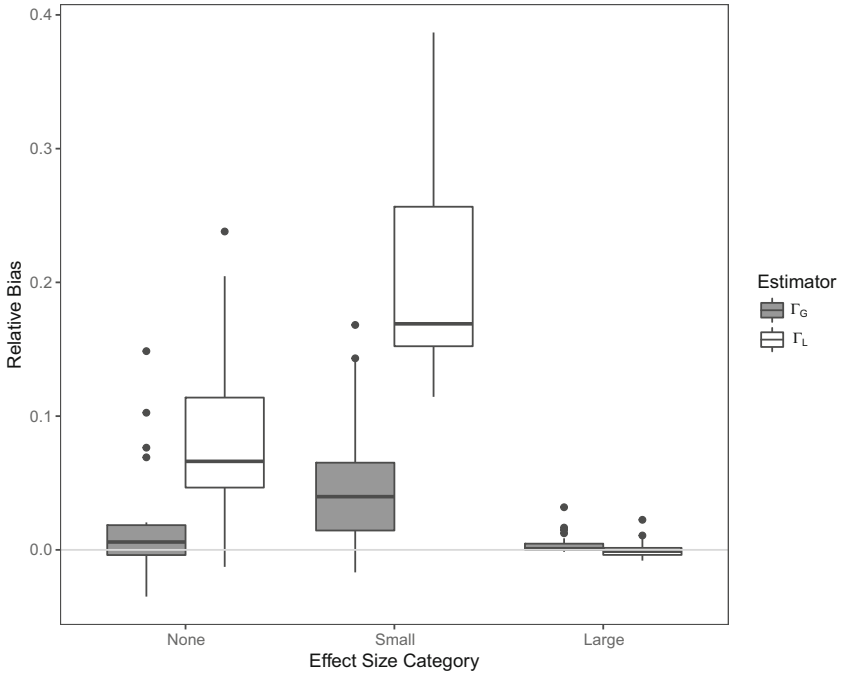

Fig. 3 Relative bias of the Γ_G and Γ_L estimators. Effect size categories are based on Table 3

statistic, i.e., LMM, appears to be most biased. The relative bias of Γ_L goes up to 40% when the effect size is small. Remarkably, the relative bias for Γ is highest for small effect sizes, and slightly lower when the effect size is zero. For large effect sizes, however, both the LMM and the GLMM estimators have very little bias. Looking deeper into the high relative bias observed in conditions where the effect size is small to zero, we see in Fig. 4 that the higher relative biases are associated with conditions where J is small and, to a lesser extent, with conditions where the underlying data are highly discrete. These observations hold true for both Γ_L and Γ_G.

From Table 5 we see that the relative MSE values hardly depend on the underlying model. The effect size category

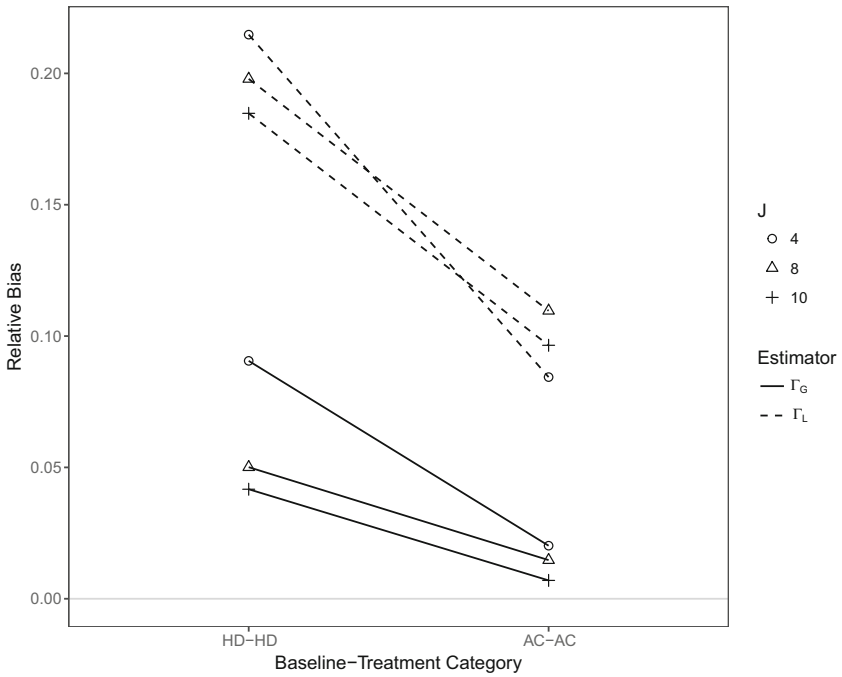

Fig. 4 Relative bias of the Γ_G and Γ_L estimators for conditions where the effect size is small to zero. Baseline-treatment categories are based on Table 3

