Between-case Incidence Rate Raito for Single Case Experimental Designs with Count Outcomes: A Monte Carlo Simulation with Conditional and Marginal Models

Haoran Li
University of Minnesota
haoranli@umn.edu
Chendong Li
Texas A&M University
cliattx@tamu.edu
Wen Luo
Texas A&M University
wluo@tamu.edu
Eunkyeng Baek
Texas A&M University
baek@tamu.edu

Author Note

Correspondence concerning this article should be addressed to Dr. Haoran Li, Department of Educational Psychology, University of Minnesota, United States.

Email: haoranli@umn.edu

Authors' Contribution Statement

Haoran Li: Conceptualization (lead), funding acquisition (supporting), methodology (lead), supervision (lead), formal analysis (supporting), software (equal), visualization (equal), writing – original draft preparation (lead), writing – review & editing (lead);

Chendong Li: Conceptualization (supporting), formal analysis (lead), software (equal), visualization (equal), writing – original draft preparation (supporting), writing – review & editing (supporting);

Wen Luo: Conceptualization (supporting), funding acquisition (lead), project administration (lead), writing – original draft preparation (supporting), writing – review & editing (supporting);

Eunkyeng Baek: Conceptualization (supporting), funding acquisition (supporting), writing – original draft preparation (supporting), writing – review & editing (supporting);

Data Availability Statement

The data used for the demonstration and the R codes are available at https://osf.io/eqa4h/?view_only=98358eeef01f4d20b043d902e8d16c0a

Funding Statement

The research reported here is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant [R305D240019] to Texas A&M University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Conflict of Interest Disclosure

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abstract

Calculating design-comparable effect sizes in single-case experimental designs (SCEDs) is essential for research synthesis to identify evidence-based practices that integrate findings from both SCEDs and group-based designs. The field has made significant progress in developing additive design-comparable effect sizes in recent years. This study aims to evaluate the statistical properties of a newly developed proportional estimator, the between-case incidence rate ratio (BC-IRR), for SCEDs with count outcomes. Two analytical frameworks are introduced for estimating BC-IRR: generalized linear mixed models (GLMMs) and generalized estimating equations (GEEs). A large-scale Monte Carlo simulation is conducted to assess the bias of point estimate, the standard error bias, mean squared error, and coverage rate. We also demonstrate the estimation of BC-IRR using GLMMs and GEEs with real data. Based on the simulation and demonstration results, we provide recommendations for applied researchers, discuss limitations, and outline directions for future research.

Keywords: generalized linear mixed model, generalized estimation equation, single case experimental design, count data, incidence rate ratio, Monte Carlo simulation

Between-case Incidence Rate Raito for Single Case Experimental Designs with Count Outcomes: A Monte Carlo Simulation with Conditional and Marginal Models

Single-case experimental designs (SCED) are experimental designs for assessing the treatment effect of an intervention on a small number of cases by repeatedly measuring outcomes over time under multiple conditions manipulated by the experimenter (Kratochwill & Levin, 2010; Onghena, 2005). These designs involve the manipulation of treatment and control conditions over time and assess causal effects by comparing outcomes observed under treatment conditions to those observed under control conditions. Because the measurement procedure in SCEDs often rely on direct observations of behaviors, count data are very common, such as number of aggressive behaviors, number of words read correctly per minute, and number of digits correct in math practices. As the distributions of count data are usually non-normal and show heteroscedasticity due to the dependent relationship between mean and variance, traditional effect sizes that assume normality and homogeneity, such as the standardized mean differences (e.g., Cohen's d type effect sizes), might not be appropriate for count data. In addition, count data typically represent the number of events occurring over a specific period of time, and an incidence rate is computed by dividing the count by the units of time. Effect sizes for count data should remain the same no matter what unit is used as the divisor to determine the incidence rate. However, standardized mean differences are highly sensitive to the units used for the divisor (Pustejovsky, 2023; Wilson, 2022). In some cases, researchers fail to report session lengths, further complicating the effect sizes calculation for count data.

To address these issues, researchers have proposed alternative effect size measures for count outcomes in SCEDs, such as the log response ratio (LRR; Pustejovsky, 2015, 2018), the Bayesian rate ratio (BRR, Natesan Batley et al., 2021) and the incidence rate ratio (IRR; Li et al.,

2025) for count data in SCEDs, all of which quantify the magnitude of treatment effects as a proportional change between baseline and treatment phases within cases. Unlike the LRR and BRR, the IRR can simultaneously account for baseline trend, trend change, overdispersion, zero-inflation, and measurement dependencies due to autocorrelation and nested data structure at the same time through the flexible framework of generalized linear mixed models (GLMMs, Li et al., 2025; Li et al., 2024; Li, 2024).

Traditionally, the IRR and its variants, such as relative rate ratio and log response ratio, have been widely adopted and discussed in meta-analyses and research synthesis methods across disciplines such as medicine, epidemiology, and evolutionary biology (e.g., Hawken et al., 2016; Lane, 2013; Li et al., 2018; Spittal et al., 2015). In the behavior and social sciences, methodologists have introduced modeling approaches including generalized liner models (GLMs) and GLMMs to analyze count data, highlighting differences in the interpretation of proportional effect sizes like the IRR in GLMs/GLMMs versus additive effect sizes from linear (mixed) models (Aiken et al., 2015; Atkins et al., 2013; Coxe et al., 2013; Grimm & Stegmann, 2019). In the SCED context, Li et al. (2024) found that GLMMs can yield accurate estimates and reliable inferential statistics for treatment effects quantified by the IRR in multiple baseline designs (MBDs) when researchers make sounded statistical decisions regarding model selections.

Despite the wide use of IRR as a standard effect size measure for count outcomes in many fields and its favorable statistical properties in SCEDs, the IRR derived from SCEDs is a within-case IRR (WC-IRR), reflecting within-case contrasts. As a result, it is not appropriate to combine WC-IRR in research synthesis with the IRRs obtained from between-group designs, such as randomized control trials (RCTs). To address this issue, researchers have developed a between-case IRR (BC-IRR) for AB designs and MBDs (Luo et al., 2025). As a marginal effect size, BC-IRR is equivalent to the IRR that would be obtained in a potential RCT when certain assumptions are met. Specifically, it is assumed that a never-treated group exists that is comparable to the treated units in a SCED in terms of individual effects, and that the model for the potential outcome under the control condition is correctly specified. Therefore, BC-IRR is a design-comparable effect size that can be used in research synthesis of SCEDs and RCTs.

The previous study (Luo et al., 2025) demonstrated the equivalence of BC-IRR and IRR based on a potential RCT using a structural model and simulations of a pseudo-population. However, the study did not investigate the statistical properties of the BC-IRR estimator in the SCED context. Moreover, it considered only one approach to estimate BC-IRR through conditional model of GLMMs, where the calculation of BC-IRR is quite complex, involving conditional estimates of immediate effect and trend change, as well as associated variance and covariance components. In this study, we consider an alternative approach to estimate BC-IRR through marginal models. By definition, BC-IRR is a marginal effect, that is, a ratio of the average outcome level at a certain session in the treatment phase across cases to the counterfactual (i.e., if the intervention is not introduced) average outcome level at the same session (Luo et al., 2025). Hence, marginal models such as generalized estimating equations (GEEs) are well-suited (Liang & Zeger, 1986; Zeger & Liang, 1986) and it is worthwhile exploring GEE as an alternative to GLMMs for estimating BC-IRR.

The purpose of the current study is to systematically evaluate and compare the statistical properties of BC-IRR estimated via the GLMM and GEE and to demonstrate the methods using R with real data. The paper is organized into five main sections. The literature review section briefly introduces the difference between conditional and marginal estimates, and formal definition of BC-IRR. The modeling and estimation section outlines the GLMM and GEE frameworks, reviews their performance in small-sample conditions, and discusses small-sample correction methods. The third section presents a simulation study evaluating the performance of both modeling approaches, focusing on the bias of point estimate, bias of standard error, mean squared error, and coverage rate of BC-IRR. The fourth section demonstrates the application of these models using real data and R code. The final section concludes with a discussion of implications, practical recommendations, study limitations, and directions for future research.

Theory of Between-case IRR

Conditional and Marginal Estimates for IRR

The different interpretations between the WC-IRR and BC-IRR in the analysis of SCED count data origins from the statistical decision on whether it is more appropriate to use a conditional model or a marginal model to analyze correlated response data. An extensive and ongoing discussion exists in disciplines such as statistics, biology, ecology, and epidemiology regarding the differences, advantages, and limitations of these two modeling approaches (Akanda & Alpizar-Jara, 2014; Fieberg et al., 2009; Heagerty, 1999; Heagerty & Zeger, 2000; Lee & Nelder, 2004; Lindsey & Lambert, 1998; Muff et al., 2016). However, the conceptual and technical nuances of both models may remain opaque to researchers in the behavioral and social sciences.

To facilitate understanding, we begin with a general illustration of conditional and marginal estimates, followed by an example specific to SCEDs. We first consider a two-level linear mixed-effects model (LMM), which can be written as: $y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{u}_j + e_{ij}$, $\boldsymbol{u}_j \sim N(0, \boldsymbol{G})$, and $e_{ij} \sim N(0, \sigma_e^2)$, where y_{ij} denotes a continuous outcome for unit i in cluster j, $\boldsymbol{x}_{ij}$ denotes a vector of covariates and $\boldsymbol{\beta}$ is the associated vector of regression coefficients, $\boldsymbol{z}_{ij}$

denotes a vector including an intercept and a subset of the unit-level covariates in $\boldsymbol{x}_{ij}$ and $\boldsymbol{u}_j$ is the associated vector of cluster random effects, assumed multivariate normally distributed with zero mean vector and variance-covariance matrix $\boldsymbol{G}$, and e_{ij} is the residual, assumed normally distributed with zero mean and residual variance σ_e^2. A distinctive feature of the LMM is that it has simple expressions for both the conditional mean response for an individual given by $E(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_j) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{u}_j$, and the marginal mean response averaged over random effects $\boldsymbol{u}_j$ given by $E(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$. The same component $\boldsymbol{x}'_{ij}\boldsymbol{\beta}$ means that in LMMs, the fixed effects $\boldsymbol{\beta}$ have both subject-specific and population-average interpretations. For example, if $\boldsymbol{x}'_{ij}$ contains only a binary treatment indicator (0 = control, 1 = treatment), then $\boldsymbol{\beta}$ represents both (a) the conditional difference in outcomes between individuals with the same random effects (e.g., typical individual with $\boldsymbol{u}_j$= $\boldsymbol{0}$) and (b) the average treatment effect across the population. However, having both conditional and marginal interpretations is generally not the case for $\boldsymbol{\beta}$ when a non-linear function is adopted in the GLMM framework.

Consider the corresponding GLMM with a non-normal outcome y_{ij}: $y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_j \sim \text{Distr}\left(\lambda_{ij}, \text{A}(\lambda_{ij})\right), g(\lambda_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{u}_j$, where λ_{ij} represents the expected value of the outcome y_{ij} conditional on the predictors $\boldsymbol{x}'_{ij}$ and random effects $\boldsymbol{u}_j$, $\text{A}(\lambda_{ij})$ is the variance function and its form depends on the theoretical distribution (Distr), and $g(\cdot)$ is a non-linear link function (e.g., log or logit link function). In this model, the conditional mean response is $E(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, \boldsymbol{u}_j) = g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{u}_j)$, while the marginal mean response, averaged over $\boldsymbol{u}_j$, is $E(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) \neq g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta})$ and often lacks a closed-form solution. Thus, the fixed effects $\boldsymbol{\beta}$ in GLMMs only have a conditional interpretation, but do not have a marginal interpretation as in the LMM. However, for count data modeled with Poisson or negative binomial distributions, approximate or numerical methods could be used to derive marginal estimates based on GLMMs.

