From Multilevel Modeling to GEE: Revisiting the Within- and Between-Person Debate with Binary Predictors and Outcomes

Ward B. Eiling (s9294163)

Department of Methodology and Statistics

Faculty of Social and Behavioural Sciences

Utrecht University

prof. dr. Ellen Hamaker & dr. Jeroen Mulder

June 25, 2025

Word count: 9,269/12,000

FETC-approved: 24-2003

Candidate Journal: Psychological Methods

Abstract

In longitudinal studies with clustered data, researchers are often interested in estimating within-person effects—how changes over time within one variable relate to changes in another. To properly isolate these effects, the multilevel linear modeling (MLM) literature has long emphasized the need to disentangle within- and between-person effects, typically using disaggregation techniques such as person-mean centering. However, these methods have been developed primarily for continuous predictors and outcomes and little attention has been paid to their application with binary predictors or outcomes, despite the prevalence of such variables in applied research. Moreover, the capacity of alternative estimation frameworks, such as Generalized Estimating Equations (GEEs), to recover these effects remains underexplored. This study addresses both gaps. First, we explain how within- and between-person effects may be understood for binary predictors and outcomes using four generative models. Second, we evaluate the performance of disaggregation methods across estimation frameworks (multilevel models vs. GEEs) and predictor and outcome types (binary vs. continuous) in retrieving within-person and contextual effects. Our results indicate that both Mundlak's contextual and hybrid approaches generalize robustly across all multilevel specifications. Person-mean centering—often considered the gold standard—performs similarly with continuous outcomes but is less effective with binary outcomes. GEEs implemented with disaggregation methods are able to recover effects when outcomes are continuous, but not when they are binary. We conclude with practical recommendations for model specification and offer directions for future research.

Keywords: clustered longitudinal data, contextual effect, person-mean centering, generalized estimating equations, multilevel models

From Multilevel Modeling to GEE: Revisiting the Within- and Between-Person Debate with Binary Predictors and Outcomes

Across a wide range of disciplines, researchers analyze clustered longitudinal data to investigate prospective—and potentially causal—relationships between variables. When analyzing such data, psychological researchers commonly use the multilevel linear modeling (MLM) framework (Bauer & Sterba, 2011). A fundamental concern within the MLM literature is that the relationship between a predictor and an outcome may differ across levels of analysis (e.g., Enders & Tofighi, 2007; Kreft et al., 1995; Raudenbush & Bryk, 2002). In the context of longitudinal data, this implies that relationships across time captured at the within-person level of a model may differ from stable, between-person relationships as captured at the between-person level of a model. This discrepancy, known as the *contextual effect*, is particularly important because failing to account for it results in an "uninterpretable blend" of within- and between-cluster relationships (Raudenbush & Bryk, 2002, p. 139). This blending obscures the within-person effect, which is typically the primary target of inference.

The MLM literature offers several *disaggregation methods* that disentangle within-person from between-person effects (Curran & Bauer, 2011). These include: (a) person-mean centering of predictor variables; and (b) entering uncentered predictors alongside cluster means as predictor of the random intercept (Bell & Jones, 2015; Kreft et al., 1995; Raudenbush & Bryk, 2002). However, the MLM literature has primarily focused on continuous predictors, with limited guidance on how to disentangle effects when dealing with *binary predictors*. As a result, applied researchers may default to intuition or omit centering altogether when analyzing categorical predictors (Yaremych et al., 2023). This issue is exacerbated in generalized linear mixed modeling (GLMM)[1] with *binary outcomes*, where the non-linearity between predictor and outcome further complicates interpretation (Austin & Merlo, 2017; Bolger & Laurenceau, 2013). This gap in methodological guidance is concerning since binary

[1] We refer to GLMM as the general estimation framework that includes MLMs for continuous outcomes and multilevel logistic models for binary outcomes.

variables are ubiquitous within the social sciences and researchers may be left unsure of how to specify their models correctly.

A well established alternative for analyzing clustered data is the Generalized Estimating Equations (GEEs) framework. It was originally developed in the biostatistics literature to accommodate longitudinal designs with non-normal outcomes (Liang & Zeger, 1986; Zeger & Liang, 1986). In biomedical studies, GEEs is widely used—and often favored over GLMMs—for modeling binary outcomes (cf. Diggle et al., 2002). An important strength of GEEs is its reliance on fewer unverifiable assumptions compared to GLMM (Hubbard et al., 2010; McNeish et al., 2017). As GEEs gains traction in psychology (e.g., McNeish et al., 2017; Muth et al., 2016), researchers familiar with MLM may question how GEEs compares to GLMMs and specifically whether the debate about separating within- from between-person effects using disaggregation method also applies when a GEEs approach is taken. While some prior work has addressed how to handle binary predictors in MLMs (Enders & Tofighi, 2007; Raudenbush & Bryk, 2002; Yaremych et al., 2023) or binary outcomes in GLMMs (Bolger & Laurenceau, 2013; Schunck & Perales, 2017), and others have compared estimation frameworks across disciplines (Ballinger, 2004; McNeish et al., 2017; Muth et al., 2016; Neuhaus et al., 1991; Yan et al., 2013), no study to date has examined how different disaggregation methods perform and compare across estimation frameworks and variable types.

This paper seeks to address this gap by offering a systematic comparison of disaggregation methods across estimation frameworks and data types. Specifically, the goal of this paper is twofold. First, we aim to clarify how we should think about within and between effects with a binary predictor and/or outcome. Second, we study the degree to which we can correctly estimate these effects with GLMM and GEEs implementations. This article is structured as follows. We outline four data generating models to conceptualize within-person and contextual effects across different measurement scales, supported by a running example. Next, we discuss estimation frameworks and methods for including predictors. We