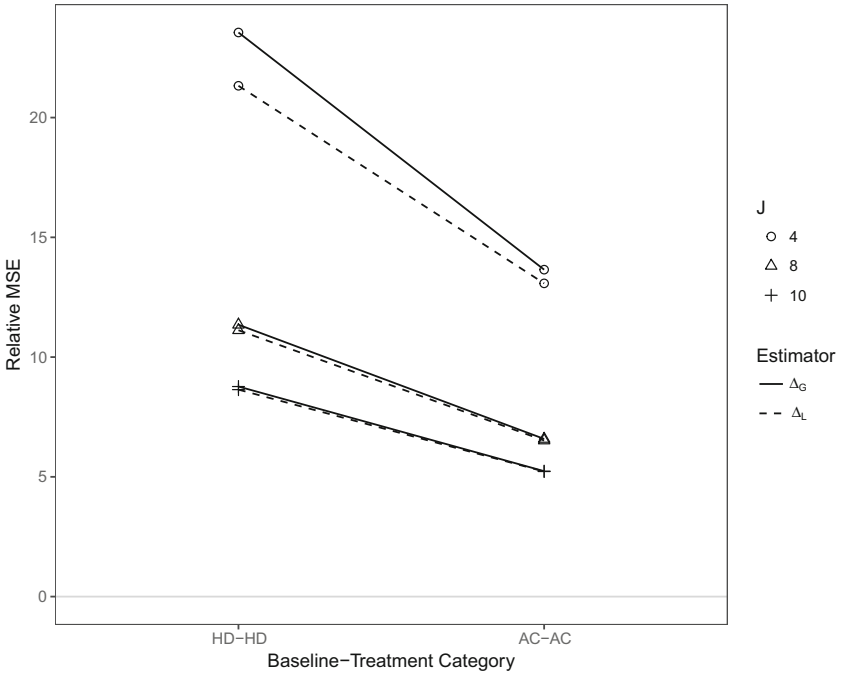

Fig. 5 Relative MSE of the Δ_G and Δ_L estimators for conditions where the effect size is small to zero. Baseline-treatment categories are based on Table 3

has the largest association and a closer look to the MSE values reveals that the relative MSE values for both Γ and Δ are highest when the effect size is small to zero. These conditions are investigated further in Figs. 5 and 6. We compare across different levels of baseline-treatment category and number of cases J because those factors yield the second and third highest η^2 values in Table 5. When J is small and/or when the underlying data are more discrete in nature (as in the HD-HD category), the relative MSEs are higher. Since the GLMM and the LMM have very similar

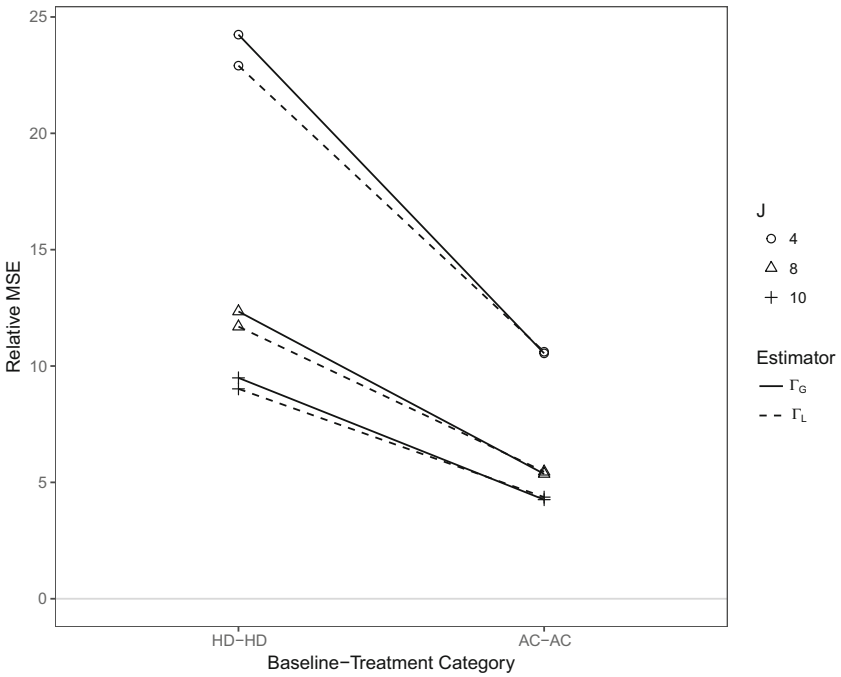

Fig. 6 Relative MSE of the Γ_G and Γ_L estimators for conditions where the effect size is small to zero. Baseline-treatment categories are based on Table 3

Table 6 Overall mean relative bias and relative mean squared error for the Δ and Γ parameter estimates

Parameter	Model	Relative MSE	Relative bias
$\hat{\Delta}^a$	GLMM	7.5580	.0434
	LMM	7.1689	.0025
$\hat{\Gamma}^b$	GLMM	6.2523	.0246
	LMM	5.9945	.0825

[a]Overall means are based on 540 conditions in which $\Delta \neq 0$

[b]Overall means are based on 432 conditions in which $\Gamma \neq 0$

relative MSEs no matter the baseline-treatment category, we cannot conclude that the LMM's relative MSE's improve (relative to those of the GLMM) when the underlying data become more continuous.