In SCEDs, consider the following Poisson GLMM for count outcomes:

$$Y_{ij} \sim Poisson\left(\lambda_{ij}\right)$$

$$\text{Level 1: } \log\left(\lambda_{ij}\right) = \beta_{0j} + \beta_{1j}Phase_{ij}$$

$$\text{Level 2: } \begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} + u_{1j} \end{cases}$$

$$\begin{bmatrix} \mu_{0j} \\ \mu_{1j} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad G = \begin{bmatrix} \sigma_{u0}^{2} & \sigma_{u0u1} \\ \sigma_{u1u0} & \sigma_{u1}^{2} \end{bmatrix}\right), \tag{1}$$

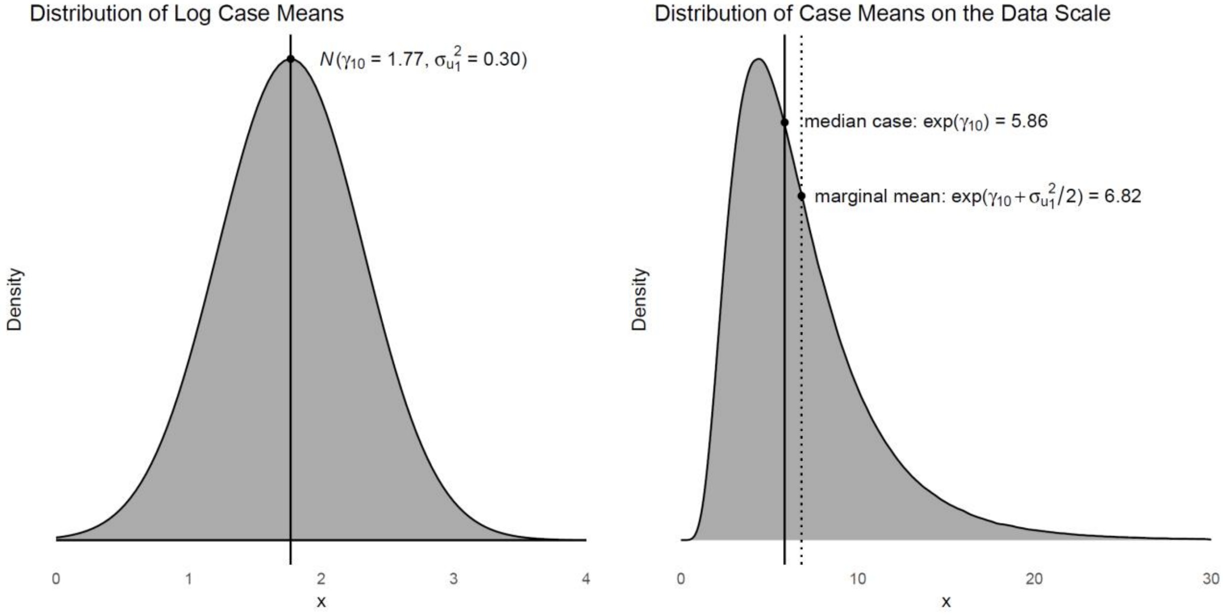
where i and j are the index of repeated measures and cases, respectively, and Phase is a binary variable (0 = baseline and 1 = treatment). The intercept γ_{00} represents the within-case bassline level and γ_{10} represents the within-case treatment effect in the log-count scale because the log link function is adopted. The effect size for treatment effect in the frequency count scale is assessed by the exp (γ_{10}), which is a with-case effect size (i.e., WC-IRR). The between-case variance components (σ_{u0}^{2} and σ_{u1}^{2}) are found on the diagonal of the variance-covariance matrix (i.e., the G matrix), representing the variance of the baseline level and the treatment effect between cases. The off-diagonal element (σ_{u0u1}) represents the covariance between baseline level and the treatment effect. To facilitate the understanding of the differences between conditional and marginal estimates, we first provide the expressions for the conditional and marginal expectations based on the Poisson GLMM in Table 1 (see Leckie et al., 2020).

For simplicity, the following illustration focuses on the conditional and marginal estimates for the treatment effect (γ_{10}). Assuming $\gamma_{10} = 1.77$ and $\sigma_{u1}^{2} = 0.30$. Hence, the conditional treatment effect for a *typical* case (i.e., a median case with $\boldsymbol{u_j} = \boldsymbol{0}$) measured by WC-

IRR is $\frac{\exp(\gamma_{00}+\gamma_{10})}{\exp(\gamma_{00})} = \exp(\gamma_{10}) = \exp(1.77) = 5.86$, whereas the marginal treatment effect for the *population* is $\frac{\exp(\gamma_{00}+\gamma_{10}+\sigma_{u0}^2/2+\sigma_{u1}^2/2)}{\exp(\gamma_{00}+\sigma_{u0}^2/2)} = \exp(\gamma_{10} + \sigma_{u1}^2/2) = \exp\left(1.77 + \frac{0.30}{2}\right) = 6.82$. A graphic illustration of the transformation between the log scale and the original data scale in terms of treatment effect estimates is depicted in Figure 1. It is noted that the marginal effect estimated from the GLMM is subject to accuracy for both fixed and random effects estimates. An alternative way to estimate the marginal effect like BC-IRR for SCED count outcomes is through direct marginal modeling. A full illustration of specifications for conditional and marginal models to estimate BC-IRR will be provided in the next section.

Figure 1

Graphical Illustration of Conditional and Marginal Estimates



Causal Definition of BC-IRR for SCEDs

The marginal treatment effect aforementioned for SCEDs could be formally defined by utilizing causal inference framework. Let us consider an AB design with *m* participants and series length of *n*. For simplicity, assume that all participants enter the treatment phase at time *T*.

This assumption is not required, as the definition also applies to multiple baseline designs in which participants enter the treatment phase at different time point. According to Luo et al. (2025), BC-IRR is defined as a causal estimand on a ratio scale:

$$\text{BC-IRR} = \frac{E[Y_B(1)]}{E[Y_B(0)]}. \quad (2)$$

The numerator represents the expected incidence rate at time *B* if a treatment strategy is implemented between times *T* and *B*. The denominator represents the expected incidence rate at time *B* if the treatment is never adopted. Both the numerator and the denominator are marginal means, representing the average incidence rates across the *m* participants under the treated and the untreated conditions. Thus, BC-IRR is a *marginal* effect.

Because the estimand is defined based on potential outcomes, it is necessary to link the potential outcomes to the observed data. The numerator can be directly linked to the observed incidence rate at time *B* because all participants receive the treatment. However, the denominator, the potential outcome under no treatment at time *B*, is unobserved and must be estimated using models based on observations from the baseline phase. A key identifiability condition, therefore, is that this potential outcome model is correctly specified. Additional details of the statistical models are presented in the next section. BC-IRR has been shown to be equivalent to the IRR that would have been obtained in a hypothetical RCT, under the assumption that a never-treated group exists in which the distribution of individual effects, defined as the time-invariant effects unique to each individual, matches that of the treated units in the SCED. Due to space limit, we do not repeat the detailed derivations here. Readers are referred to Luo et al. (2025) for the complete logical and mathematical arguments.

Estimation Models for BC-IRR

To estimate BC-IRR as defined in the previous section, we consider two methods, one derived from the conditional/subject-specific model of GLMM, the other based on the marginal/population-average model via the GEE approach.

Generalized Linear Mixed Model

GLMMs are a set of conditional models that integrate the frameworks of generalized linear models (GLM) and linear mixed models (LMM). Hence, GLMMs could take the source of clustering or nested data structure in SCEDs into account by directly modeling between-case variation using random effects. With an extension to generalized framework, GLMM could also deal with non-normal outcomes in SCEDs, such as binary, percentage/proportion, count, and rate data using various distributions (e.g., Poisson, negative binomial, binomial, beta-binomial).

We used the following Poisson GLMM to derive BC-IRR from count outcomes in an AB design:

$$\begin{aligned} &y_{ij} \sim \text{Poisson}\left(\lambda_{ij}\right) \\ &\log\left(\lambda_{ij}\right) = \beta_{0j} + \beta_{1j}Time_{ij} + \beta_{2j}Phase_{ij} + \beta_{3j}Time'_{ij}Phase_{ij} \end{aligned} \tag{3}$$

where $Phase_{ij}$ is an indicator variable equal to 0 for control phase and 1 for treatment phase; $Time_{ij}$ is a time variable representing the time (measured either in calendar units or sessions) elapsed from a reference time point. Researchers can choose any time point as a reference time point, which changes the meaning of the intercept β_{0j} and its associated variance-covariance components. $Time'_{ij}$ is another centered time variable which equals zero for the first session in the treatment phase. Based on this coding scheme, $e^{\beta_{0j}}$ represents the expected outcome level at the last session in baseline for case j, $e^{\beta_{1j}}$ represents the proportional change per session in the baseline phase, $e^{\beta_{2j}}$ represents the immediate proportional change in the outcome level after the treatment (i.e., the within-case IRR for immediate effect), and $e^{\beta_{3j}}$ represents the trend/slope change after the treatment (i.e., the within-case IRR for trend change). An example coding matrix for all the variables is shown in Table 2, where C is the reference time point, the treatment is implemented at time $T = A$, and the outcome Y is measured at time point B.

Based on a multivariate normal assumption for all random coefficients (i.e., $\beta_{0j} - \beta_{3j}$)

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \end{bmatrix} \sim \text{MVN} \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{20} \\ \gamma_{30} \end{bmatrix} \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u0u1} & \sigma_{u0u2} & \sigma_{u0u3} \\ \sigma_{u0u1} & \sigma_{u1}^2 & \sigma_{u1u2} & \sigma_{u1u3} \\ \sigma_{u0u2} & \sigma_{u1u2} & \sigma_{u2}^2 & \sigma_{u2u3} \\ \sigma_{u0u3} & \sigma_{u1u3} & \sigma_{u2u3} & \sigma_{u3}^2 \end{bmatrix} \tag{4}$$

, the expression of marginal exception in Table 1, and the coding matrix in Table 2

$$\text{BC-IRR } = \frac{E[Y_B(1)]}{E[Y_B(0)]} = e^{\gamma_{20} + \frac{1}{2}\sigma_{u2}^2 + \sigma_{u0u2} + (B-C)\sigma_{u1u2} + \gamma_{30}(B-\text{A}) + \frac{1}{2}(B-A)^2\sigma_{u3}^2 + (B-A)\sigma_{u0u3} + (B-A)\sigma_{u2u3} + (B-C)(B-A)\sigma_{u1u3}}, \tag{5}$$

where $B - C$ represents the time elapsed between the reference time point (C) and the outcome measurement occasion B, while $B - A$ represents the time elapsed between the introduction of treatment and the outcome measurement occasion. To compute BC-IRR, the estimates of the conditional effects (γ) and variance/covariance components (σ) from the GLMM are substituted into equation 5. The full derivation process for equation 5 can be found in Appendix A in Luo et al. (2025).

The Poisson GLMM in equation 3 assumes that the conditional mean and variance for the outcome are identical. However, when count data exhibit more variability than predicted by the Poisson model, overdispersion occurs. Failing to account for overdispersion leads to poor model fit and misleading statistical inferences due to underestimated standard errors of regression coefficients (Agresti, 2018; Hilbe, 2011, 2014; Li et al., 2025). Pustejovsky et al. (2023) found that overdispersion is not uncommon for count outcomes in SCEDs. To accommodate overdispersion, the following Poisson-Gamma mixture model, also called negative binomial model (NB), has been introduced to analyze overdispersed count data in SCEDs (Li et al., 2025;

Li et al., 2024).

$$y_{ij} \sim Poisson\ (\lambda_{ij})$$

$$\log(\lambda_{ij}) = \beta_{0j} + \beta_{1j}Time_{ij} + \beta_{2j}Phase_{ij} + \beta_{3j}Time'_{ij}Phase_{ij} + e_{ij} \qquad (6)$$

$$\exp(e_{ij}) \sim \text{Gamma}(\theta, 1/\theta), \theta > 0.$$

With the log link function and the same multivariate assumption for random coefficients β, the calculation for BC-IRR as shown in equation 5 is exactly the same for the NB model.

Marginal Model via Generalized Estimating Equation

Rather than explicitly modeling the clustering mechanism via random effects as is done with the conditional models of GLMM, marginal models treat within-subject association as a nuisance characteristic of data that should be accounted for to make correct inferences about changes in the population mean response (Liang & Zeger, 1986). A marginal model for longitudinal data has the following three-part specification (Fitzmaurice, 2012): (1) The marginal expectation $E(y_{ij}|\boldsymbol{x}_{ij}) = \mu_{ij}$, depends only on the covariates $\boldsymbol{x}_{ij}$, through a known link function $g(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$; (2) the variance of Y_{ij}, conditional on the covariates, is given by $Var(y_{ij}|\boldsymbol{x}_{ij}) = \phi v(\mu_{ij})$, where $v(\mu_{ij})$ is a known variance function of the conditional mean μ_{ij} and ϕ is a scale parameter; (3) the covariance matrix for the within-subject responses, conditional on the covariates, is given by $\boldsymbol{V}_j = \phi\sqrt{\boldsymbol{A}}_j\text{Corr}(\boldsymbol{Y}_j)\sqrt{\boldsymbol{A}_j}$, where $\boldsymbol{A}_j$ is a diagonal matrix with components $v(\mu_{ij})$, and $\text{Corr}(\boldsymbol{Y}_j)$ is the correlation matrix as the function of a vector of parameters $\boldsymbol{\alpha}$ that estimate the correlations of observations within subjects.