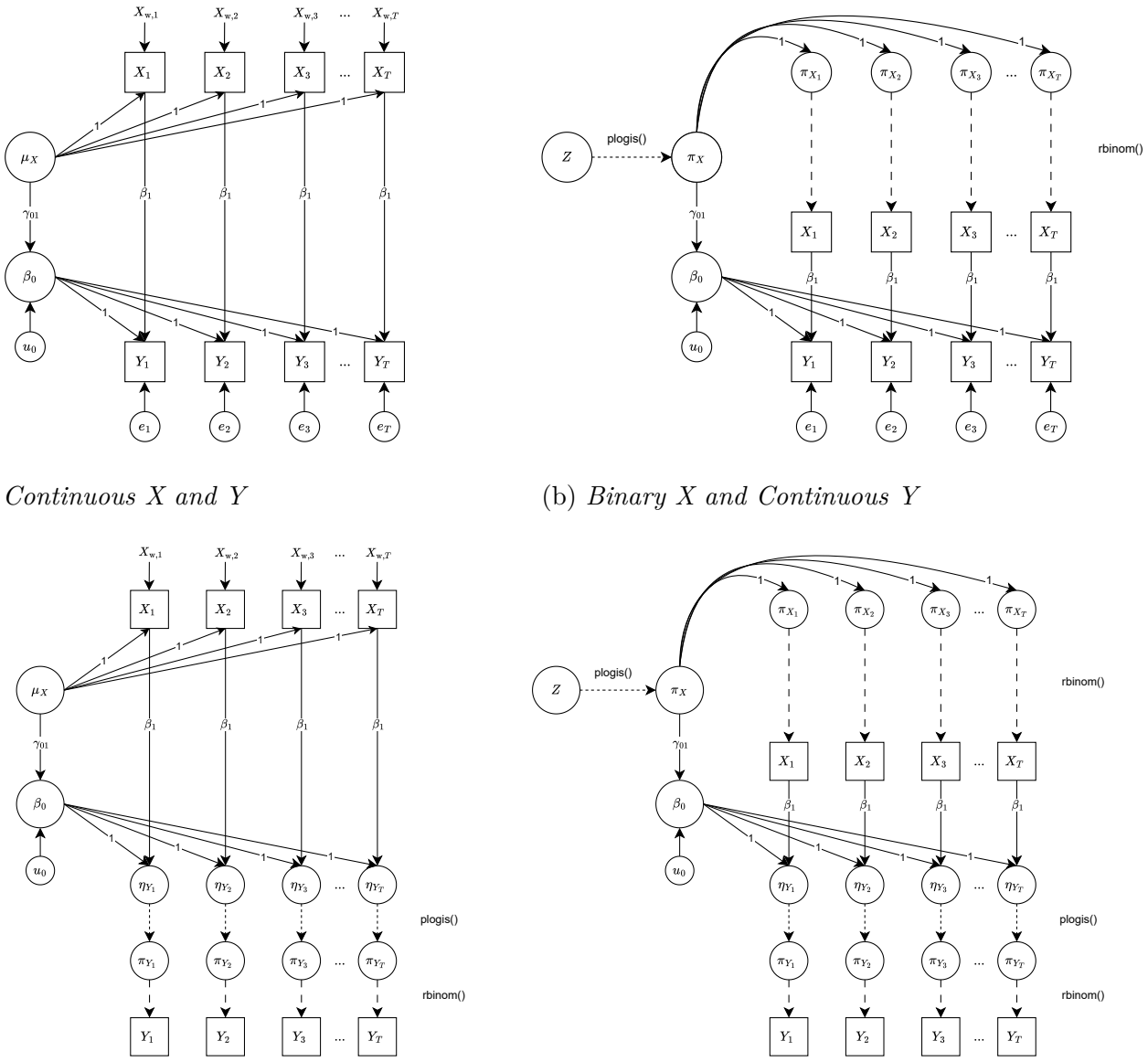
then present a simulation study in which we assess estimation performance across variations in (a) estimation strategy (GLMM vs. GEEs), (b) disaggregation methods, (c) measurement levels (binary vs. continuous), and (d) sample size—number of clusters and number of measurements per cluster. We evaluate when common modeling choices may lead to biased estimates and provide practical guidance for applied researchers working with clustered binary data. We conclude with a summary of key findings, limitations, and directions for future research.

Data Generating Models

In this section, we describe the four data generating models (DGMs) used to investigate contextual as well as within- and between-person effects in the context of binary predictors and/or outcomes. These models also form the basis of the simulation study. The DGMs differ in the measurement level of the predictor X and the outcome Y, and are visualized in Figure 1. In all four DGMs, we consider a single time-varying predictor X, which reflects both stable between-person differences and within-person variation over time. Similarly, Y is a time-varying outcome that exhibits both between-person differences and within-person fluctuations. Y is affected by X at both levels. For a given individual, a time-point-specific increase in X leads to a change in Y, which we refer to as the within-person effect. Additionally, individuals with higher person-level averages of X obtain systematically greater changes in Y than predicted by the within-person effect alone, suggesting a contextual spill-over effect.

Figure 1

Path diagrams of Generative Models with Within-Person (β_1) and Contextual Effect (γ_{01})



(a) *Continuous X and Y*

(b) *Binary X and Continuous Y*

(c) *Continuous X and Binary Y*

(d) *Binary X and Y*

Note. Path diagrams of the four data-generating models, varying by predictor type X (columns) and outcome type Y (rows), each either continuous or binary. In all models, the outcome is regressed on the predictor, and the person-level predictor mean μ_X predicts the outcome intercept β_0, yielding a within-person slope β_1 and a contextual effect γ_{01}. Circles represent latent variables, squares observed variables. Solid arrows denote linear relations; dotted arrows indicate logit transformations (`plogis` in `R`); dashed arrows indicate Bernoulli sampling with probability π (`rbinom` in `R`).

To illustrate these DGMs, we use a running example grounded in prior empirical work on the effect of mindfulness (X_{it}) on anger (Y_{it}), where X and Y vary over individuals i and time points t. Numerous studies have shown that, across individuals, those with higher overall mindfulness report lower overall levels of anger (Baer & Sauer, 2011; Borders et al., 2010; Brown & Ryan, 2003; Eisenlohr-Moul et al., 2016; Kashdan et al., 2016). Mindfulness also shows substantial within-person variability over time (Brown & Ryan, 2003; Kiken et al., 2015), and evidence suggests that on days when individuals report higher-than-usual mindfulness, they also report lower levels of state anger (Eisenlohr-Moul et al., 2016). Hence, both within- and between-person associations are negative, though effect sizes depend on the specific operationalization (Eisenlohr-Moul et al., 2016). Moreover, between-person associations tend to be larger, implying that a one-unit difference in average mindfulness between individuals is associated with a greater difference in anger-related outcomes than a one-unit increase in moment-to-moment fluctuations.