To summarize the results of the relative bias and the relative MSEs of the Δ and Γ parameter estimates, Table 6 shows the overall means of both measures. This table confirms again that the relative MSE is very similar for both models, but it also shows a slight disadvantage for the model which cannot directly estimate the parameter. The latter observation is more clear for the relative bias, as was also clear from Figs. 2 and 3.

Inference

Because the type 1 error rates are calculated as the proportion rejections in conditions where the nominal effect is zero, and because in that case the effect size factor variable only has one level, effect size was left out in the ANOVA analysis. Based on the η^2 values in Table 5, we look further into how the baseline-treatment category has an effect on the type 1 error rate ($\eta^2 = .2487$). From Table 7, it is clear that the type I error rate of the GLMM is higher than that of the LMM. In conditions where the underlying data are highly discrete, the type I error rate of the GLMM improves, but a closer look to the data revealed that the type I error rate of the LMM is consistently closer to the nominal $\alpha = .05$ than the type I error rate of the GLMM.

The power naturally depends mostly on the effect size ($\eta^2 = .8835$, Table 5), because power generally increases for larger, more noticeable effects. The impact of all other factors falls below our 14% cutoff for η^2. However, to study what sample sizes are needed to reach an acceptable power

Table 7 Type I error rates

	Type I error rate	
Baseline-treatment category	GLMM[a]	LMM[b]
HD - HD	.07	.04
AC - AC	.12	.05

[a]$H_0 : \Gamma = 0$

[b]$H_0 : \Delta = 0$

Table 8 Proportion rejections in function of Γ

	Power					
	$J = 4$		$J = 8$		$J = 10$	
Γ	GLMM[a]	LMM[b]	GLMM[a]	LMM[b]	GLMM[a]	LMM[b]
−0.0822	.13	.04	.11	.05	.11	.05
−0.045	.10	.03	.09	.04	.08	.04
1.1706	.99	.43	1.00	.89	1.00	.91
1.2077	1.00	.56	1.00	.93	1.00	.95

[a]$H_0 : \Gamma = 0$

[b]$H_0 : \Delta = 0$

Note. The p values were obtained through an approximate Wald F-test with Satterthwaite denominator degrees of freedom (LMM) and an approximate Wald Z-test (GLMM) with $\alpha = .05$

Table 9 Proportion rejections in function of Δ

	Power					
	$J = 4$		$J = 8$		$J = 10$	
Δ	GLMM[a]	LMM[b]	GLMM[a]	LMM[b]	GLMM[a]	LMM[b]
0.0921	.06	.02	.06	.03	.06	.03
0.1	.07	.02	.07	.03	.07	.03
0.1713	.09	.03	.07	.04	.07	.04
0.1842	.08	.03	.07	.04	.07	.04
0.1967	.06	.02	.05	.03	.05	.04
0.2	.09	.03	.09	.04	.08	.04
0.2799	.08	.02	.06	.03	.06	.04
0.3427	.11	.03	.09	.04	.08	.04
0.3935	.08	.02	.07	.04	.07	.04
0.5598	.11	.03	.08	.04	.08	.05
0.9212	.15	.04	.10	.04	.10	.05
1	.17	.05	.13	.05	.12	.05
1.7135	.17	.04	.11	.05	.10	.05
1.9673	.15	.04	.10	.04	.09	.05
2.7992	.16	.04	.11	.05	.09	.05
5	.98	.46	1.00	.87	1.00	.90
5.35	.98	.38	1.00	.87	1.00	.91
10	1.00	.54	1.00	.92	1.00	.93
10.7	.99	.44	1.00	.90	1.00	.93
50	1.00	.63	1.00	.95	1.00	.96
53.5	1.00	.50	1.00	.93	1.00	.96

[a]$H_0 : \Gamma = 0$

[b]$H_0 : \Delta = 0$

Note. The p values were obtained through an approximate Wald F-test with Satterthwaite denominator degrees of freedom (LMM) and an approximate Wald Z-test (GLMM) with $\alpha = .05$

level, we also consider the number of cases J as a factor when analyzing the power of Γ in Table 8 and of Δ in Table 9. Our choice of J rather than I (the number of measurements) is based on the higher η^2 value in Table 5 ($\eta^2 = .0137$ for J versus $\eta^2 = .0003$ for I). From Tables 8 and 9, we can indeed see that the power increases as the sample size increases. For large (absolute) values of Δ and Γ, the power approaches 1. Comparing the two models, we see that the power of the GLMM is consistently higher than that of the LMM for both Γ and Δ. However, the power of the LMM reaches the commonly accepted 80% threshold (Cohen, 1988) when $J \geq 8$ in conditions where the effect is large ($\Delta \geq 5$ or $|\Gamma| \geq 1.1706$). When $J = 4$ the power of the LMM stays below .63 (Δ) or .56 (Γ) even for the largest effects.

Discussion

With this simulation study, we wanted to see whether the GLMM consistently outperforms the LMM, and, if not, in which cases the LMM has an acceptable performance. Three aspects of both models have been considered to assess their performance: goodness of fit, fixed effect parameter recovery, and inference.