When the outcomes in the marginal model are discrete (e.g., binary, count, or ordinal categorical data), there is no convenient specification of the joint distribution of $\boldsymbol{Y}_j$. Hence, it requires an alternative to maximum likelihood estimation and leads to a method known as generalized estimation equations (GEE). GEE approach essentially views the model as a single-level model and apply statistical corrections to produce standard error estimates that account for clustering (Liang & Zeger, 1986; Zeger & Liang, 1986). The advantage of GEE is that the covariance structure of the outcome does not have to be correctly specified, meaning that there are far fewer assumptions required compared to random effects models (Zeger et al., 1988). For the estimation, GEE requires researchers to specify a working correlation matrix (i.e., $\mathrm{Corr}(\boldsymbol{Y}_j)$ and then start an iterative estimation process using weighted least squares (quasi-likelihood) by (1) estimating the regression coefficients ($\boldsymbol{\beta}$) as if data are independent, (2) estimating the parameters ($\boldsymbol{\alpha}$) in the working correlation matrix to handle clustering, (3) using the estimated working correlation matrix to update the outcome variable covariance matrix ($\boldsymbol{V}_j$), (4) using $\boldsymbol{V}_j$ to update estimates of $\boldsymbol{\beta}$; and (5) continuing this iteration process until convergence and then applying sandwich estimate estimators for stander errors. For an approachable illustration of the full estimation procedure for GEE, readers are referred to McNeish and Stapleton (2016).

In SCEDs, BC-IRR could be estimated by specifying a marginal model using the GEE approach with an appropriate working correlation matrix. The marginal model is specified as follows for count outcomes:

$$E\left(y_{ij}\middle|\boldsymbol{x}_{ij}\right) = \lambda_{ij}, \log\left(\lambda_{ij}\right) = \beta_0 + \beta_1 Time_{ij} + \beta_2 Phase_{ij} + \beta_3 Time'_{ij} Phase_{ij}. \tag{7}$$

$$Var\left(y_{ij}\middle|\boldsymbol{x}_{ij}\right) = \phi v\left(\lambda_{ij}\right)$$

$$\boldsymbol{V}_j = \phi \boldsymbol{A}_j^{\frac{1}{2}} \boldsymbol{R}_j(\boldsymbol{\alpha}) \boldsymbol{A}_j^{\frac{1}{2}}$$

where the Poisson variance function $v\left(\lambda_{ij}\right) = \lambda_{ij}$ is adopted and the scaler parameter ϕ could handle potential overdispersion in SCED count outcomes. The working correlation structure $\boldsymbol{R}_j(\boldsymbol{\alpha})$ describes the correlation pattern of observations within-subject with $\boldsymbol{\alpha}$ as a vector of association parameters depending on the correlation structure. In our study, we adopt the "independence" correlation structure to avoid convergence issues encountered with very small sample sizes in SCEDs. We note that other correlation structures including, but not limited to, "unstructured", "autoregressive", "Toeplitz", and "exchangeable" are available for the GEE approach. Based on equations (2) and (7),

$$\text{BC-IRR} = e^{(\beta_2+\beta_3(B-A))}, \tag{8}$$

where β_2 and β_3 represent the marginal estimates of the immediate effect and trend change, respectively. A key advantage of the GEE over GLMM is that it does not require specifying the level-2 variance-covariance structure of random effects. The GEE approach could provide robust estimates of standard errors if all relevant predictors are included, regardless of the proper specification of the working correlation matrix of the outcome measurements (McNeish & Harring, 2017).

Delta Methods for Standard Errors

As BC-IRR is essentially a combination or a transformation of parameter estimates related to the immediate effect and regarding trend change, we need to further calculate its standard error. As outlined in equations (5) and (8), the transformation is relatively complex for the GLMM and more intuitive for the GEE approach. To get standard errors for the transformed parameter, the delta method is typically used. The delta method is essentially calculating the variance of the Taylor series approximation of a function (G) of random variables ($\boldsymbol{x}$ = (x1, x2, ...)) with known variance. The first two terms of the Talor expansion are an approximation for the function G,

$$G(\boldsymbol{x}) \cong G(\boldsymbol{\mu}_x) + \left(\frac{d(\,G(\boldsymbol{\mu}_x))}{d(\boldsymbol{x})}\right)^T (\boldsymbol{x} - \boldsymbol{\mu}_x)\,, \tag{9}$$

where $\boldsymbol{\mu}_x$ is the mean vector of $\boldsymbol{x}$ and $\frac{d(\,G(\boldsymbol{\mu}_x))}{d(\boldsymbol{x})}$ is a vector of partial derivatives of $G(\boldsymbol{x})$ at point $\boldsymbol{\mu}_x$. We can then approximate the variance of $G(\boldsymbol{x})$,

$$\operatorname{var}\big(G(\boldsymbol{x})\big) \cong \left(\frac{d(\,G(\boldsymbol{\mu}_x))}{d(\boldsymbol{x})}\right)^T \operatorname{cov}(\boldsymbol{x}) \left(\frac{d(\,G(\boldsymbol{\mu}_x))}{d(\boldsymbol{x})}\right), \tag{10}$$

where $\operatorname{cov}(\boldsymbol{x})$ is the known variance-covariance matrix of $\boldsymbol{x}$. The square root of the outcome based on equation 10 is the standard error of the transformed parameter estimate. To calculate the standard errors of BC-IRR estimated by the GEE approach as shown in equation (8), $G(\boldsymbol{x}) = G(\boldsymbol{\beta}) = e^{(\beta_2+\beta_3(B-A))}$ and $\operatorname{cov}(\boldsymbol{x}) = \operatorname{cov}(\boldsymbol{\beta}) = \operatorname{cov}(\beta_2,\ \beta_3)$. Similarly, to obtain the standard error of BC-IRR estimated by the GLMM as shown in equation (5), $G(\boldsymbol{x}) = G(\boldsymbol{\gamma}, \boldsymbol{\sigma})$ and $\operatorname{cov}(\boldsymbol{x}) = \operatorname{cov}(\boldsymbol{\gamma}, \boldsymbol{\sigma})$. The calculations via the delta method can be conducted using the *deltamethod* function in the R package *msm* (Jackson, 2011).

Performance of GLMM with Small Sample Size

Recent studies have systematically evaluated the performance of GLMMs in the SCED context (Declercq et al. 2019; Li et al., 2025; Li et al., 2024). They have demonstrated that GLMMs are promising methods to deal with overdispersed and zero-inflated SCED count data. These studies also provide tailored guidelines for model selection based on the properties of SCED count data. For example, Declercq et al. (2019) found that when count data are not overdispersed, Poisson GLMMs can produce accurate estimates of treatment effects evaluated by WC-IRR. In recent simulations studies, Li et al. (2024) also found that GLMMs can provide accurate estimates and reliable inferential statistics for treatment effects measured by WC-IRR in multiple baseline designs, when researchers select appropriate distributions (e.g., Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial). In addition, there are few studies directly comparing the performance of GLMM and GEE under small sample size in more general contexts (e.g., McNeish & Stapleton, 2016; McNeish & Harring, 2017). The authors found that GLMM outperforms GEE in terms of standard error bias, coverage rate, and statistical power for both continuous and binary outcomes. More specifically, for both continuous and binary outcomes with a small number of clustering units (≤ 50), GLMM with degrees of freedom adjusted method (i.e., Kenward-Roger adjustment) showed very minimum bias in the standard error estimates of fixed effects and had the coverage rate close to the nominal level of 95%.

Performance of GEE with Small Sample Size

Unlike solid evidence provided for the statistical properties of GLMMs, there have been no direct studies of GEE applying to SCED context. However, the performance of GEE for correlated binary and continuous data has been well documented. In the large sample framework, it was known that estimates of regression parameters of GEE are generally unbiased regardless of the working correlation matrix, but the misspecifications of the working correlation matrix led to biased estimates of the standard errors of the regression estimates. Thus, it is typical practice to obtain standard error based on the robust or sandwich estimator proposed by Liang and Zeger (1986) for the variance-covariance matrix of regression coefficients (Zorn, 2001), which is consistent even under misspecification of the working correlation matrix.

With small sample size, previous research found that GEE does not perform well (McNeish & Stapleton, 2016; McNeish & Harring, 2017; Thompson et al., 2021). Specifically, McNeish and Stapleton (2016) found that GEEs had relatively large bias for fixed effects estimates regarding level-2 predictors and cross-level interactions when dealing with a very small number of clustering units (e.g., n = 4). The 95% confidence interval based on the classical GEE with the sandwich estimator for standard errors often produced significant under-coverage across different levels of number of units (n = 4 to 14).

To enhance both parameter estimates and standard errors of the estimates under small sample size, several bias correction methods have been proposed (Paul & Zhang, 2014, Thompson et al., 2021). For example, Paul and Zhang (2014)'s study demonstrated that by using bias adjusted approaches, unbiased parameter estimates can be obtained for small to moderate sample size (n = 15 to 50). Some other researchers have proposed various corrections for the sandwich estimator to obtain accurate statistical inferences for the GEE regression parameter estimators under small sample sizes (e.g., FG, Fay & Graubard, 2001; GST, Gosho et al., (2014); KC, Kauermann & Carroll, 2001; MK, MacKinnon, 1985; MD, Mancl & DeRouen, 2001; MBN, Morel et al., 2003; PAN, Pan, 2001; WL, Wang & Long, 2011). Recently, Thompson et al. (2021) compared the performance of several methods to improve the standard error estimates of GEE with binary outcomes in small samples (n= 6 to 54). They found that some bias-corrected standard error methods (i.e., KC and FG standard errors) outperformed others to mitigate this bias when used with an independent working correlation matrix and were robust to the misspecification of the working correlation matrix. However, Thompson et al. (2021) noted significant overcorrections for the FG standard error with a very number of clusters (n = 6), which is also consistent with other findings (Leyrat et al., 2018; Scott et al., 2017).

Small-sample Correction Methods

For GLMMs, researchers could use Kenward-Roger (Kenward & Roger, 1997, 2009) or Satterthwaite approach (Satterthwaite, 1946) to adjust for standard errors and degrees of freedom of fixed effect estimates with small sample sizes. Numerous studies have provided illustration and/or evaluation of these small sample corrections for multilevel models in psychological and educational research (Ferron et al., 2009; Luke, 2017; McNeish, 2017; McNeish & Stapleton, 2016; Stroup, 2013). However, researchers in social sciences are not familiar with small-sample corrections for the GEE approach, which is an important issue to consider in the calculation of the standard error of BC-IRR.

As mentioned in the review of performance, the sandwich estimator for the standard error in GEE is downwardly biased when the sample size is small. In our study, we chose four corrections that could potentially improve small-sample performance and are able to account for unbalanced designs (e.g., different number of measurements across cases) based on previous studies. To understand these corrections, the sandwich estimator of standard errors of $\boldsymbol{\beta}$ based on equation (7) is provided:

$$\boldsymbol{\Phi}_{LZ} = \left(\sum_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1} \boldsymbol{M} \left(\sum_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1} \tag{11}$$

where $\boldsymbol{M}$ is the "meat" of the "sandwich": $\boldsymbol{M} = \sum_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \hat{\boldsymbol{r}}_j \hat{\boldsymbol{r}}'_j \boldsymbol{V}_j^{-1} \boldsymbol{D}_j$, with $\boldsymbol{D}_j = \boldsymbol{X}_j' \boldsymbol{A}_j$, $\boldsymbol{V}_j = \phi \boldsymbol{A}_j^{\frac{1}{2}} \boldsymbol{R}_j(\boldsymbol{\alpha}) \boldsymbol{A}_j^{\frac{1}{2}}$, $\hat{\boldsymbol{r}}_j \hat{\boldsymbol{r}}'_j = \left(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_j\right)\left(\boldsymbol{Y}_j - \hat{\boldsymbol{\mu}}_j\right)'$. When sample size is small, $\hat{\boldsymbol{r}}_j \hat{\boldsymbol{r}}'_j$ is downwardly biased and leads to underestimated standard errors of $\boldsymbol{\beta}$.