DGM 1: Continuous Predictor and Outcome

To illustrate the DGM for a continuous predictor and outcome, we consider the case of how mindfulness relates to daily experiences of anger (see Figure 1a). Each individual possesses a stable level of trait mindfulness, denoted as a person-specific latent mean $\mu_{X,i}$. This latent trait is assumed to vary across individuals according to a normal distribution $\mu_{X,i} \sim \mathcal{N}(0, \sigma^2_{X,\text{b}})$, where $\sigma^2_{X,\text{b}}$ reflects the between-person variance in mindfulness. The observed X for individual i at occasion t, denoted X_{it}, is a combination of an individual's stable trait level and a momentary deviation, and can thus be expressed as:

$$X_{it} = \mu_{X,i} + X_{\text{w},it}. \tag{1}$$

Here, $X_{\text{w},it}$ represents the within-person score in mindfulness, capturing occasion-specific fluctuations around the person's trait level. In Figure 1a, the within-person score is obtained as $X_{\text{w},it} = X_{it} - \mu_{X,i}$. $X_{\text{w},it}$ varies over time according to a normal distribution $X_{\text{w},it} \sim \mathcal{N}(0, \sigma^2_{X,\text{w}})$, where $\sigma^2_{X,\text{w}}$ reflects within-person variance in mindfulness. Daily anger, denoted

Y_{it}, is given by

$$Y_{it} = \beta_{0i} + \beta_1 X_{it} + e_{it}, \tag{2}$$

where β_{0i} is an individual-specific intercept capturing average levels of anger, and $e_{it} \sim \mathcal{N}(0, \sigma_e^2)$ is a person- and time-specific residual with variance σ_e^2. The individual-specific intercept β_{0i} is a function of a person's average mindfulness ($\mu_{X,i}$) through

$$\beta_{0i} = \gamma_{00} + \gamma_{01}\mu_{X,i} + u_{0i}, \tag{3}$$

where γ_{00} is the grand mean of anger (given that $\mathrm{E}[\mu_{X,i}] = 0$), and γ_{01} is the contextual effect of trait mindfulness on anger. The person-specific residual $u_{0i} \sim \mathcal{N}(0, \sigma_u^2)$ is normally distributed with variance σ_u^2, and represents residual heterogeneity in average anger between individuals after accounting for average levels of mindfulness. By substituting Equation 3 into Equation 2, we obtain the combined expression:

$$Y_{it} = \gamma_{00} + \gamma_{01}\mu_{X,i} + u_{0i} + \beta_1 X_{it} + e_{it}. \tag{4}$$

In this DGM, there are two ways in which mindfulness (X) influences anger (Y). To make this explicit, we substitute Equation 1 into Equation 4, yielding:

$$Y_{it} = \gamma_{00} + (\gamma_{01} + \beta_1)\mu_{X,i} + u_{0i} + \beta_1 X_{\mathrm{w},it} + e_{it}. \tag{5}$$

First, occasion-specific changes in mindfulness ($X_{\mathrm{w},it}$) affect anger (Y_{it}) through the coefficient β_1, which represents the within-person effect: how anger changes on a given occasion when an individual is more or less mindful than usual. Between-person differences in average mindfulness ($\mu_{X,i}$) also influence anger, both through the term β_1 and via the contextual effect γ_{01}. The between-person effect—the impact of a person's typical mindfulness level ($\mu_{X,i}$) on average anger level—is given by $\beta_{\mathrm{b}} = \gamma_{01} + \beta_1$ (Mundlak, 1978).[2] If there is no contextual effect ($\gamma_{01} = 0$), the within- and between-person effects are equal and captured by β_1. Conversely, the presence of a contextual effect ($\gamma_{01} \neq 0$) implies that the between-person

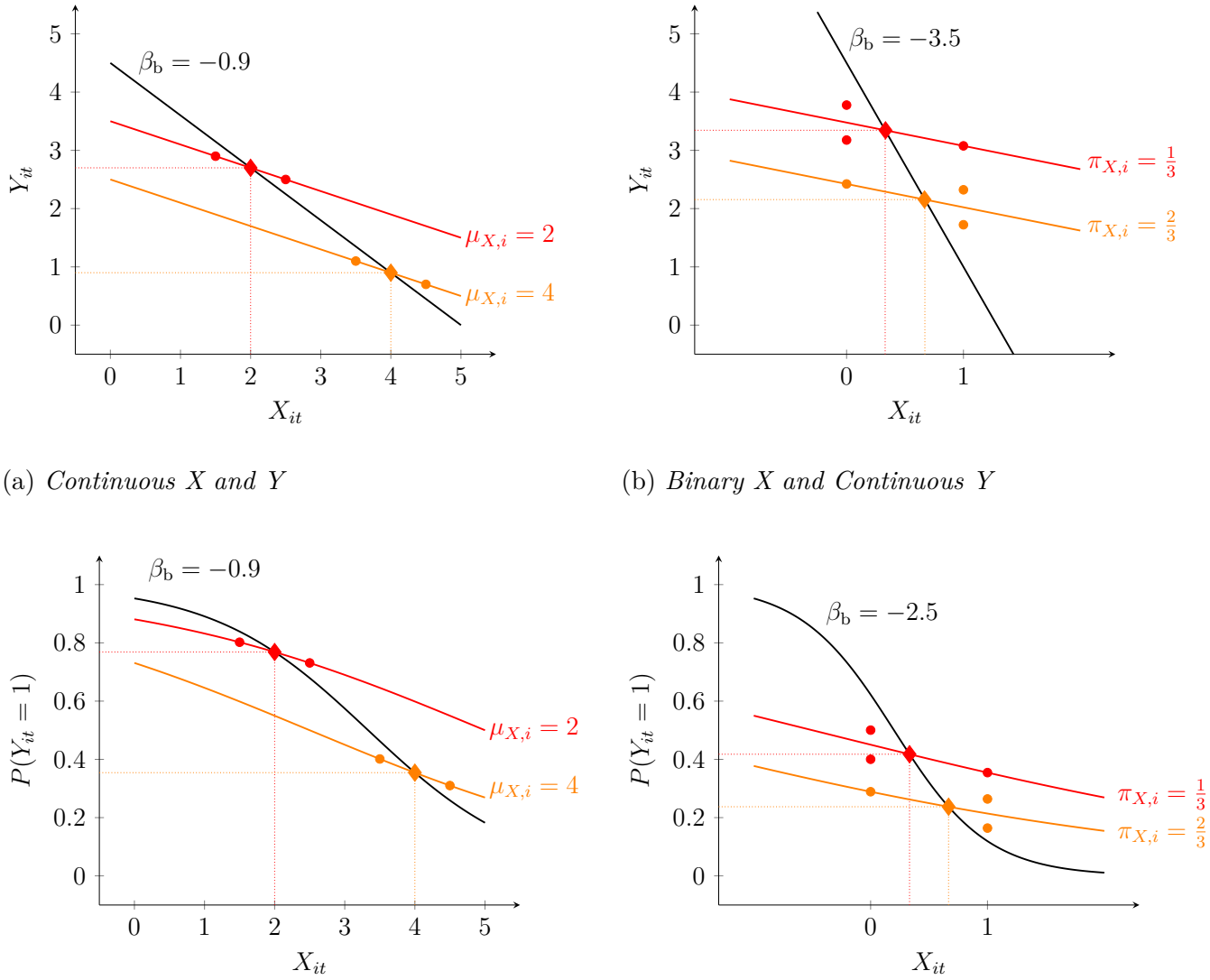
[2] This equivalence does not hold in the presence of *random slopes*; see Snijders and Bosker (2011).