In terms of goodness of fit, the LMM does in general not perform as well as the GLMM. In Fig. 1, a vast majority of the S_{AIC} scores lies above 0, indicating that the AIC of the GLMM is generally lower than the AIC of the LMM according to Eq. 11. Only when the baseline and treatment average responses are relatively high and when the number of cases is very small ($J = 4$) does the LMM achieve a goodness of fit comparable to that of the GLMM. In conditions with very sparse information, the more complex GLMM has a disadvantage compared to the LMM. Additionally, the LMM has the advantage that the baseline and treatment phase averages of the underlying count data are high and that the LMM therefore provides a good normal approximation of the data.

To assess the performance of both models in terms of fixed effect parameter recovery, we compared their parameter estimators $\hat{\Delta}_G$ vs. $\hat{\Delta}_L$ and $\hat{\Gamma}_G$ vs. $\hat{\Gamma}_L$. The most important measure of quality of an estimator is the MSE because it encompasses both the bias and the variance. A qualitative estimator should have an MSE as small as possible, i.e., a bias of zero and a small variance. From Table 6 and from Figs. 5 and 6 it is clear that the MSEs of the estimators of both models are on average very alike, with a slight advantage for the model which can directly estimate the parameter (i.e., the LMM for Δ and the GLMM for Γ).

In terms of inference, the first step in comparing the performance of the LMM with the GLMM is to look at the type I error rate. As seen in Table 7, the type I error rate of the LMM is better under control than the rate when using the GLMM. Although this might seem surprising, similar good behavior of less complex albeit misspecified (generalized) linear mixed models on small sample data has been observed (Bell et al., 2014). The more complex models, even though theoretically better fit to model the data, might function poorly when making too many estimates from too few pieces of information (Muth et al., 2016). Since the type I error rate of the LMM is under control, the next step is to look at its power. From Tables 8 and 9 it is clear that the LMM does not obtain the same power as the GLMM, not even for large effects. Only when the effect size and the number of cases J are large ($\Delta \geq 5$ or $\Gamma \geq 1.1706$, and $J \geq 8$) does the power of the LMM reach a level of 80%. This was true for all values of I (the number of measurements) considered in our simulation.

For applied research, a crucial next question is when is it acceptable to use an LMM to analyze single-case count data? In terms of goodness of fit, the LMM only yields acceptable AICs (i.e., AICs as low or lower than those of the GLMM) if the count data are well approximated by a normal distribution in both the baseline and the treatment phase and if the sample size (and especially the number of cases J) is very small. However, even in those conditions the LMM obtains a goodness of fit that is only 10% worse than that of the GLMM (Fig. 1). If this is considered acceptable, we recommend using the LMM in situations where the estimated effect size and the number of cases are reasonably large ($J \geq 8$), to ensure an acceptable power and unbiased fixed effect estimates.

When it comes to selecting an effect size to express the fixed effect, applied researchers need to determine whether they have a specific interest in either the additive effect expressed by Δ or the effect expressed by Γ. It makes sense to opt for the additive effect as expressed by Δ because it is more easily interpretable. Moreover, its estimate $\hat{\Delta}$ is readily available from the applied LMM as it does not need any transformation. Since this simulation study has provided some quantitative evidence of the good performance of the Δ_L estimator in terms of relative bias and relative MSE, the use of the LMM to model single-case count data to obtain an estimate for Δ would not be discouraged, even though it is an overly simplified model. Inference based on Δ_L is valid, because the type I error rate of Δ_L is under control and behaves well in all conditions. Again, caution is advised when doing inference based on Δ_L if the effect size or the number of cases is small, since then the power might not be acceptable.

When practitioners want to estimate the effect size Γ, it is preferable to use the GLMM to avoid the manipulations required to get the Γ_L estimate from the LMM (as illustrated in Appendix B). The GLMM will result in a slightly higher bias for Γ, but a lower MSE, compared to modeling a

LMM and estimating Δ. Even when using the GLMM to estimate Γ, however, there might be up to 10% relative bias in the estimates for some conditions, particularly when the effect size is not large and the amount of data for estimation is limited. When the sample sizes increase (i.e., the measurement series gets longer and the number of participants increases), this bias disappears.

If practitioners decide to model SCED count data using the GLMM with a Poisson distribution (Eq. 1), they need to be aware of the assumptions associated with the Poisson distribution (Winkelmann, 2008). First of all, the length of time intervals or session during which the counts are measured has to be the same across the entire time series. Applied researchers might do this already intuitively to make counts comparable over sessions, or based on good practices recommended by single-case handbooks (Ayres & Gast, 2010). In case the time series includes sessions of different lengths, the GLMM (1) can be adjusted to account for this by including an offset (Casals et al., 2015). As such, the outcome modeled is a rate rather than a count.

Another assumption to be taken into account when modeling a Poisson distribution is that the rate of occurrence across each time interval or session has to be constant; that is, the probability of occurrence of the measured event should be constant throughout each time interval. This assumption might be violated when an observed participant is disturbed by an external event or factor during a measuring session and when this disturbance has a temporary impact on the measured outcome. For example, when measuring problem behavior in a classroom environment, an observed participant might show temporarily increased problem behavior when a classmate initiates a fight with the participant during a measuring session. To lessen the likelihood of external factors impacting the rate of occurrence, practitioners can try to keep the length of measuring sessions short.