Two classes of corrected sandwich estimator have been proposed to deal with small samples: residual-based correction and design-based corrections (McNeish & Stapleton, 2016). Residual based corrections account for the small-sample bias by adding matrixes to the innermost part of $\boldsymbol{M}$ in the sandwich estimators, which could inflate the standard errors. For example, $\boldsymbol{\Phi}_{KC}$ is a bias-corrected sandwich estimator proposed by Kauermann and Carroll (2001), which rewrites the estimator from equation 11 such that

$$\begin{gathered} \boldsymbol{\Phi}_{KC} = \left(\sum_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1} \boldsymbol{M}_{KC} \left(\sum_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1} \\ \boldsymbol{M}_{KC} = \sum_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} (\boldsymbol{I} - \boldsymbol{H}_j)^{-\frac{1}{2}} \hat{\boldsymbol{r}}_j \hat{\boldsymbol{r}}'_j (\boldsymbol{I} - \boldsymbol{H}_j')^{-\frac{1}{2}} \boldsymbol{V}_j^{-1} \boldsymbol{D}_j, \end{gathered} \tag{12}$$

where $\boldsymbol{H}_j = \boldsymbol{D}_j \left(\sum_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1}$.

Other commonly used residual-based corrections include $\boldsymbol{\Phi}_{MD}$ proposed by Mncl and Deroued (2001):

$$\begin{aligned}
\boldsymbol{\Phi}_{MD} &= \left(\sum\nolimits_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1} \boldsymbol{M}_{MD} \left(\sum\nolimits_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1} \\
\boldsymbol{M}_{MD} &= \sum\nolimits_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} (\boldsymbol{I} - \boldsymbol{H}_j)^{-1} \hat{\boldsymbol{r}}_j \hat{\boldsymbol{r}}_j' (\boldsymbol{I} - \boldsymbol{H}_j')^{-1} \boldsymbol{V}_j^{-1} \boldsymbol{D}_j,
\end{aligned} \tag{13}$$

and $\boldsymbol{\Phi}_{FG}$ proposed by Fay and Graubard (2001):

$$\begin{aligned}
\boldsymbol{\Phi}_{FG} &= \left(\sum\nolimits_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1} \boldsymbol{M}_{FG} \left(\sum\nolimits_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1} \\
\boldsymbol{M}_{FG} &= \sum\nolimits_{j=1}^{J} Diag\{(1 - min\{b, \boldsymbol{Q}_{jj}\})^{-1/2}\} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} (\boldsymbol{I} - \boldsymbol{H}_j)^{-\mathbf{1}} \hat{\boldsymbol{r}}_j \hat{\boldsymbol{r}}_j' (\boldsymbol{I} - \\
&\boldsymbol{H}_j')^{-1} \boldsymbol{V}_j^{-1} \boldsymbol{D}_j Diag\{(1 - min\{b, \boldsymbol{Q}_{jj}\})^{-1/2}\},
\end{aligned} \tag{14}$$

where $\boldsymbol{Q} = \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j \left(\sum_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1}$ and b is the an upper bound for the correction by default has a value of 0.75.

On the other hand, design-based corrections take a different form and include additional additive terms to the sandwich estimator rather than modifying the “meat” of the “sandwich”. The Morel-Bokossa-Neerchal correction ($\boldsymbol{\Phi}_{MBN}$, Morel et al., 2003) is a commonly adopted design-based correction in applied studies, which comprises of the sandwich estimator and an additive term:

$$\begin{aligned}
\boldsymbol{\Phi}_{MBN} &= \boldsymbol{\Phi}_{LZ} + \delta\phi\left(\sum\nolimits_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1} \\
\delta &= \begin{cases} P/(J-P) \text{ if} J > (d+1)P \\ \quad 1/d \text{ otherwise} \end{cases},
\end{aligned} \tag{15}$$

where P is the number of predictors, J equals to the number of clusters, and $\phi = \max$ (r, $trace\left(\left(\sum_{j=1}^{J} \boldsymbol{D}_j' \boldsymbol{V}_j^{-1} \boldsymbol{D}_j\right)^{-1} \boldsymbol{M}\right)/P$). The default values for d and r are 2 and 1, respectively (Morel et al., 2003).

Simulation Study

Data Generation

Data were simulated using a negative binomial GLMM as shown in equation 6. We varied the following design factors, including series length (I), number of cases (J), immediate effect (γ_{20}), trend change (γ_{30}), variance components (σ_{u0}^2, σ_{u1}^2, σ_{u2}^2, σ_{u3}^2), dispersion (θ), and the time elapsed between intervention and measurement sessions ($B - A$). The chosen conditions and associated rationale are similar to the previous simulation studies with SCED count data (Declercq et al., 2019; Li et al., 2024). Specifically, the number of cases was 4 or 8, and the series length was 10 or 20. This represented a common setting based on reviews of single-case studies (Pustejovsky et al., 2023; Shadish & Sullivan, 2011). Following the MBD, the start of the treatment was staggered across cases: I = 10 (starting points of the intervention: 3, 4, 6, 7) and I = 20 (starting points of the intervention: 6, 8, 12, 14). The trend in the baseline phase γ_{10} = log(1.00) and thus there is no trend in the baseline phase. The immediate effect γ_{20} was set to was set to log(1.00), log (2.00), log (6.00), or log (12.00), representing the zero, small, medium, and large treatment conditions, respectively. For trend change γ_{30}, the zero, medium, and large effects were set to log(1.00), log(1.05) and log(1.10), which leads to the frequency count increasing by 0%, 5%, and 10% for each measurement in the treatment phase, respectively.

The between-case variance components for the intercept (σ_{u0}^2), baseline trend (σ_{u1}^2), immediate effect (σ_{u2}^2) and trend change (σ_{u3}^2) were paired as [0.1, 0, 0.1,0], [0.3, 0.0003, 0.3, 0.0003], or [0.5, 0.0005, 0.5, 0.0005]. Following the same practice in Moeyaert et al. (2016), the correlations between all random effects were set as -0.5. The 1:1000 ratio of σ_{u1}^2 and σ_{u3}^2 versus σ_{u0}^2 and σ_{u2}^2 was also determined based on the reanalysis of SCED count data in previous meta-analyses (Dart et al., 2014; Ledford et al., 2018; Verschuur et al., 2014). Table 3 summarizes all design conditions. The full factorial design resulted in 864 conditions, each with 2,000 independent replications.

To analyze the simulated data, we fitted the negative binomial GLMMs by using the R package *glmmTMB* (Brooks et al., 2017) and the GEE models with independent error structures via the R package *gee* (Carey, 2024). In GEE models, we applied four small-sample corrections in addition to the sandwich estimator, GEE_LZ ($\boldsymbol{\Phi}_{LZ}$), to get standard errors of coefficients, including GEE_KC ($\boldsymbol{\Phi}_{KC}$) , GEE_MD ($\boldsymbol{\Phi}_{MD}$), GEE_FG ($\boldsymbol{\Phi}_{FG}$), and GEE_MBN ($\boldsymbol{\Phi}_{MBN}$). All corrections were implemented using the R package *geesmv* (Wang et al., 2015). For GLMMs, we allowed all variance components and only one covariance between random intercept and immediate effect to be freely estimated to avoid convergence issues. Hence, the fitted GLMMs were under-specified regarding the level-2 variance-covariance structure, which allows us to examine the robustness of GLMMs for the misspecification. For both GLMM and GEE, each model was estimated using two reference time points (C), namely, the first and last sessions of the baseline phase, whereas the data were consistently generated with C at the last session of the baseline phase. This allows us to examine whether BC-IRR is robust to different coding strategies adopted for $Time_{ij}$. Standard errors of BC-IRRs for both models were computed via the delta method using the R package *msm* (Jackson, 2011). The simulation code is available at https://osf.io/eqa4h/?view_only=98358eeef01f4d20b043d902e8d16c0a.

Performance Measure

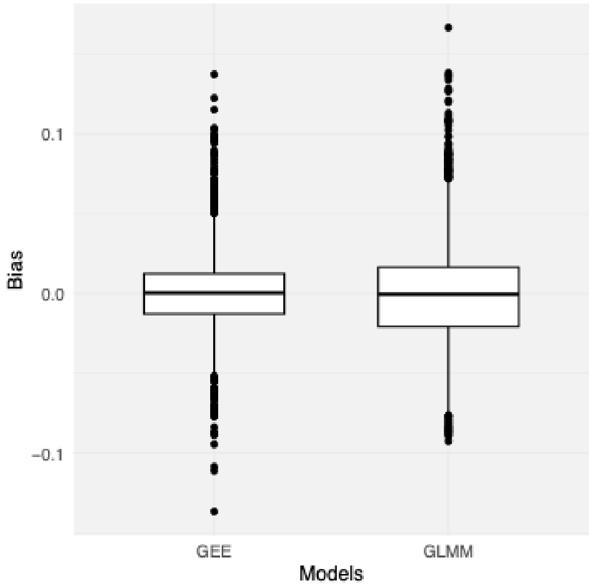
For each simulated condition, estimator performance was evaluated with four metrics. Bias of the BC-IRR was computed as the mean difference between the estimated and true BC-IRR across the 2,000 replications, as well as the mean-squared error (MSE), which is the sum of the estimator's variance and squared bias. We calculated the bias of the model-based standard error (SE) from the GEE approach as the difference between the mean reported SE and the empirical standard deviation of BC-IRR estimates. Finally, we evaluated the 95% confidence-interval coverage for the small-sample corrections found effective in the preceding section, defined as the proportion of replications whose Wald interval for BC-IRR contains the true value. Sampling variability in coverage was calculated with exact binomial error bounds, $.95 \pm 1.96\sqrt{\frac{0.95(1-0.95)}{2000}} = (.94, .96)$, so deviations outside this range showed under- or over-coverage beyond Monte-Carlo error.

Results

We first described overall results of simulation, followed by illustrating impactful factors on performance measures. To provide an overall comparison of the two estimation strategies (i.e., GLMM versus GEE), Figure 2 presents boxplots of bias across all simulation conditions for both GLMM and GEE. For each model, the median bias was very close to zero, indicating that both estimators were unbiased. The interquartile range (IQR) of bias for the GLMM was 0.03 and was 0.04 for the GEE, indicating that bias remained negligible under most design conditions.

Figure 2

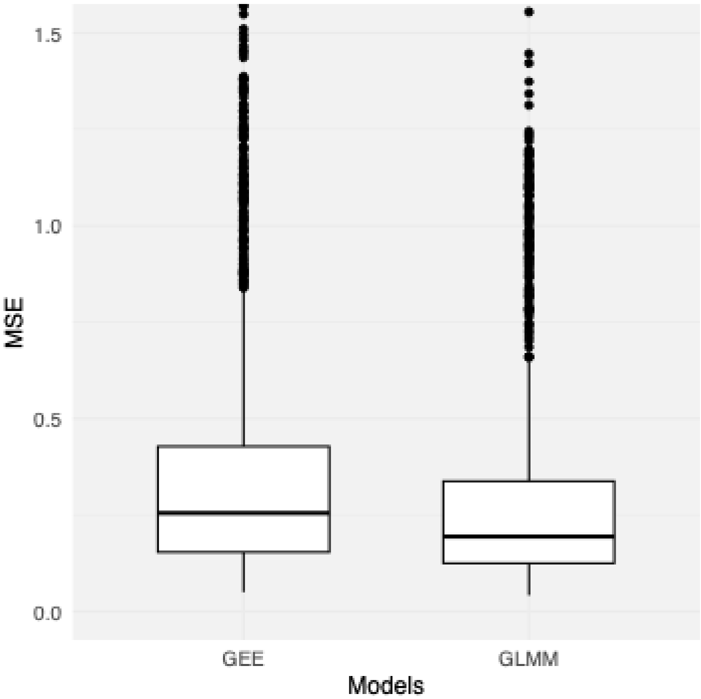
Boxplot of Bias of BC-IRR for GLMM and GEE



The boxplot of MSE across conditions is shown in Figure 3. The median MSE for the GLMM and GEE were 0.19 and 0.26, respectively. The IQR for the GLMM was 0.21, compared to 0.27 for the GEE. Overall, GLMM showed slightly better efficiency, if not at all, than GEE for BC-IRR estimators.