effect differs from the within-person effect. A positive γ_{01} indicates that individuals who are more mindful on average experience more anger than would be expected based only on the within-person relationship. A negative γ_{01} implies that individuals who are more mindful on average experience less anger than would be expected based solely on the within-person relationship.

Empirical work has found that the between-person effect tends to be larger in magnitude than the within-person effect ($\beta_\text{b} > \beta_1$; Eisenlohr-Moul et al., 2016). As both effects are negative, this implies the presence of a *negative* contextual effect. To reflect this, we employ the following parameter values: $\gamma_{00} = 4.5, \quad \gamma_{01} = -0.5, \quad \beta_1 = -0.4$, which implies $\beta_\text{b} = -0.5 - 0.4 = -0.9$. As visualized in Figure 2a, these parameters yield a stronger negative association at the between-person level than at the within-person level. We assume that the within-person effect is identical for both individuals and that these individuals vary only in their average levels of mindfulness and anger. Specifically, the individual depicted in orange has a distribution centered around a higher trait level of mindfulness ($\mu_{X,i} = 4$) than the individual depicted in red ($\mu_{X,i} = 2$). Due to the negative between-person effect of X on Y, it follows that the individual in orange also has lower overall anger levels.

Figure 2

Illustration of Within- and Between-Person Effects in Generative Models with $\beta_1 = -0.4$



(a) *Continuous X and Y*

(b) *Binary X and Continuous Y*

(c) *Continuous X and Binary Y*

(d) *Binary X and Y*

Note. Within-person effects (β_1) and between-person effects (β_b) are shown for data-generating mechanisms varying by predictor type X (columns) and outcome type Y (rows), each either continuous or binary. For continuous outcomes, relationships are linear; for binary outcomes, the Y-axis reflects the probability of $Y = 1$, and relationships follow a logistic curve. The solid black line denotes the between-person effect; colored lines represent within-person effects for two individuals. Circles are hypothetical observations; diamonds mark where each within-person slope intersects the between-person slope. Thin dotted lines indicate each person's mean on X and Y.

DGM 2: Binary Predictor and Continuous Outcome

To illustrate the DGM for a binary predictor and continuous outcome, we consider the case where daily engagement in a mindfulness practice (e.g., meditation), coded as 1 if practiced and 0 otherwise, influences experiences of anger (see Figure 1b). Each individual is characterized by a latent propensity to engage in mindfulness on a given day, denoted Z_i, which varies across individuals according to a normal distribution $Z_i \sim \mathcal{N}(0, \sigma_Z^2)$. This continuous latent trait is transformed via a logistic function to yield the individual-specific probability $\pi_{X,i} \in (0, 1)$—taking on values strictly between 0 and 1—of engaging in mindfulness:

$$\pi_{X,i} = \frac{e^{Z_i}}{1 + e^{Z_i}}. \tag{6}$$

The observed mindfulness practice score $X_{it} \in \{0, 1\}$ for individual i at occasion t is a *binary* variable determined by the underlying probability $\pi_{X,i}$ and a Bernoulli sampling process: $X_{it} \sim \text{Bernoulli}(\pi_{X,i})$. In this way, each person's binary predictor values vary across time due to probabilistic sampling, and the extent of within-person variability is a function of their underlying propensity $\pi_{X,i}$. Individuals with extreme trait probabilities (close to 0 or 1) will show little within-person variation, whereas those with moderate probabilities will exhibit more fluctuation.

Anger on day t for person i, denoted Y_{it}, is determined by the person's observed mindfulness practice X_{it}, a person-specific intercept β_{0i}, and a residual e_{it} (see Equation 2). The random intercept β_{0i} is a function of $\pi_{X,i}$ via

$$\beta_{0i} = \gamma_{00} + \gamma_{01}\pi_{X,i} + u_{0i}, \tag{7}$$

where γ_{00} represents the extrapolated anger level for a hypothetical individual with an extreme aversion to mindfulness practice (with $\pi_{X,i} = 0$), γ_{01} is the contextual effect of average mindfulness engagement, and $u_{0i} \sim \mathcal{N}(0, \sigma_u^2)$ captures person-level heterogeneity unexplained by $\pi_{X,i}$. Substituting Equation 7 into Equation 2 yields the composite model:

$$Y_{it} = \gamma_{00} + \gamma_{01}\pi_{X,i} + u_{0i} + \beta_1 X_{it} + e_{it}. \tag{8}$$

This DGM captures two distinct effects of mindfulness practice (X) on anger (Y). Within-person fluctuations in mindfulness engagement influence anger (Y_{it}) through β_1, reflecting how anger changes on days with versus without mindfulness practice. Between-person differences in the general tendency to practice mindfulness ($\pi_{X,i}$) influence anger both through β_1 and via the contextual effect γ_{01}. Accordingly, the contextual effect γ_{01} captures whether individuals who typically engage more in mindfulness experience systematically different anger levels than expected based on their day-to-day engagement alone.