A final assumption of the Poisson distribution is that the events occurring in different time intervals should be independent. This assumption is violated when autocorrelation is present in the data. Practitioners can try to avoid this from happening by making sure their measuring sessions are far enough apart in time.

The results presented in this study have limitations inherent to all simulation studies, i.e., they are conditional on the simulation design and parameter values used. Because this study is the first of its kind, we have used the most basic GLMM design to simulate data and all simulation conditions were exclusively based on sample size and nominal parameter values. Naturally, there is much more to a GLMM design than these two aspects, and the many GLMM design extensions could all provide starting points for further exploration of the impact of model misspecifications for count data. These extensions include: (1) using alternative probability distributions to sample the dependent variable from, such as the binomial or other discrete distributions (to model discrete proportions or percentages), zero-inflated distributions and distributions fit for over-dispersed data; (2) specifying a specific covariance structure and as such modeling autocorrelation, rather than using an unstructured covariance matrix like in this study; (3) simulating data with variable exposure (i.e., the frequency of the behavior of interest is not tallied across the same period of time at each measurement occasion); (4) including linear or non-linear trends in the simulated data and in the fitted models; (5) using different single-case design types, e.g., alternating treatments or phase changes with reversal, rather than the multiple baseline AB design used in this study; and (6) simulating unbalanced data.

We focused mainly on the average treatment effect when comparing the results of the LMM and the GLMM estimations. This is in line with common practice, where applied researchers who are combining SCED count data are usually primarily interested in the average treatment effects (as expressed by Γ and Δ in this study), rather than in the individual treatment effects or the variance components. Moreover, just like average treatment effect estimations, individual effect and variance component estimations are not comparable between the LMM and the GLMM. Attempting to compare them would involve a similar and arguably even more complex method of transformation as illustrated for Γ and Δ. This is beyond the scope of this study.

Finally, we want to point out that inference results of the GLMM are based on an approximate Wald Z-test, which is likely to misspecify the sampling distribution of the Wald statistic as normal, especially in small samples. As explained in Appendix B, this was due to a lack of available procedures in the `lme4` package in R. In SAS, the PROC GLIMMIX procedure does include the option to set different degrees of freedom approximations to adjust for small sample sizes. It would be very useful to reanalyze our simulated datasets in SAS to see whether the inference results lead to substantially different conclusions from the conclusions we drew based on the R p values.

Conclusions

This simulation study showed that the GLMM in general does not substantially outperform the LMM, except in terms of the goodness of fit criteria. For the small sample sizes that we have considered, and which are common to SCED count datasets, we have found that the LMM does equally well as the GLMM in terms of fixed effect parameter recovery. In terms of inference, the type I error rates of the LMM are more under control than those of the GLMM. The power

of the LMM is generally lower than the power of the GLMM, but the LMM might provide acceptable power for SCED samples with a sufficient number of cases. This simulation provided some evidence that the GLMM might not necessarily be the better choice when it comes to very sparse SCED count data due to the model being too complex to estimate. Evidence for relatively better performance of the LMM if the expected count responses in baseline and/or treatment phases are relatively high was not so clear. Based on our results, we have provided some guidelines for applied researchers. Reviewers or meta-analysts using mixed modeling to combine SCED studies should be well aware of the effects of misspecifying their mixed model for discrete data. Their model choice should be well considered based on the type of raw data included and on the sample sizes.

Appendix A

Derivation of the Γ_G statistic

For the GLMM, Eqs. 5, 10 and 7 can be used to define a Γ_G statistic because they express E (B), E (Tx) and Var (B) respectively. Additionally, we need an expression for Var (Tx), which will be derived first before going on to define a Γ_G. Recall that Var (Tx) equals $\mathrm{Var}\left[\exp\left(\beta_{0j}\right)\exp\left(\beta_{1j}\right)\right]$.

$$\begin{aligned}\mathrm{Var}\left[\exp\left(\beta_{0j}\right)\exp\left(\beta_{1j}\right)\right] &= \mathrm{Cov}\left[\exp\left(\beta_{0j}\right)^2,\exp\left(\beta_{1j}\right)^2\right]\\ &+\left\{\mathrm{Var}\left[\exp\left(\beta_{0j}\right)\right]+\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]^2\right\}\\ &\times\left\{\mathrm{Var}\left[\exp\left(\beta_{1j}\right)\right]+\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2\right\}\\ &-\left\{\mathrm{Cov}\left[\exp\left(\beta_{0j}\right),\exp\left(\beta_{1j}\right)\right]\right.\\ &\left.+\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]\right\}^2\end{aligned} \tag{23}$$

The first step is to expand the first covariance term on the right-hand side by using the fact that $\left[\exp\left(2\beta_{0j}\right),\exp\left(2\beta_{1j}\right)\right]^{\mathsf{T}}$ also follows a lognormal distribution.