Figure 3

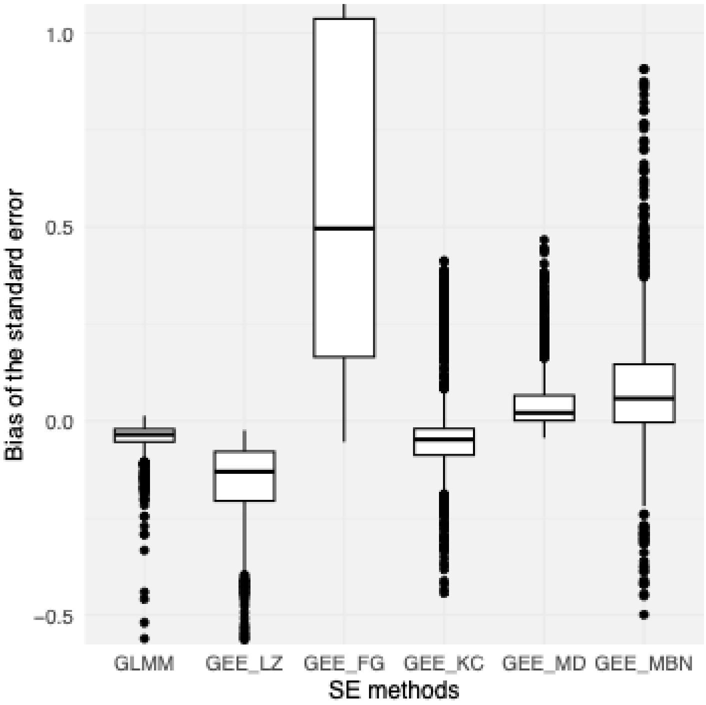
Boxplot of MSE of BC-IRR for GLMM and GEE



In Figure 4, we compare the bias of the standard error of BC-IRR using six different methods. The GLMM, along with the KC and MD corrections for sandwich estimators in GEE, yielded minimal median bias (-0.04 for the GLMM, -0.05 for the KC, and 0.02 for the MD) and narrow IQRs (0.03 for the GLMM, 0.07 for the KC, and 0.07 for the MD) for the standard error estimates. The GEE_FG correction showed the largest bias (median = 0.50) and the largest variability across conditions (IQR = 0.87), reflecting systematic over-correction of the standard error. The GEE_LZ and GEE_MBN methods showed relatively small median bias, -0.13 and 0.06, respectively, but both were not stable across conditions with wide IQRs (0.13 and 0.15, respectively).

Figure 4

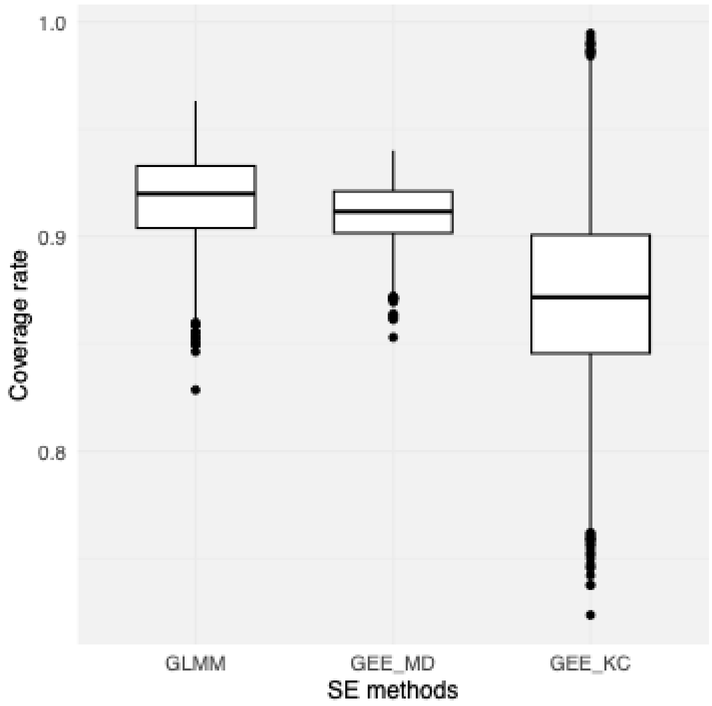
Boxplot of the Standard Error Bias of BC-IRR



As GLMM, GEE_KC, and GEE_MD yielded more accurate estimates and higher stability across all conditions for standard errors than other estimators, we only calculated the coverage rate of BC-IRR for the three estimators as shown in Figure 5. The median coverage rates for BC-IRR under the GLMM and the MD-corrected SEs are nearly identical (.92 vs. .91), with tight IQRs of .03 and .02, respectively. In comparison, the KC adjustment yields a slightly lower median coverage rate (.87) and wider IQR (.06).

Figure 5

Boxplot of the Coverage Rate of BC-IRR



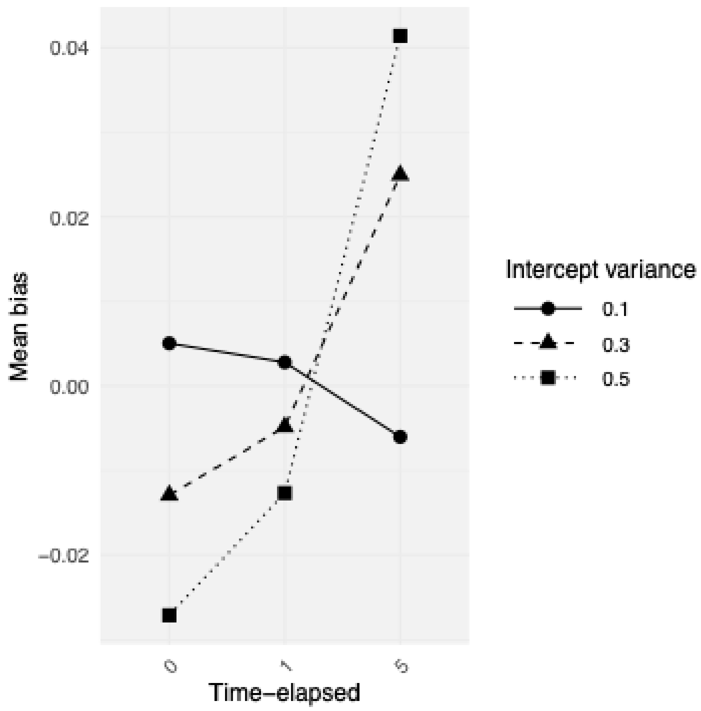
Next, to evaluate the impact of design factors on performance measures, we conducted mixed two-way ANOVA in which the *between-condition* factors were series length I, number of cases J, immediate-effect magnitude γ_{20}, trend change magnitude γ_{30}, between-case variance of the intercept σ_{00}^2, and dispersion θ, while the *within-condition* factor included the estimation strategy (only in ANOVAs with bias and MSE), standard error estimators (only in ANOVAs with SE bias and coverage rate), time elapsed ($B - A$) and reference time point (C). Main effects and all two-way interactions were tested, with η^2 statistics reported. All effects that were reported in the Results section had statistical significance at the α = .05 level. To concentrate on the most consequential sources of variation, we used scree-type plots of the ordered η^2 values to find main effects and interaction terms that fall to the left of the "elbow" point, which is analogous to using the scree plots to determine the number of factors in exploratory factor analyses (Li, 2024).

The bias of BC-IRR was mainly impacted by the time elapsed ($F(2, 554) = 2518.41$,

$\eta^2 = .19$) and the interaction of the intercept variance and the time elapsed ($F(4, 554) = 1344.96, \eta^2 = .20$). As depicted in Figure 6, when the between-case variance of the intercept was set at its smallest level ($\sigma_{00}^2 = 0.1$), the bias remained at the lowest level across different conditions of time elapsed. For the medium and large intercept-variance levels ($\sigma_{00}^2 = 0.3$ or 0.5), the mean bias increases as time elapsed increases. Although the main and interaction effects were statistically significant, there were no meaningful differences with the bias ranging from -0.03 to 0.04, indicating that the BC-IRR was practically unbiased throughout the design space.

Figure 6

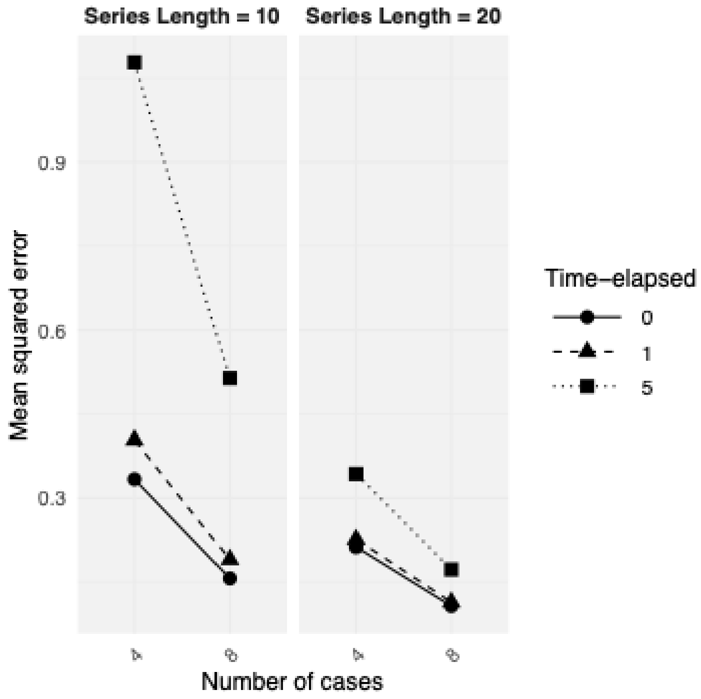
Bias of BC-IRR as a Function of Time Elapsed and Variance of the Intercept



The MSE of BC-IRR was mainly impacted by time elapsed ($F(2, 554) = 3059.73, \eta^2 = .24$), series length ($F(1, 237) = 4458.61, \eta^2 = .18$), and the number of cases ($F(1, 237) = 3567.22, \eta^2 = .14$). As shown in Figure 7, MSE increases when the time elapsed increases from zero to five. Regarding series length, MSE tends to decrease as the series length increased from 10 to 20. Similarly, increasing the number of cases consistently lowers the MSE across all design conditions.

Figure 7

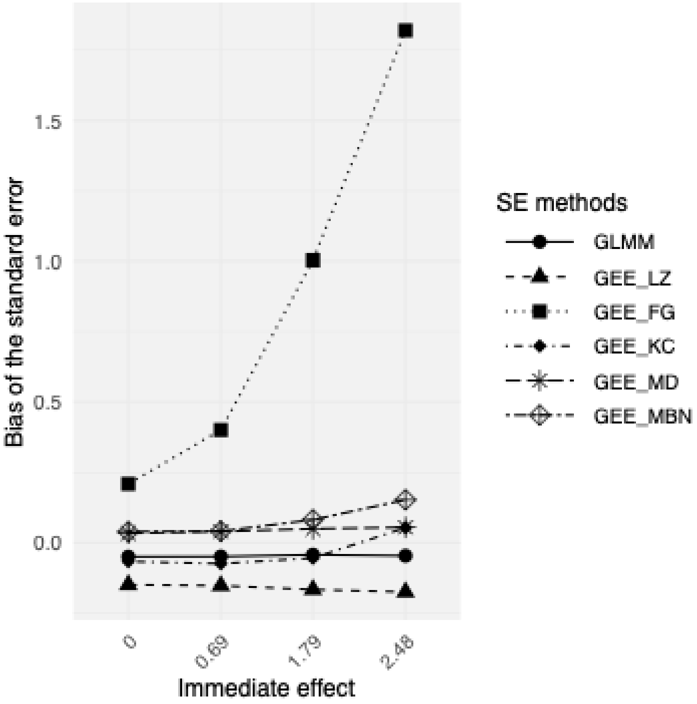
MSE of BC-IRR as a Function of Time Elapsed, Number of Cases, and Series Length



The bias of the standard error for BC-IRR was impacted mostly by standard error methods ($F(5, 1385) = 1127.53, \eta^2 = .30$) and its interaction with immediate effect ($F(15, 1385) = 171.78, \eta^2 = .14$). The GEE_FG correction exhibited the largest positive bias, which deteriorated exponentially as the immediate effect increased from 0 to 2.48. GEE_MBN correction showed overestimation while GEE_LZ shows underestimation across different settings. The GLMM estimator, as well as the KC and MD corrections for GEE's SE showed minimal bias and were relatively insensitive to the magnitude of the immediate effect.