Let us consider a plausible generative structure, where $\gamma_{00} = 4.5, \quad \gamma_{01} = -3.1, \quad \beta_1 = -0.4$, resulting in a between-person effect of $\beta_{\text{b}} = -3.5$. This pattern, visualized in Figure 2b, suggests that individuals who engage in mindfulness more frequently benefit from reduced anger than expected based on their day-to-day practice alone. The two individuals vary only in their propensity to engage in mindfulness and average anger. The person indicated in orange has a higher overall probability of practicing mindfulness ($\pi_{X,i} = \frac{2}{3}$) than the person indicated in red ($\pi_{X,i} = \frac{1}{3}$). Given the negative between-level effect of X on Y, the person in orange also has lower overall anger than the person in red.

DGM 3: Continuous Predictor and Binary Outcome

To illustrate the DGM for a continuous predictor and a binary outcome, we consider how daily levels of mindfulness relate to the probability of experiencing an anger episode, coded as 1 if an anger episode occurred and 0 otherwise (see Figure 1c). The continuous predictor X_{it} is generated as described in DGM 1 (see Equation 1). The binary outcome $Y_{it} \in \{0, 1\}$ reflects whether an anger episode occurred for person i on day t. The log-odds of an anger episode, $P(Y_{it} = 1)$, is determined by a linear function of observed mindfulness and a person-specific intercept:

$$\text{logit}\left[P(Y_{it} = 1)\right] = \beta_{0i} + \beta_1 X_{it}. \tag{9}$$

The expression for the random intercept β_{0i} is the same as in DGM 1 (see Equation 3). Substituting Equation 3 into Equation 9 yields the combined expression:

$$\text{logit}\left[P(Y_{it}=1)\right] = \gamma_{00} + \gamma_{01}\mu_{X,i} + u_{0i} + \beta_1 X_{it}. \tag{10}$$

The use of a logistic link function, unlike the linear function in DGM 1 and 2, entails that the coefficients are expressed on the log-odds scale. The within-person effect β_1 reflects how day-to-day deviations in mindfulness relate to fluctuations in the log-odds of experiencing anger. The contextual effect γ_{01} indicates whether individuals who are more inclined to practice mindfulness on average differ in their overall log-odds of anger, over and above the within-person association.

Let us consider a generative structure where $\gamma_{00} = 3$, $\gamma_{01} = -0.5$, and $\beta_1 = -0.4$, so that the between-person effect is $\beta_\text{b} = -0.9$. This pattern, visualized in Figure 2c, reflects that individuals with higher average mindfulness are less likely to experience anger episodes, beyond what is expected from daily fluctuations alone, holding constant u_{0i}. While the within-person effect is of the same size as in DGMs 1 and 2, the curves are now logistic rather than linear. The two persons differ only in their average mindfulness levels and propensity to experience anger episodes. Specifically, the person shown in orange has a distribution centered around a higher trait level of mindfulness ($\mu_{X,i} = 4$) than the person shown in red ($\mu_{X,i} = 2$). Due to the negative between-person effect of X on Y, it follows that the person in orange also has a lower overall probability of experiencing an anger episode.

To clarify the interpretation of coefficients, we convert the log-odds to probabilities using the inverse logit transformation. We begin by substituting the parameter values into Equation 10:

$$\text{logit}\left[P(Y_{it}=1)\right] = 3 - 0.5\mu_{X,i} - 0.4X_{it} + u_{0i}.$$

We first consider the within-person effect by examining a change in X_{it} while $\mu_{X,i}$ remains constant. Suppose Clara (red color in Figure 2c) has an average mindfulness score of $\mu_{X,i} = 2$, and on a given day scores $X_{it} = 2$ (i.e., at her mean). In interpreting model coefficients, we

must evaluate the random effect u_{0i} at some fixed value; setting it to zero is conventional, as it corresponds to the average individual in terms of their unexplained baseline risk. Thus, suppose Clara has $u_{0i} = 0$, indicating no residual deviation in her overall propensity to experience anger beyond what is explained by her average mindfulness. In that case, her log-odds of experiencing an anger episode are:

$$\text{logit}\left[P(Y_{it} = 1)\right] = 3 - 0.5 \cdot 2 - 0.4 \cdot 2 = 1.2,$$

which can be converted to the probability of an anger episode using the inverse logit transformation:

$$P(Y_{it} = 1) = \frac{e^{1.2}}{1 + e^{1.2}} \approx 0.77.$$

Now suppose Clara scores one point above her average (i.e., $X_{it} = 3$). Then:

$$\text{logit}\left[P(Y_{it} = 1)\right] = 1.2 - 0.4 = 0.8,$$

$$P(Y_{it} = 1) = \frac{e^{0.8}}{1 + e^{0.8}} \approx 0.69.$$

Thus, a one-unit increase in mindfulness relative to her usual level decreases the probability of an anger episode from 77% to 69%, illustrating the within-person effect.

Next, we examine the between-person effect by considering a change in $\mu_{X,i}$ while holding $X_{\text{w},it}$ constant. Consider Clara ($\mu_{X,i} = 2$) and Maya ($\mu_{X,i} = 4$; orange color in Figure 2c), both at their respective means ($X_{it} = \mu_{X,i}$). Supposing Maya has $u_{0i} = 0$, her log-odds of an anger episode are:

$$\text{logit}\left[P(Y_{it} = 1)\right] = 3 - 0.5 \cdot 4 - 0.4 \cdot 4 = -0.6,$$

$$P(Y_{it} = 1) = \frac{e^{-0.6}}{1 + e^{-0.6}} \approx 0.35.$$

Thus, given that $u_{0i} = 0$, Maya's higher higher typical mindfulness results in a much lower probability of experiencing anger than Clara (35% vs. 77%), illustrating the between-person effect.

DGM 4: Binary Predictor and Outcome

To illustrate the DGM for a binary predictor and a binary outcome, we consider how engaging in a mindfulness practice (1 = yes, 0 = no) relates to the probability of experiencing an anger episode (1 = yes, 0 = no) on the same day (see Figure 1d). The predictor X_{it} and the person-specific probability $\pi_{X,i}$ are generated as described in DGM 2 (see Equation 6).