$$\begin{aligned}&\mathrm{Cov}\left[\exp\left(\beta_{0j}\right)^2,\exp\left(\beta_{1j}\right)^2\right]\\ &=\mathrm{Cov}\left[\exp\left(2\beta_{0j}\right),\exp\left(2\beta_{1j}\right)\right]\\ &=\mathrm{E}\left[\exp\left(2\beta_{0j}\right)\right]\mathrm{E}\left[\exp\left(2\beta_{1j}\right)\right]\left[\exp\left(4\sigma_{u01}\right)-1\right]\\ &=\left\{\mathrm{Var}\left[\exp\left(\beta_{0j}\right)\right]+\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]^2\right\}\\ &\times\left\{\mathrm{Var}\left[\exp\left(\beta_{1j}\right)\right]+\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2\right\}\\ &\times\left[\exp\left(4\sigma_{u01}\right)-1\right]\end{aligned} \tag{24}$$

The third term on the right-hand side of Eq. 23 can be expanded as follows based on Eq. 9.

$$\begin{aligned}&\mathrm{Cov}\left[\exp\left(\beta_{0j}\right),\exp\left(\beta_{1j}\right)\right]+\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]\\ &=\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]\left[\exp\left(\sigma_{u01}\right)-1\right]\\ &\quad+\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]\\ &=\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]\exp\left(\sigma_{u01}\right)\end{aligned} \tag{25}$$

Substituting Eqs. 24 and 25 into Eq. 23, one obtains:

$$\begin{aligned}&\mathrm{Var}\left[\exp\left(\beta_{0j}\right)\exp\left(\beta_{1j}\right)\right]\\ &=\left\{\mathrm{Var}\left[\exp\left(\beta_{0j}\right)\right]+\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]^2\right\}\\ &\quad\times\left\{\mathrm{Var}\left[\exp\left(\beta_{1j}\right)\right]+\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2\right\}\\ &\quad\times\exp\left(4\sigma_{u01}\right)\\ &\quad-\mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]^2\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2\exp\left(2\sigma_{u01}\right)\end{aligned}$$

All the required expressions for deriving the Γ_G statistic are now available and they are summarized below:

$$\begin{aligned}\mathrm{E}\left(\mathrm{B}\right) &= \mathrm{E}\left[\exp\left(\beta_{0j}\right)\right]\\ \mathrm{Var}\left(\mathrm{B}\right) &= \mathrm{E}\left(\mathrm{B}\right)^2\left[\exp\left(\sigma_{u0}^2\right)-1\right]\\ \mathrm{E}\left(\mathrm{Tx}\right) &= \mathrm{E}\left(\mathrm{B}\right)\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]\exp\left(\sigma_{u01}\right)\\ \mathrm{Var}\left(\mathrm{Tx}\right) &= \left[\mathrm{Var}\left(\mathrm{B}\right)+\mathrm{E}\left(\mathrm{B}\right)^2\right]\left\{\mathrm{Var}\left[\exp\left(\beta_{1j}\right)\right]\right.\\ &\quad\left.+\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2\right\}\exp\left(4\sigma_{u01}\right)-\mathrm{E}\left(\mathrm{Tx}\right)^2\end{aligned}$$

When substituting the above expressions into Eq. 18 for Γ, one obtains:

$$\begin{aligned}\Gamma_G &= \log\left\{\left[\frac{\mathrm{E}\left(\mathrm{Tx}\right)}{\mathrm{E}\left(\mathrm{B}\right)}\right]^2\sqrt{\frac{\mathrm{E}\left(\mathrm{B}\right)^2+\mathrm{Var}\left(\mathrm{B}\right)}{\mathrm{E}\left(\mathrm{Tx}\right)^2+\mathrm{Var}\left(\mathrm{Tx}\right)}}\right\}\\ &= \log\left\{\frac{\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2\exp\left(2\sigma_{u01}\right)}{\sqrt{\left\{\mathrm{Var}\left[\exp\left(\beta_{1j}\right)\right]+\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2\right\}\exp\left(4\sigma_{u01}\right)}}\right\}\\ &= \log\left\{\frac{\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2}{\sqrt{\mathrm{Var}\left[\exp\left(\beta_{1j}\right)\right]+\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2}}\right\}\\ &= \log\left\{\frac{\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2}{\sqrt{\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2\left[\exp\left(\sigma_{u1}^2\right)-1\right]+\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]^2}}\right\}\\ &= \log\left\{\frac{\mathrm{E}\left[\exp\left(\beta_{1j}\right)\right]}{\sqrt{\exp\left(\sigma_{u1}^2\right)}}\right\}\\ &= \log\left[\frac{\exp\left(\gamma_{10}+\frac{\sigma_{u1}^2}{2}\right)}{\exp\left(\frac{\sigma_{u1}^2}{2}\right)}\right]\\ &= \log\left[\exp\left(\gamma_{10}+\frac{\sigma_{u1}^2}{2}-\frac{\sigma_{u1}^2}{2}\right)\right]\\ &= \gamma_{17}\end{aligned} \tag{26}$$