Figure 8

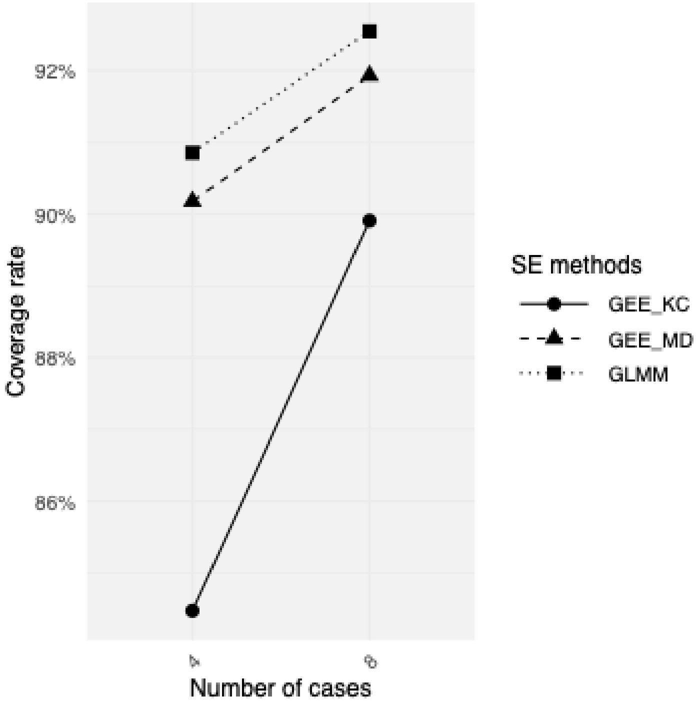
Bias of SE as a Function of SE Methods and Immediate Effect



From preceding analysis of SE bias, we only include the standard errors based on GLMM well as KC and MD corrections for GEE, all of which showed minimal bias, for calculating the coverage rate. The SE methods ($F(2, 554) = 6905.16, \eta^2 = .34$) and the number of cases ($F(1, 237) = 3750.37, \eta^2 = .19$) were largely associated with coverage rate. In Figure 9, with four cases the GLMM's SE and MD-based intervals had coverage rates close to the nominal .95 band (mean = .91 and .90, respectively), whereas KC-based interval showed under coverage (mean = .85). Increasing the number of cases to eight, improved coverage rates for each method, with GLMM to .93, MD to .92, and KC to .90. Overall, all methods showed under-coverage across conditions.

Figure 9

Coverage Rate of BC-IRR as a Function of SE methods and Number of Cases.



An Empirical Example

In this section we demonstrated the estimation of BC-IRR using data from a single-case study reported by Ota and DuPaul (2002). We first described the empirical data set, then outlined the model specification and step-by-step procedure for computing BC-IRR together with its standard errors and confidence intervals and finally interpreted the resulting effect sizes. The annotated R script used for the analysis is available at https://osf.io/eqa4h/?view_only=98358eeef01f4d20b043d902e8d16c0a.

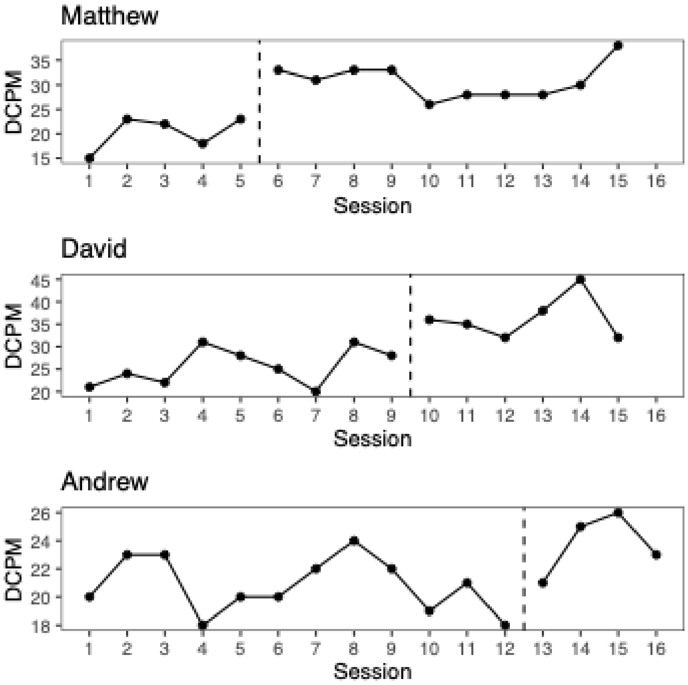
Data

Ota and DuPaul (2002) evaluated a computer-assisted mathematics program with a game format for three fourth- to sixth-grade students who have attention-deficit hyperactivity disorder (ADHD). A MBD across participants was used: each student completed a series of baseline typical mathematic instructions sessions followed by an intervention phase in which student used

20-minute computer sessions three to four times per week. For every session the investigators recorded two curriculum-based measures including Digits Correct per Minute (DCPM) and Problems Correct per Minute (NPCPM), For an illustration purpose, we used the DCPM outcome only and raw data of the three participants were extracted from the graphs presented in the original paper as shown in Figure 10.

Figure 10.

Performance in DCPM for Three Students with ADHD During Baseline and a Computer-Assisted Intervention, Adapted from Ota and DuPaul (2002).



Model Specification

The equations 5 and 8 define BC-IRR within a fully specified GLMM and GEE models that account for the baseline trend, immediate effect, and trend change. However, analysts often fit parsimonious models in practice. A reduced specification may reflect substantive knowledge about real data patterns or be imposed by non-convergence issues for complicated models. For demonstration purposes, we illustrated the computation of BC-IRRs using GLMM and GEE under three realistic scenarios with increasing complexities. For GEE models, we used two different SE methods including GEE_KC and GEE_MD, which showed more favorable statistical properties in the simulation study. In Scenario 1, we assume that there is no baseline trend or trend change, while the immediate effect varies across individuals. The GLMM is specified as follows (Model 1):

$Y_{ij} \sim Poisson\left(\lambda_{ij}, \theta\right)$ or $Y_{ij} \sim Negative\ Binomial\left(\lambda_{ij}, \theta\right)$

Level 1: $\log\left(\lambda_{ij}\right) = \beta_{0j} + \beta_{1j}Phase_{ij}$

$$\text{Level 2: } \begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} + u_{1j} \end{cases} \quad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u0u1} \\ \sigma_{u1u0} & \sigma_{u1}^2 \end{bmatrix}\right). \tag{16}$$

Based on this GLMM, the formula for BC-IRR is

$$\text{BC-IRR} = e^{\gamma_1 + \frac{1}{2}\sigma_{u1}^2 + \sigma_{u0u1}}. \tag{17}$$

The GEE is specified as follows with $\boldsymbol{R}_j(\boldsymbol{\alpha})$ being the independence working correlation structure (Model 2):

$$E\left(y_{ij} \middle| \boldsymbol{x_{ij}}\right) = \lambda_{ij}, \log\left(\lambda_{ij}\right) = \beta_0 + \beta_1 Phase_{ij}. \tag{18}$$

$$Var\left(y_{ij} \middle| \boldsymbol{x_{ij}}\right) = \phi v\left(\lambda_{ij}\right)$$

$$\boldsymbol{V_j} = \phi \boldsymbol{A}_j^{\frac{1}{2}} \boldsymbol{R}_j(\boldsymbol{\alpha}) \boldsymbol{A}_j^{\frac{1}{2}}$$

Based on the specified GEE, the formula for BC-IRR is

$$\text{BC-IRR} = e^{\beta_1}, \tag{19}$$

In Scenario 2, we assume that there is a fixed baseline trend, a fixed trend change, and varying immediate effects. The GLMM model is specified as follows (Model 3):

$Y_{ij} \sim Poisson\left(\lambda_{ij}, \theta\right)$ or $Y_{ij} \sim Negative\ Binomial\left(\lambda_{ij}, \theta\right)$

Level 1: $\log\left(\lambda_{ij}\right) = \beta_{0j} + \beta_{1j}Time_{ij} + \beta_{2j}Phase_{ij} + \beta_{3j}Time'_{ij}Phase_{ij}$

$$\text{Level 2: } \begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} \\ \beta_{2j} = \gamma_{20} + u_{2j} \\ \beta_{3j} = \gamma_{30} \end{cases} \begin{bmatrix} u_{0j} \\ u_{2j} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u0u2} \\ \sigma_{u2u0} & \sigma_{u2}^2 \end{bmatrix} \right). \tag{20}$$

Based on this GLMM, the formula for BC-IRR is simplified to

$$\text{BC-IRR} = e^{\gamma_2 + \gamma_3(B - \mathrm{A}) + \frac{1}{2}\sigma_{u2}^2 + \sigma_{u0u2}} . \tag{21}$$

The GEE is specified as follows with $\boldsymbol{R_j}(\boldsymbol{\alpha})$ being the independence working correlation structure (Model 4):

$$E(y_{ij}|\boldsymbol{x_{ij}}) = \lambda_{ij}, \log(\lambda_{ij}) = \beta_0 + \beta_1 Time_{ij} + \beta_2 Phase_{ij} + \beta_3 Time'_{ij} Phase_{ij}. \tag{22}$$

$$Var(y_{ij}|\boldsymbol{x_{ij}}) = \phi v(\lambda_{ij})$$

$$\boldsymbol{V_j} = \phi \boldsymbol{A}_j^{\frac{1}{2}} \boldsymbol{R}_j(\boldsymbol{\alpha}) \boldsymbol{A}_j^{\frac{1}{2}}$$

Based on the specified GEE, the formula for BC-IRR is

$$\text{BC-IRR} = e^{(\beta_2 + \beta_3(B - A))}, \tag{23}$$

In Scenario 3, we assume that there are varying baseline trends, immediate effects, and trend changes, and only the covariance between the random intercept and the random immediate effect is modeled. The model is specified as follows (Model 5):

$$Y_{ij} \sim Poisson\ (\lambda_{ij}, \theta) \text{ or } Y_{ij} \sim Negative\ Binomial\ (\lambda_{ij}, \theta)$$

$$\text{Level 1: } \log(\lambda_{ij}) = \beta_{0j} + \beta_{1j} Time_{ij} + \beta_{2j} Phase_{ij} + \beta_{3j} Time'_{ij} Phase_{ij}$$

$$\text{Level 2: } \begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} + u_{1j} \\ \beta_{2j} = \gamma_{20} + u_{2j} \\ \beta_{3j} = \gamma_{30} + u_{3j} \end{cases} \begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u0}^2 & 0 & \sigma_{u0u2} & 0 \\ 0 & \sigma_{u1}^2 & 0 & 0 \\ \sigma_{u2u0} & 0 & \sigma_{u2}^2 & 0 \\ 0 & 0 & 0 & \sigma_{u3}^2 \end{bmatrix} \right). \tag{24}$$

Based on this model, the formula for BC-IRR is simplified to

$$\text{BC-IRR} = e^{\gamma_2 + \gamma_3(B - \mathrm{A}) + \frac{1}{2}\sigma_{u2}^2 + \frac{1}{2}(B - A)^2 \sigma_{u3}^2 + \sigma_{u0u2}}. \tag{25}$$

The GEE (Model 6) for Scenario 3 is the same as Scenario 2 (see equation 22) and the resulting BC-IRR formula is the same as well (see equation 23).

The coding of the variables used in these models is the same as that shown in Table 2. For the second and third scenarios where $B - A$ (i.e., time elapsed between the introduction of treatment and the outcome measurement occasion) should be identified for BC-IRR calculation, multiple durations were applied including $B - A = 3$ (shortest treatment duration of this study) and $B - A = 9$ (longest treatment duration of this study). Models were fitted and analyzed using *glmmTMB* package version 1.1.11 (Brooks, et al., 2017), *gee* package version 4.13.29 (Carey, 2024), *geesmv* version 1.3 (Wang, 2015), and *saws* version 0.9.7.0 (Fay & Graubard, 2001) in R version 4.5.0. Standard errors of BC-IRRs were computed based on the delta method, using the *msm* package in R (Jackson, 2011).

Convergence and Test of Overdispersion

To determine whether a negative binomial model should be used in GLMM-based BC-IRR calculation, overdispersion was tested after fitting the Poisson model in each scenario for GLMM model. Then the Pearson's Chi-squared test was conducted to test overdispersion. For Scenario 1 and 2, the test results indicated that there was no significant overdispersion in the count outcome based on the Poisson model (*Model 1*: $\chi^2 = 19.98\ , df = 41, p = .99$, *Model 3*: $\chi^2 = 19.71, df = 39, p = .99$. Thus, Poisson models were selected for Scenario 1 and 2. For Scenario 3, the Poisson GLMM (Model 6) failed to converge due to non-positive-definite Hessian matrix when estimating fixed-effect standard error. All GEE models with Poisson mean specification converged successfully across the three scenarios.

Results

Table 4 summarizes the parameter estimates from both the conditional GLMMs and the marginal GEEs, together with their corresponding BC-IRRs. For brevity, we do not interpret coefficients of GLMM and GEE models. Our focus is on the calculation and interpretation of the final BC-IRR values. Detailed explanations of the GLMM parameters can be found in Li et al. (2023). The GEE coefficients are interpreted and used to obtain BC-IRR.