The probability of an anger episode, $P(Y_{it} = 1)$, is determined by a logistic function of a person's observed mindfulness practice X_{it} and a person-specific intercept β_{0i} (see Equation 9). The expression for the random intercept β_{0i} is the same as in DGM 2 (see Equation 7). By plugging Equation 7 into Equation 9, the composite model becomes:

$$\text{logit}\left[P(Y_{it} = 1)\right] = \gamma_{00} + \gamma_{01}\pi_{X,i} + u_{0i} + \beta_1 X_{it}. \tag{11}$$

As in DGM 3, coefficients are interpreted on the log-odds scale. The within-person effect captured by β_1 reflects the average change in the log-odds of experiencing anger on days with versus without mindfulness practice. The contextual effect γ_{01} captures the association between individuals' average engagement in mindfulness practice and the log-odds of experiencing anger, beyond daily fluctuations.

For a substantive generative structure, we consider $\gamma_{00} = 0.5$, $\gamma_{01} = -2.1$, and $\beta_1 = -0.4$, so that $\beta_{\text{b}} = -2.5$. As shown in Figure 2d, this implies that individuals who tend to practice mindfulness more frequently are less likely to experience anger episodes, beyond what is expected from daily fluctuations alone. As in DGM 3, we have a logistic curve. The two persons differ only in their general tendency to practice mindfulness and propensity to experience anger episodes. Specifically, the person shown in orange has a greater overall probability of practicing mindfulness ($\pi_{X,i} = \frac{2}{3}$) than the person shown in red ($\pi_{X,i} = \frac{1}{3}$). Due to the negative between-person effect of X on Y, it follows that the person in orange also has a lower overall probability of experiencing an anger episode.

Estimation Frameworks

To analyze the hierarchical longitudinal data generated by the specified DGMs, we consider two primary estimation frameworks: GLMMs and GEEs. Within the GLMM framework, we employ a standard MLM for continuous outcomes and a multilevel logistic model for binary outcomes. For the GEEs framework, we consider three commonly used implementations, which will be described in more detail later.

For the specification of the predictor in each of these estimation strategies, we employ one method known to yield an uninterpretable blend and three methods known to disentangle within- from between-person effects in the context of the MLM (Curran & Bauer, 2011). While these methods are well established for MLMs with continuous predictors, it is less clear how they apply to MLMs with binary predictors, to multilevel logistic models, and to GEEs. Accordingly, we first discuss these methods within an MLM with a continuous predictor applied to data from DGM 1, providing a clear reference point. We then extend this discussion to binary predictors applied to data of DGM 2. Next, we introduce the multilevel logistic model used to analyze data from DGM 3 and DGM 4, and outline how the implementation of the four disaggregation methods differs in this context. Finally, we describe how these methods are adapted when applied within the GEEs framework.

Generalized Linear Mixed Model

The GLMM extends the Generalized Linear Model (GLM) to accommodate hierarchical or clustered data structures (Breslow & Clayton, 1993; Hoffman, 2015; Stroup, 2012). Like the GLM, the GLMM allows the analyst to specify an appropriate distribution for the outcome variable (e.g., normal, binomial, Poisson) and a link function that defines the relationship between the linear predictor and the expected outcome. However, unlike the GLM, which assumes independence of observations, the GLMM explicitly accounts for dependence among observations within clusters—such as repeated measures within individuals—by incorporating random effects.

This estimation framework generalizes the MLM by permitting non-normal outcomes

and non-identity link functions, thereby broadening its applicability to a wide range of data types. In this paper, we focus on two widely used instantiations of the GLMM: (1) the MLM which is a special case of the GLMM with a normally distributed outcome and an identity link function and (2) the multilevel logistic model which is a GLMM with a binary outcome, modeled using a binomial distribution and a logit link function.

Multilevel Linear Model

In MLM, it is well established that when using continuous, time-varying predictors, researchers interested in within-person effects must disentangle these from between-person effects. This distinction is critical because a model that does not explicitly separate these sources of variation will produce a conflated estimate that mixes the two, impairing interpretability and potentially leading to incorrect conclusions (e.g., Enders & Tofighi, 2007; Kreft et al., 1995; Raudenbush & Bryk, 2002).

If the within- and between-person associations are identical—or there is no between-person variability in X—there is no conflation and bias is absent. However, in typical psychological research, one cannot assume that contextual effects are absent as level-specific effects can differ drastically (Curran & Bauer, 2011; Robinson, 1950). In fact, detecting and modeling such contextual effects is often one of the key motivations for applying MLMs. In the context of the MLM with a continuous predictor, we review four approaches to specifying the predictor variable. We begin with a method that conflates within- and between-person effects, followed by three methods designed to isolate the within-person component.

Uncentered (UC) Method. We fit the following simple two-level model to data generated by DGM 1 with a time-varying continuous predictor X_{it} and a random intercept:

$$Y_{it} = \beta_{0i} + \beta_1^* X_{it} + e_{it}, \tag{12}$$

$$\beta_{0i} = \gamma_{00} + u_{0i}. \tag{13}$$

Substituting the between-person equation into the within-level model yields:

$$Y_{it} = \gamma_{00} + u_{0i} + \beta_1^* X_{it} + e_{it}. \tag{14}$$

In this formulation, the estimate of β_1^* reflects a blend of the within- and between-person associations. The degree to which each component contributes to the estimate depends on the number of time points per individual (T) and the amount of variability at each level (Mundlak, 1978; Neuhaus & Kalbfleisch, 1998; Raudenbush & Bryk, 2002). As T increases, the estimate tends to be more influenced by the within-person effect, but the conflation remains unless explicitly modeled.

The source of this conflation can be understood by considering a causal path diagram (see Figure 1a), where the latent mean of the predictor μ_X, serves as a common cause of both X_{it} and Y_{it} (via its influence on the random intercept β_{0i}). In other words, μ_X acts as a confounder in the relationship between X_{it} and Y_{it} (e.g., Berlin et al., 1999). This confounding disappears under two special cases: (1) when there is no contextual effect (i.e., the path from μ_X to β_0 is zero), or (2) when there is no between-person variation in μ_X (i.e., all individuals have the same latent mean on the predictor). Outside of these conditions, failing to account for μ_X results in biased estimation of the within-person effect.