Appendix B

R code samples

The SCED count data are stored in an R data frame `my_data`, with columns i indicating the measurement, j indicating the case, D indicating the phase (0 for baseline and 1 for treatment) and Y containing the dependent variable's outcomes. We use the `lmer()` and `glmer()` functions from the `lme4` package. The model is specified by a `formula` object as used in other regression functions in R, with the addition of a definition of the random part.

```
library(lme4)

LMM = lmer(Y ~ 1 + D + (1 + D | j), data = my_data)

GLMM = glmer(Y ~ 1 + D + (1 + D | j), family = "poisson", data = my_data)
```

The GLMM (Eq. 1) and the LMM (Eq. 2) have two coefficients: an intercept (denoted as 1 in the R formula) and the phase variable D. The expression `1 + D` defines the fixed part of the model: the expected value for Y is modeled by a mean intercept (`1`) and a mean treatment effect (`D`). The expression `(1 + D | j)` states that both the intercept and the treatment effect can vary randomly over cases. As part of the analysis of the results we first calculate the AIC and BIC of the fitted model objects `LMM` and `GLMM`:

```
AIC_LMM = AIC(LMM)

AIC_GLMM = AIC(GLMM)

BIC_GLMM = BIC(LMM)

BIC_GLMM = BIC(GLMM)
```

Next, we retrieve the fixed effect estimations $\hat{\gamma}_{00}$ and $\hat{\gamma}_{10}$ from the GLMM and $\hat{\gamma}^*_{00}$ and $\hat{\gamma}^*_{10}$ from the LMM:

```
gamma_00_LMM = as.numeric(fixef(LMM)[1])

gamma_10_LMM = as.numeric(fixef(LMM)[2])

gamma_00_GLMM = as.numeric(fixef(GLMM)[1])

gamma_10_GLMM = as.numeric(fixef(GLMM)[2])
```

The variance component estimations ($\hat{\sigma}_{u0}$, $\hat{\sigma}_{u1}$ and $\hat{\sigma}_{u01}$ from the GLMM and $\hat{\sigma}^*_{u0}$, $\hat{\sigma}^*_{u1}$ and $\hat{\sigma}^*_{u01}$ from the LMM) are retrieved as elements from the estimated covariance matrix Σ:

```
sigma_LMM = as.data.frame(VarCorr(LMM))

sigma_GLMM = as.data.frame(VarCorr(GLMM))


sigma_u0_LMM = sigma_LMM$vcov[1]

sigma_u1_LMM = sigma_LMM$vcov[2]

sigma_u01_LMM = sigma_LMM$vcov[3]

sigma_u0_GLMM = sigma_GLMM$vcov[1]

sigma_u1_GLMM = sigma_GLMM$vcov[2]

sigma_u01_GLMM = sigma_GLMM$vcov[3]
```

With the fixed and random effect parameter estimations we can calculate $\hat{\Delta}_G$ (15), $\hat{\Delta}_L$ (17), $\hat{\Gamma}_G$ (20) and $\hat{\Gamma}_L$ (22).

```
Delta_LMM = gamma_10_LMM
Delta_GLMM =
  exp(gamma_00_GLMM + ((sigma_u0_GLMM^2)/2))*
  (exp(gamma_10_GLMM + ((sigma_u1_GLMM^2)/2) + sigma_u01_GLMM) - 1)
Gamma_LMM =
  log(
    (((gamma_00_LMM + gamma_10_LMM)/gamma_00_LMM)^2)*
      sqrt(
        (gamma_00_LMM^2 + sigma_u0_LMM^2)/
        (
          (gamma_00_LMM + gamma_10_LMM)^2 +
          (sigma_u0_LMM^2) + (sigma_u1_LMM^2) + 2*sigma_u01_LMM
        )
      )
    )
Gamma_GLMM = gamma_10_GLMM
```

Finally, we retrieve the p values for the fixed effects. The `lme4` package does not by default include small sample adjustments for (G)LMMs like the Satterthwaite method of approximating the degrees of freedom. For the LMM, the `lmerTest` package (Kuznetsova et al., 2017) provides the `calcSatterth` function which returns p values based on an approximate Wald F-test with Satterthwaite degrees of freedom. For GLMMs, no R package providing small sample adjustments is currently available and therefore we rely on the approximate Wald Z-test. The associated p values are readily available when applying the `summary` method on the `glmerMod` object. The `coef` method returns a matrix with rows for each fixed coefficient and columns containing the estimates, standard errors, z values, and p values.

```
library("lmerTest")
p_LMM = as.numeric(calcSatterth(LMM,matrix(c(0,1),1,2))$pvalue)
p_GLMM = coef(summary(GLMM))[2,4]
```

References

Abad, A. A., Litière, S., & Molenberghs, G. (2010). Testing for misspecification in generalized linear mixed models. *Biostatistics*, *11*(4), 771–786. https://doi.org/10.1093/biostatistics/kxq019.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Parzen, E., Tanabe, K., & Kitagawa, G. (Eds.) *Selected Papers of Hirotogu Akaike*. https://doi.org/10.1007/978-1-4612-1694-0_15, (pp. 199–213). New York: Springer.

Ayres, K., & Gast, D. L. (2010). Dependent measures and measurement procedures. In Gast, D. L. (Ed.) *Single subject research methodology in behavioral sciences*, (pp. 129–165). New York: Routledge.