GLMM results

For Scenario 1, based on equation 17, the parameters used to calculate BC-IRR are the immediate effect ($\gamma_{10} = .30$), the variance of the random effects associated with immediate effect ($\sigma^2_{u1} = 0.004$), and the covariance between the baseline intercept and immediate effect ($\sigma_{u0u1} = 0.005$). The estimated BC-IRR was exp(0.30 + 0.004/2 + 0.005) = 1.36, indicating that the expected (marginal) rate ratio between the baseline and the treatment phase across all cases is 1.36. This means that the DPCM on average increased by 36% after the treatment across participants (SE = 0.10, 95% CI = 1.17-1.45). In this scenario, BC-IRR remains constant across all durations of the treatments, because there is no trend in the treatment phase.

For Scenario 2, based on equation 21, the parameters used to calculate BC-IRR are the immediate effect ($\gamma_{20} = 0.27$), trend change ($\gamma_{30} = .008$), the variance of the random effect associated with immediate effect ($\sigma^2_{u2} = 0.003$), and the covariance between the baseline intercept and shift in level treatment effect ($\sigma_{u0u2} = 0.005$). For example, when B-A=3, the estimated BC-IRR was exp[0.27 + (0.008 * 3) + 0.003/2 + .005] = 1.34, indicating the DPCM increased by 34% after 4 sessions of intervention (SE = 0.18, 95% CI = 1.00-1.69). This effect is comparable to the effect that would have been obtained from a potential RCT in which participants in the treatment group are consecutively treated for 4 sessions and the outcomes measured immediately after session 4 are compared with those from a never-treated control group.

For Scenario 3, the GLMM did not converge due to non-positive-definite Hessian matrix when estimating standard errors of fixed effects.

GEE results

For Scenario 1, based on equation 19, the parameter used to calculate BC-IRR is marginal shift in level treatment effect ($\beta_1 = 0.33$). The estimated BC-IRR was exp(0.33) = 1.39, indicating the DPCM increased by 39% averagely after the treatment across participants (KC-adjusted SE = 0.02, 95% CI = 1.34-1.44; MD-adjusted SE = 0.02, 95% CI = 1.35-1.43).

For Scenario 2 and 3 using GEE model, based on equation 23, the parameter used to calculate BC-IRR is the marginal immediate effect ($\beta_2 = 0.27$) and the marginal trend change ($\beta_3 = 0.002$). For example, when B-A=3, The estimated BC-IRR was exp(0.27 + 0.002*4) = 1.31, indicating the DPCM increased 31% averagely after 5 sessions (KC-adjusted SE = 0.09, 95% CI = 1.13-1.49; MD-adjusted SE = 0.10, 95% CI = 1.12-1.50). Same with GLMM-based BC-IRR, this effect is comparable to the effect that would have been obtained from a potential RCT in which participants in the treatment group are consecutively treated for 4 sessions and the outcomes measured immediately after session 4 are compared with those from a never-treated control group.

Discussion

The present study systematically evaluated the statistical properties of BC-IRR when estimated via GLMMs and GEEs. Through a Monte Carlo simulation, we assessed the bias, MSE, and converge rate of BC-IRR estimators and compared the performance of various estimators for their standard errors with GLMM and GEE. We also included a real-data demonstration by calculating BC-IRR, its standard errors and confidence intervals under different model specifications based on researchers' prior knowledge under commonly encountered scenarios.

Overall, the simulation results suggest that both GLMM and GEE approaches yield accurate estimates of BC-IRR across various conditions. Although BC-ICC tends to be slightly overestimated when the time elapsed $B - A$ increases and the intercept variance becomes relatively large, the magnitudes of the bias are still trivial and meaningless in practice. Similarly, the standard errors of BC-IRR based on GLMM, and two corrections with GEE_KC and GEE_MD are relatively unbiased, whereas biased standard errors estimate are found in the sandwich estimator GEE_LZ, and the other two corrections with GEE_FG and GEE_MBN. Notably, the GEE_FG shows significant overcorrections for standard errors in small samples, which aligns with previous findings (Leyrat et al., 2018; Scott et al., 2017; Thompson et al., 2021). Both point estimates and standard error estimates of BC-IRR are not sensitive to the different choices of the reference point C and the misspecification of the level-2 variance-covariance structure (only for GLMM). The robustness of BC-IRR against potential misspecification allows researchers to adjust model specifications to resolve convergence issues without sacrificing estimation accuracy. Regarding the efficiency of BC-IRR estimators, GLMMs are slightly more efficient than GEE, while significant less efficiency is observed for both when the series length is 10 and the $B - A$ is 5.

For interval estimates, BC-IRR shows substantial under-coverage relative to the nominal level of 95% across all examined methods. When the number of cases is 8, the coverage rates are close to the nominal level for GLMM and GEE_MD. This finding is consistent with the known limitations of Wald-type confidence intervals in small samples. The Wald confidence intervals are adopted due to the lack of well-defined degrees of freedom for BC-IRR under either GLMM or GEE. We also caution that the interval estimate of BC-IRR should not be used for the purpose of statistical inferences of treatment effectiveness for a single study, not only because of the under-coverage found in the simulation, but also the inconsistency between the design in the null hypothesis (i.e., a hypothetical group-based design) and the data obtained from a single-case study. For statistical inferences in a single study, researchers should adopt WC-IRR, to evaluate the effectiveness of treatments on a typical case (i.e., all random effects = 0) or cases with the same random effects.

Based on the theoretical framework, simulation results, and observed convergence issues, we provide the following recommendations for researchers to calculate BC-IRR and obtain standard errors in their own study: (1) when researchers use GLMM to calculate BC-IRR and its standard error, determine the complexity of level-2 covariance structures based on prior theory and/or visual analyses. Starting from Scenario 2 or 3 is recommended to mitigate convergence problems; (2) if convergence issues are encountered or level-2 variance-covariance structures are difficult to specify, consider using GEE with an independent working correlation structure. For more accurate standard error estimates, GEE_KC or GEE_MD are recommended; and (3) When using a range of time elapsed $B - A$ values, it is important to note the longest treatment duration that is observed across all cases. We recommend computing BC-IRRs for $B - A$ values within the range. Choosing a $B - A$ that is longer than what was actually observed requires extrapolating the time trend beyond the data and the accuracy of the estimates could be questionable.

Limitations and Future Directions

Given that the calculation for BC-IRR relies on multiple fixed effects and variance components in the GLMM and that the undesirable statistical properties found for the GEE approach with small sample sizes in previous studies, the results from our study under the SCED context are very encouraging for further methodological studies. The findings are constrained by the selected simulation conditions, which, although typical in MBDs, may not generalize to other research contexts or designs. To release the full potential of calculating BC-IRR with an aim to synthesize results from both SCEDs and group-based designs, there are several important issues yet to be investigated. A thorough investigation of these issues is beyond the scope of this study, but we discuss them here to suggest directions for future research.

Bayesian Approach

In practice, researchers applying BC-IRR in SCEDs using GLMMs may face convergence challenges, or difficulty specifying variance-covariance structures. While GEEs offer a marginal modeling approach that can address some of these concerns, experienced researchers may still prefer GLMMs when prior theory or visual inspection supports their use, especially for improved efficiency. Bayesian estimation presents a promising alternative in this context. R packages such as *brms* (Bürkner, 2017) and *GLMMadaptive* (Rizopoulos, 2022) allow Bayesian GLMM fitting and offer tools to derive marginal estimates from conditional ones. Future studies should investigate the convergence properties and small-sample bias of BC-IRR estimates under a Bayesian framework.

Degrees of Freedom of BC-IRR

Although standard error estimates of BC-IRR via GLMM, GEE_KC, and GEE_MD are relatively unbiased, their associated Wald confidence intervals exhibited poor coverage. A critical unresolved issue is how to appropriately define degrees of freedom for BC-IRR, given its reliance on multiple components from GLMM or GEE. The Satterthwaite approximation (Satterthwaite, 1946) may offer a viable solution, given its favorable statistical properties found in dealing with small samples in SCED contexts (citations). Deriving Satterthwaite-type degrees of freedom for composite effect sizes like BC-IRR could be considered for future methodological work.

Outcome Scales

This study focused on count outcomes, which are frequently observed in SCEDs. However, other outcome types, such as rates, proportions, percentages, and ordinal or continuous scores from surveys, are also common. These varied outcome scales complicate synthesis across studies, as different distributions (e.g., Gaussian, Poisson, Binomial) entail different assumptions about the relationship between the mean and variance. Multilevel modeling frameworks have been adapted to produce appropriate effect sizes for these outcome types, yet the incompatibility of obtained effect sizes can hinder meta-analyses aiming to aggregate diverse findings. The GEE-based BC-IRR, which makes no distributional assumptions, may provide a solution. Future research should explore the performance of GEE based proportional effect sizes when applied to count, rate, proportion/percentage, and continuous outcomes.

References

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Akanda, M. A. S., & Alpizar-Jara, R. (2014). Estimation of capture probabilities using generalized estimating equations and mixed effects approaches. *Ecology and Evolution*, *4*(7), 1158−1165. https://doi-org.ezp2.lib.umn.edu/10.1002/ece3.1000

Atkins, D. C., Baldwin, S. A., Zheng, C., Gallop, R. J., & Neighbors, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors, 27*(1), 166–177. https://doi.org/10.1037/a0029508

Aiken, L. S., Mistler, S. A., Coxe, S., & West, S. G. (2015). Analyzing count variables in individuals and groups: Single level and multilevel models. *Group Processes & Intergroup Relations*, *18*(3), 290–314. https://doi.org/10.1177/1368430214556702

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Carey, V. J. (2024). gee: Generalized Estimation Equation Solver (Version 4.13-29) [R package]. Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.gee

Coxe, S., West, S. G., & Aiken, L. S. (2013). Generalized linear models. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (pp. 26–51). Oxford University Press.

Dart, E. H., Collins, T. A., Klingbeil, D. A., & McKinley, L. E. (2014). Peer management interventions: A meta-analytic review of single-case research. *School Psychology Review*, *43*(4), 367–384. https://doi.org/10 .17105/SPR-14-0009.1

Declercq, L., Jamshidi, L., Fernández-Castilla, B., Beretvas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (2019). Analysis of singlecase 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

Fay, M. P., & Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, *57*, 1198–1206. https://doi.org/10.1111/j.0006-341X.2001.01198.x

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*, 372−384. https://doi.org/10.3758/BRM.41.2.372

Fieberg, J., Rieger, R. H., Zicus, M. C., & Schildcrout, J. S. (2009). Regression modelling of correlated data in ecology: subject-specific and population averaged response patterns. *Journal of Applied Ecology*, *46*(5), 1018−1025. https://doi.org/10.1111/j.1365-2664.2009.01692.x

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis*. John Wiley & Sons.