Centering Within Clusters (CWC). One solution is centering the predictor X within clusters, using the sample mean $\bar{X}_i$ for person i and the within-person centered predictor $(X_{it} - \bar{X}_i)$. This method ensures that the within-person slope is not confounded by between-person differences (e.g., Hoffman, 2015; Kreft et al., 1995; Raudenbush & Bryk, 2002). The model is then specified as:

$$Y_{it} = \beta_{0i} + \beta_{\mathrm{w}}(X_{it} - \bar{X}_i) + e_{it}, \tag{15}$$

$$\beta_{0i} = \gamma_{00} + u_{0i}. \tag{16}$$

Substituting the second equation into the first gives:

$$Y_{it} = \gamma_{00} + u_{0i} + \beta_{\mathrm{w}}(X_{it} - \bar{X}_i) + e_{it}. \tag{17}$$

Since $(X_{it} - \bar{X}_i)$ contains no between-person variance, β_{w} estimates the within-person slope.

The Hybrid (HB) Method. The Hybrid method extends the CWC approach by including the cluster mean $\bar{X}_i$ as a predictor of the person-specific intercept, thereby esti-

mating the between-person slope β_{b}:

$$\beta_{0i} = \gamma_{00} + \beta_{\text{b}}\bar{X}_i + u_{0i}. \tag{18}$$

The full model then becomes:

$$Y_{it} = \gamma_{00} + \beta_{\text{b}}\bar{X}_i + u_{0i} + \beta_{\text{w}}(X_{it} - \bar{X}_i) + e_{it}. \tag{19}$$

Mundlak's Contextual (MuCo) Method. An alternative approach, known as Mundlak's model (Mundlak, 1978), closely aligns with the structure of the DGMs but relies on estimating $\mu_{X,i}$ with the observed person-level mean $\bar{X}_i$. This approach retains the raw predictor X_{it} in the within-person equation while including $\bar{X}_i$ in the expression of the intercept:

$$Y_{it} = \beta_{0i} + \beta_1 X_{it} + e_{it}, \tag{20}$$

$$\beta_{0i} = \gamma_{00} + \gamma_{01}\bar{X}_i + u_{0i}. \tag{21}$$

Substituting the second equation into the first gives:

$$Y_{it} = \gamma_{00} + \gamma_{01}\bar{X}_i + u_{0i} + \beta_{\text{w}}X_{it} + e_{it}. \tag{22}$$

At first glance, it may seem unintuitive that the regression coefficient for the raw, uncentered time-varying predictor X_{it} in a model that includes the person-mean $\bar{X}_i$ represents the within-person effect. However, this result follows from the statistical relationship between X_{it} and $\bar{X}_i$, which are inherently correlated (cf. Hoffman, 2015). When both terms are included in the model, their shared variance in predicting the outcome is partialled out, and each coefficient reflects its unique contribution to the prediction. Specifically, the coefficient for X_{it} captures the within-person effect β_1 and the coefficient for $\bar{X}_i$, in turn, captures the contextual effect γ_{01}. Given the absence of random slope and no interactions between X_{it} and $\bar{X}_i$, the between-person effect is given by $\beta_{\text{b}} = \gamma_{01} + \beta_1$ (Raudenbush & Bryk, 2002).

Application of Disaggregation Methods to Binary Predictors. Although binary predictors differ substantively from continuous variables, the statistical principles underlying their treatment in MLMs are analogous. As shown by Enders and Tofighi (2007), the algebra and logic of centering does not require us to distinguish between continuous and categorical predictors. Consequently, the same four disaggregation methods used for continuous predictors can be applied to a binary predictor generated from DGM 2 without modification. However, the application of methods has distinct interpretational implications when involving binary compared to continuous predictors as discussed in DGM 3 and 4.

Additionally, modeling binary predictors introduces practical challenges due to their bounded and discrete nature. When many individuals have a person-mean $\bar{X}_i$ (i.e., the proportion of time points with $X_{it} = 1$) close to 0 or 1, within-person variability is limited, leading to unstable estimates of the within-person effect. Between-person predictors such as $\bar{X}_i$ are further constrained to at most $T+1$ discrete values, resulting in truncated distributions that can attenuate correlations and bias contextual effects (Asparouhov & Muthén, 2019). For instance, with $T = 5$, $\bar{X}_i$ can only take values like 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. When between-person variability is low, the effective range of $\bar{X}_i$ is restricted further. Moreover, since within-person variance $\text{Var}(X_{it}) = \bar{X}_i(1-\bar{X}_i)$ is a deterministic function of the between-person mean, the assumption of independence between levels is violated, potentially biasing both within- and between-person estimates (Asparouhov & Muthén, 2019).

Multilevel Logistic Model

When the outcome variable is binary (e.g., experiencing an anger episode), a linear model is no longer appropriate due to the bounded nature of probabilities (between 0 and 1). To analyze data structures—such as those generated in DGM 3 and 4—we employ the multilevel logistic model, a specific form of the GLMM using a binomial distribution and a logit link function (Anderson & Aitkin, 1985; Stiratelli et al., 1984). In logistic models, the

linear predictor maps onto the probability of a binary event via the logit function:

$$\text{logit}\left[P(Y_{it}=1)\right] = \log\left(\frac{P(Y_{it}=1)}{1-P(Y_{it}=1)}\right),$$

which transforms probabilities $P(Y_{it} = 1) \in (0, 1)$ onto the unbounded log-odds scale $(-\infty, \infty)$. As a consequence, the parameters and partition of variance do not have the intuitive interpretation familiar from linear models (Merlo et al., 2006). Fixed effects describe changes in log-odds rather than in the outcome variable itself. For interpretability, log-odds can be converted back into probabilities using the inverse logit transformation as shown in DGM 3.

To explain how disaggregation methods apply to the multilevel logistic model, we revisit the four methods introduced in the linear case and illustrate them using a continuous predictor (DGM 3). The same logic generalizes to the binary predictor case (DGM 4).

Uncentered (UC) Method. We begin with a model including the raw predictor:

$$\text{logit}\left[P(Y_{it}=1)\right] = \gamma_{00} + u_{0i} + \beta_1^* X_{it}. \tag{23}$$

Here, β_1^* reflects a conflated estimate combining within- and between-person components.