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

Bates, D., et al. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.10.18637/jss.v067.i01.

Bell, B. A., et al. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology*, *10*(1), 1–11. https://doi.org/10.1027/1614-2241/a000062.

Capanu, M., Gönen, M., & Begg, C. B. (2013). An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Statistics in Medicine*, *32*(26), 4550–4566. https://doi.org/10.1002/sim.5866.

Casals, M., et al. (2015). Parameter estimation of Poisson generalized linear mixed models based on three different statistical principles: A simulation study. *Statistics and Operations Research Transactions*, *39*(2), 281–308.

Claeskens, G., & Jansen, M. (2015). Model selection and model averaging. In *International encyclopedia of the social & behavioral sciences*. (2nd ed., pp. 647–652). Oxford: Elsevier.

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

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.10.3758/BRM.42.4.930.

Gelman, A., & Hill, J. (2009). *Data analysis using regression and multilevel/hierarchical models*, (11th ed.). Cambridge: Cambridge University Press.

Gill, J. (2001). *Generalized linear models: A unified approach*. Thousand Oaks (CA): Sage Publications.

Gumedze, F. N., & Dunne, T. T. (2011). Parameter estimation and inference in the linear mixed model. *Linear Algebra and its Applications*, *435*(8), 1920–1944. https://doi.org/10.1016/j.laa.2011.04.015.

Hembry, I., et al. (2015). Estimation of a nonlinear intervention phase trajectory for multiple-baseline design data. *Journal of Experimental Education*, *83*(4), 514–546. https://doi.org/10.1080/00220973.2014.907231.

Hox, J. J. (2010). *Multilevel analysis: techniques and applications*, (2nd ed.). New York: Routledge.

Jamshidi, L., et al. (2017). Review of single-subject experimental design meta-analyses and reviews: 1985-2015. Manuscript submitted for publication.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. New York: Springer.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions*, (3rd ed.)., p. 646). New York: Wiley. https://doi.org/10.1002/0471715816.

Kazdin, A. E., & Kopel, S. A. (1975). On resolving ambiguities of the multiple-baseline design: Problems and recommendations. *Behavior Therapy*, *6*(5), 601–608. https://doi.org/10.1016/S0005-7894(75)80181-X.

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.

Li, P., & Redden, D. T. (2015). Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology*, *15*(1), 38. https://doi.org/10.1186/s12874-015-0026-x.

McCullagh, P., & Nelder, J. A. (1999). *Generalized linear models*, (2nd ed.). London: Chapman and Hall.

Moeyaert, M., et al. (2013). 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.

Moeyaert, M. et al. (2014). Three-level analysis of single-case experimental data: Empirical validation. *The Journal of Experimental Education*, *82*(1), 1–21. https://doi.org/10.1080/00220973.2012.745470.

Molenberghs, G., Renard, D., & Verbeke, G. (2002). A review of generalized linear mixed models. *Journal de la Société Française de statistique*, *143*(1), 53–78.

Muth, C., et al. (2016). Alternative models for small samples in psychological research: Applying linear mixed effects models and generalized estimating equations to repeated measures data. *Educational and Psychological Measurement*, *76*(1), 64–87. https://doi.org/10.1177/0013164415580432. http://epm.sagepub.com/cgi/doi/10.1177/0013164415580432.

Nelson, K. P., & Leroux, B. G. (2008). Properties and comparison of estimation methods in a log-linear generalized linear mixed model. *Journal of Statistical Computation and Simulation*, *78*(3), 367–384. https://doi.org/10.1080/10629360601023599.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *The Clinical Journal of Pain*, *21*(1), 56–68. https://doi.org/10.1097/00002508-200501000-00007.

R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, (2nd ed.)., p. 485). Thousand Oaks: Sage Publications.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110–114.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

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*(3), 385–405. https://doi.org/10.1037/a0032964.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment & Intervention*, *2*(3), 188–196. https://doi.org/10.1080/17489530802581603.

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.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*, (2nd ed.). London: Sage Publications.

Stroup, W.alter.W. (2013). *Generalized linear mixed models: Modern concepts, methods and applications*. Boca Raton: CRC Press.

Swanson, H., & Sachse-Lee, C. (2000). A meta-analysis of single-subject-design intervention research for students with LD. *Journal of Learning Disabilities*, *33*(2), 114–136.

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. (2007). The aggregation of single-case results using hierarchical linear models. *Behavior Analyst Today*, *8*(2), 52–57. https://doi.org/10.1037/h0100613.

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.

Winkelmann, R. (2008). *Econometric analysis of count data*, (4th ed.). Berlin: Springer. isbn: 9783540776482.

Yau, K.elvin.K. W., & Kuk, A.nthony.Y. C. (2002). Robust estimation in generalized linear mixed models. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *64*(1), 101–117.

Zhou, X.-H., Perkins, A. J., & Hui, S. L. (1999). Comparisons of software packages for generalized linear multilevel models. *The American Statistician*, *53*(3), 282–290. https://doi.org/10.2307/2686112.