Gosho, M., Sato, Y., & Takeuchi, H. (2014). Robust covariance estimator for small-sample adjustment in the generalized estimating equations: A simulation study. *Science Journal of Applied Mathematics and Statistics*, *2*(1), 20−25. https://doi.org/10.11648/j.sjams.20140201.13

Grimm, K. J., & Stegmann, G. (2019). Modeling change trajectories with count and zero-inflated outcomes: Challenges and recommendations. *Addictive Behaviors*, *94*, 4−15. https://doi.org/10.1016/j.addbeh.2018.09.016

Hawken, S., Potter, B. K., Little, J., Benchimol, E. I., Mahmud, S., Ducharme, R., & Wilson, K. (2016). The use of relative incidence ratios in self-controlled case series studies: an overview. *BMC Medical Research Methodology*, *16*, 1−9. https://doi-org.ezp1.lib.umn.edu/10.1186/s12874-016-0225-0

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, *55*(3), 688−698. https://doi.org/10.1111/j.0006-341X.1999.00688.x

Heagerty, P. J., & Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, *15*(1), 1−26. https://doi.org/10.1214/ss/1009212671

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Jackson, C. H. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, *38*(8), 1–29. https://doi.org/10.18637/jss.v038.i08

Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, *96*(456), 1387–1396. https://doi.org/10.1198/016214501753382309

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997. https://doi:10.2307/2533558

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*, 2583–2595. https://doi:10.1016/j.csda.2008.12.013

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*(2), 124–144. https://doi.org/10.1037/a0017736

Lane, P. W. (2013). Meta-analysis of incidence of rare events. *Statistical Methods in Medical Research*, *22*(2), 117−132. https://doi.org/10.1177/0962280211432218

Leckie, G., Browne, W. J., Goldstein, H., Merlo, J., & Austin, P. C. (2020). Partitioning variation in multilevel models for count data. *Psychological Methods, 25*(6), 787–801. https://doi.org/10.1037/met0000265

Ledford, J. R., King, S., Harbin, E. R., & Zimmerman, K. N. (2018). Antecedent social skills interventions for individuals with ASD: What works, for whom, and under what conditions? *Focus on Autism and Other Developmental Disabilities*, *33*(1), 3–13. https://doi.org/10.1177/1088357616634024

Lee, Y. & Nelder, J.A. (2004) Conditional and marginal models: another view. Statistical Science, *19*(2), 219–238. https://doi.org/10.1214/088342304000000305

Leyrat, C., Morgan, K. E., Leurent, B., & Kahan, B. C. (2018). Cluster randomized trials with a small number of clusters: which analyses should be used?. *International Journal of Epidemiology*, *47*(1), 321−331. https://doi.org/10.1093/ije/dyx169

Li, H. (2024). Model selection of GLMMs in the analysis of count data in single-case studies: A Monte Carlo simulation. *Behavior Research Methods*, *56*(7), 7963−7984. https://doi.org/10.3758/s13428-024-02464-7

Li, H., Li, J., Yu, X., Zheng, H., Sun, X., Lu, Y., ... & Bi, X. (2018). The incidence rate of cancer in patients with schizophrenia: a meta-analysis of cohort studies. *Schizophrenia Research*, *195*, 519−528. https://doi.org/10.1016/j.schres.2017.08.065

Li, H., Luo, W., & Baek, E. (2024). Multilevel modeling in single-case studies with zero-inflated and overdispersed count data. *Behavior Research Methods*, *56*(4), 2765−2781. https://doi.org/10.3758/s13428-024-02359-7

Li, H., Luo, W., Baek, E., Thompson, C. G., & Lam, K. H. (2025). Multilevel modeling in single-case studies with count and proportion data: A demonstration and evaluation. *Psychological Methods, 30*(4), 815–842. https://doi.org/10.1037/met0000607

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22. https://doi.org/10.1093/biomet/73.1.13

Lindsey, J. K., & Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine*, *17*(4), 447−469. https://doi.org/10.1002/(sici)1097-0258(19980228)17:4<447::aid-sim752>3.0.co;2-g

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*, 1494−1502. https://doi.org/10.3758/s13428-016-0809-y

Luo, W., Li, H., Baek, E., & Li, C. (2025, July 23). Between-case incidence rate ratio: A design comparable effect size for count outcomes in single case experimental designs. https://doi.org/10.31219/osf.io/9gdkn_v1

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3), 305−325. https://doi.org/10.1016/0304-4076(85)90158-7

Mancl, L. A., & DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, *57*(1), 126−134. https://doi.org/10.1111/j.0006-341X.2001.00126.x

McGillycuddy, M., Warton, D. I., Popovic, G., & Bolker, B. M. (2025). Parsimoniously fitting large multivariate random effects in glmmTMB. *Journal of Statistical Software*, *112*(1), 1–19. https://doi.org/10.18637/jss.v112.i01

McNeish, D. M., & Harring, J. R. (2017). Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics - Simulation and Computation*, *46*(2), 855–869. https://doi.org/10.1080/03610918.2014.983648

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, *51*(4), 495−518. https://doi.org/10.1080/00273171.2016.1167008

McNeish, D. (2017). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research*, *52*(5), 661−670. https://doi.org/10.1080/00273171.2017.1344538

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2016). The misspecification of the covariance structures in multilevel models for single-case data: A Monte Carlo simulation study. *The Journal of Experimental Education*, *84*(3), 473 −509. https://doi.org/10.1080/00220973.2015.1065216

Morel, J. G., Bokossa, M. C., & Neerchal, N. K. (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *45*(4), 395−409. https://doi.org/10.1002/bimj.200390021

Muff, S., Held, L., & Keller, L. F. (2016). Marginal or conditional regression models for correlated non-normal data?. *Methods in Ecology and Evolution*, *7*(12), 1514−1524. https://doi.org/10.1111/2041-210X.12623

Natesan Batley, P., Shukla Mehta, S., & Hitchcock, J. H. (2021). A Bayesian rate ratio effect size to quantify intervention effects for count data in single case experimental research. *Behavioral Disorders*, *46*(4), 226–237. https://doi.org/10.1177/0198742920930704

Onghena, P. (2005). Single-case designs. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (Vol. 3, pp. 1850–1854). Chichester, UK: John Wiley & Sons.

Ota, K. R., & DuPaul, G. J. (2002). Task engagement and mathematics performance in children with attention-deficit hyperactivity disorder: Effects of supplemental computer instruction. *School Psychology Quarterly, 17*(3), 242–257. https://doi.org/10.1521/scpq.17.3.242.20881

Pan, W. (2001). On the robust variance estimator in generalised estimatin equations. *Biometrika*, *88*(3), 901-906. https://www.jstor.org/stable/2673458

Paul, S., & Zhang, X. (2014). Small sample GEE estimation of regression parameters for longitudinal data. *Statistics in Medicine*, *33*(22), 3869–3881. https://doi.org/10.1002/sim.6198

Pustejovsky, J. E. (2015). Measurement-comparable effect sizes for single-case studies of free operant behavior. *Psychological Methods, 20(3),* 342–359. https://doi.org/10.1037/met0000019

Pustejovsky, J. E. (2018). Using response ratios for meta analyzing single-case designs with behavioral outcomes. *Journal of School Psychology, 68,* 99–112. https://doi.org/10.1016/j.jsp.2018.02.003

Pustejovsky, J. E., Swan, D. M., & English, K. W. (2023). An examination of measurement procedures and characteristics of baseline outcome data in single-case research. *Behavior Modification*, *47*(6), 1423−1454. https://doi.org/10.1177/0145445519864264

Rizopoulos, D. (2022). GLMMadaptive: generalized linear mixed models using adaptive Gaussian quadrature. *R package version 0.8-5*. https://CRAN.R-project.org/package=GLMMadaptive

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110−114. https://doi.org/10.2307/3002019

Scott, J. M., deCamp, A., Juraska, M., Fay, M. P., & Gilbert, P. B. (2017). Finite-sample corrected generalized estimating equation of population average treatment effects in stepped wedge cluster randomized trials. *Statistical Methods in Medical Research*, *26*(2), 583−597. https://doi.org/10.1177/0962280214552092

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

Spittal, M. J., Pirkis, J., & Gurrin, L. C. (2015). Meta-analysis of incidence rate data in the presence of zero events. *BMC Medical Research Methodology*, *15*, 1−16. https://doi: 10.1186/s12874-015-0031-0

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

Thompson, J. A., Hemming, K., Forbes, A., Fielding, K., & Hayes, R. (2021). Comparison of small-sample standard-error corrections for generalised estimating equations in stepped wedge cluster randomised trials with a binary outcome: A simulation study. *Statistical Methods in Medical Research*, *30*(2), 425−439. https://doi.org/10.1177/0962280220958735

Verschuur, R., Didden, R., Lang, R., Sigafoos, J., & Huskens, B. (2014). Pivotal response treatment for children with autism spectrum disorders: A systematic review. *Review Journal of Autism and Developmental Disorders*, *1*(1), 34–61. https://doi.org/10.1007/s40489-013-0008-z

Wang, M. (2015). geesmv: Modified Variance Estimators for Generalized Estimating Equations (Version 1.3) [R package]. Comprehensive R Archive Network. https://doi.org/10.32614/CRAN.package.geesmv

Wang, M., & Long, Q. (2011). Modified robust variance estimator for generalized estimating equations with improved small-sample performance. *Statistics in Medicine*, *30*(11), 1278−1291. https://doi.org/10.1002/sim.4150

Wilson, D. B. (2022). The relative incident rate ratio effect size for count-based impact evaluations: When an odds ratio is not an odds ratio. *Journal of Quantitative Criminology*, *38*, 323−341. https://doi.org/10.1007/s10940-021-09494-w

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, *42*, 121–130. https://doi.org/10.2307/2531248

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, *44*, 1049–1060. https://doi.org/10.2307/2531734

Zorn, C. J. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, 470–490. https://doi.org/10.2307/2669353

Table 1

Expressions for the Conditional and Marginal Expectations of the Count Outcome in a Two-level Random Coefficients GLMM

	Conditional Expectation	Marginal Expectation
Expression	$\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{u}_j)$	$\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{G}\boldsymbol{z}_{ij})$

Table 2

Example Coding Matrix for a Case with Count Data

Measurement Occasions (I)	$Time_{ij}$	$Phase_{ij}$	$Time'_{ij}Phase_{ij}$ $Time'_{ij} = I - (C + 1)$	$B - A$	Outcome (Y)
1	-5	0	0	−	1
2	-4	0	0	−	3
3	-3	0	0	−	2
4	-2	0	0	−	4
5	-1	0	0	−	2
6 (C)	0	0	0	−	3
7 (A)	1	1	0	0	10
8	2	1	1	1	9
9	3	1	2	2	11
10	4	1	3	3	6
11	5	1	4	4	9
12	6	1	5	5	8

Table 3

Summary of Factors and Conditions for Simulated Count Data

Parameter	Value
Series length (I)	10 (starting points of the intervention: 3, 4, 6, 7) or 20 (starting points of the intervention: 6, 8, 12, 14)
Number of cases (J)	4 or 8
Intercept (γ_{00})	log (2)
Baseline trend (γ_{10})	log(1.00)
Immediate effect (γ_{20})	log(1.00), log(2.00), log(6.00) or log(12.00)
Trend change (γ_{30})	log(1.00), log(1.05) or log(1.10)
Between-case variance	
[intercept (σ_{u0}^2), baseline trend (σ_{u1}^2), immediate effect (σ_{u2}^2), trend change (σ_{u3}^2)]	[0.1, 0, 0.1,0], [0.3, 0.0003, 0.3, 0.0003] or [0.5, 0.0005, 0.5, 0.0005]
Correlations between all random effects	−0.5
Dispersion parameter (θ)	5 or 10
Time elapsed (B−A)	0, 1 or 5

Table 4

Parameter Estimates and BC-IRR for Models 1 to 6

Model	Parameter	Estimate	S.E.	95% CI
Scenario 1				
Model 1-GLMM	γ_{00}	3.11	0.06	
	γ_{10}	0.30	0.07	
	$\sigma_{u_0}^2$	0.006		
	$\sigma_{u_1}^2$	0.004		
	$\sigma_{u_1 u_2}$	0.005		
	BC-IRR	1.36	0.10	(1.17, 1.55)
Model 2-GEE, KC-adjusted SE	β_0	3.11	0.08	
	β_1	0.33	0.02	
	BC-IRR	1.39	0.02	(1.34, 1.44)
Model 2-GEE, MD-adjusted SE	β_0	3.11	0.10	
	β_1	0.33	0.02	
	BC-IRR	1.39	0.02	(1.35, 1.43)
Scenario 2				
Model 3-GLMM	γ_{00}	3.12	0.08	
	γ_{10}	0.002	0.01	
	γ_{20}	0.27	0.11	
	γ_{30}	0.008	0.02	
	$\sigma_{u_0}^2$	0.006		
	$\sigma_{u_2}^2$	0.003		
	$\sigma_{u_0 u_2}$	0.005		
	BC-IRR (B – A = 3)	1.34	0.18	(1.00, 1.69)
	BC-IRR (B – A = 9)	1.41	0.32	(0.79, 2.03)
Model 4-GEE, KC-adjusted SE	β_0	3.14	0.12	
	β_1	0.007	0.01	
	β_2	0.27	0.09	
	β_3	0.002	0.03	
	BC-IRR (B – A = 3)	1.31	0.09	(1.13, 1.49)
	BC-IRR (B – A = 9)	1.32	0.25	(0.82, 1.82)
Model 4-GEE, MD-adjusted SE	β_0	3.14	0.14	
	β_1	0.007	0.01	
	β_2	0.27	0.15	
	β_3	0.002	0.05	
	BC-IRR (B – A = 3)	1.31	0.10	(1.12, 1.50)
	BC-IRR (B – A = 9)	1.32	0.42	(0.51, 2.14)
Scenario 3				
(Model 5 – GLMM)	Non-convergence			
Model 6 – GEE	Same results as Model 4			

Data Availability Statement

The data and R code used for the simulation and demonstration are available at https://osf.io/eqa4h/?view_only=98358eeef01f4d20b043d902e8d16c0a

Funding Statement

The research reported here is supported by the Institute of Education Sciences, U.S. Department of Education, through Grant [R305D240019] to Texas A&M University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Conflict of Interest Disclosure

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.