Centering Within Clusters (CWC). To isolate the within-person effect, we center the predictor around the individual's mean:

$$\text{logit}\left[P(Y_{it}=1)\right] = \gamma_{00} + u_{0i} + \beta_{\text{w}}(X_{it} - \bar{X}_i). \tag{24}$$

Now, β_{w} captures the effect of deviations from a person's own average.

The Hybrid (HB) Method. To also obtain an estimate of the between-person effect, we extend the CWC method to include the cluster mean:

$$\text{logit}\left[P(Y_{it}=1)\right] = \gamma_{00} + \beta_{\text{b}}\bar{X}_i + u_{0i} + \beta_{\text{w}}(X_{it} - \bar{X}_i). \tag{25}$$

This formulation allows β_{w} to represent the within-person effect and β_{b} the between-person effect—how typical levels of the predictor relate to overall likelihood of the outcome.

Mundlak's Contextual (MuCo) Method. An alternative formulation includes both the raw predictor and its person-mean:

$$\text{logit}\left[P(Y_{it}=1)\right] = \gamma_{00} + \gamma_{01}\bar{X}_i + u_{0i} + \beta_1 X_{it}. \tag{26}$$

As in the linear case, this is algebraically equivalent to the hybrid model, since the logit does not alter the underlying structure of the model and relationships. Thus, the between-person effect is recovered as $\beta_{\text{b}} = \gamma_{01} + \beta_1$.

Generalized Estimating Equations

GEEs were introduced by Liang and Zeger (1986) as an extension of GLMs for analyzing correlated, non-normally distributed outcomes—most notably, longitudinal and clustered data. In this study, we apply GEEs with an identity link for continuous outcomes (DGMs 1 and 2) and a logit link for binary outcomes (DGMs 3 and 4). Unlike GLMMs, the most common GEEs approaches do not include random effects. Instead, they account for within-cluster correlation through a so-called *working correlation structure*. Since GEEs do not model random effects, they make fewer unverifiable assumptions, which can be beneficial in applied settings (Hubbard et al., 2010; McNeish et al., 2017).

Marginal versus Conditional Inference. In the biomedical literature, GEEs are typically described as *marginal models*, whereas GLMMs are referred to as *conditional models* (e.g., Diggle et al., 2002). The central distinction lies in how the standard interpretation of regression coefficients—*holding all other variables constant*—is applied: in GEEs, this applies only to observed predictors, rendering estimates *marginal* with respect to the random intercept; in GLMMs, it also includes the random effects (e.g., u_{0i}), yielding estimates *conditional* on them. Marginal models achieve this by integrating over the distribution of the random effects, effectively averaging out subject-specific deviations.

Crucially, under the identity (linear) link, marginalization over the random effects yields the same fixed-effect slopes in GEEs as in GLMMs—that is, marginal and conditional models produce equivalent estimates. However, this equivalence breaks down for nonlinear

links such as the logistic (Neuhaus et al., 1991).[3] Accordingly, GEEs with logit link (like standard GLMs) estimate *population-averaged* effects (Diggle et al., 2002): the expected log-odds change in the mean outcome for a one-unit increase in a predictor, averaged across the population, holding other observed variables constant. In contrast, GLMMs with a logit link yield *conditional-on-the-random-effect* estimates: the expected log-odds change in the outcome for a one-unit increase in a predictor, holding both observed predictors and random effects (in this case u_{0i}) constant. For a GLMM with random intercept, it is conventional to set the random intercept residual to zero ($u_{0i} = 0$; see DGM 3), reflecting 'person-specific' inference for an *average individual* in terms of their unexplained baseline propensity for the outcome.

Importantly, whether a marginal or conditional interpretation is preferable depends on the scientific aim of the study (Diggle et al., 2002; Neuhaus et al., 1991). Consider the example of a longitudinal study examining the effect of mindfulness on the occurrence of anger episodes (a binary outcome). A marginal question suitable for GEEs might be: "How does practicing mindfulness affect the overall likelihood of experiencing anger episodes in the general population?" Here, the focus is on population-average effects, which are suitable for informing public health interventions. A conditional question suitable for GLMM might be: "For a given individual, how does their likelihood of experiencing an anger episode change following an increase in mindfulness?" Here, the focus is on understanding 'person-specific' within-person processes.

Working Correlation Structures and Computational Considerations. Regardless of the type of outcome variable, GEEs account for within-cluster correlation through a user-specified working correlation matrix R, which encodes the expected correlation pattern among repeated measurements Y_{it} within the same cluster i. Common choices include

[3] An interactive RShiny app illustrating the relationship between marginal and conditional effects in multilevel logistic models is available at: wardeiling.shinyapps.io/GLMM_population-averaged-and-person-specific-interpretations.

the independence, exchangeable, and first-order autoregressive (AR(1)) structures. To illustrate these three structures, let us consider a dataset with five time points, resulting in a 5×5 working correlation matrix, where ρ denotes the correlation between repeated observations:

$$R_{\text{indep}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad R_{\text{exch}} = \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}, \quad R_{\text{AR(1)}} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

The *independence* structure R_{indep} assumes no within-cluster correlation and is often used as a baseline. The *exchangeable* (or compound symmetric) structure R_{exch}[4] assumes a constant correlation between all pairs of measurements within a cluster. The *AR(1)* structure $R_{\text{AR(1)}}$ assumes that correlations decay exponentially with increasing time lag, which is often appropriate for regularly spaced longitudinal data. While correctly specifying the working correlation can improve efficiency, GEEs estimates tend to remain consistent even under misspecification of this matrix (Ballinger, 2004; Zeger et al., 1988).

The simplicity of GEE estimation has notable advantages, but also presents trade-offs when compared to GLMMs. It is generally less computationally intensive, particularly with simpler working correlation structures. However, convergence criteria are less straightforward (see Hardin & Hilbe, 2012, p. 92), and model comparison is more complex. For an accessible introduction to the computational and methodological aspects of GEEs, including iterative estimation and model selection, see McNeish et al. (2017) and Ballinger (2004).

Application of Disaggregation Methods to GEEs. Although GEEs are widely used for analyzing clustered longitudinal data, the application of disaggregation methods to separate within- and between-person effects has received little attention in this framework. An exception is the study by Begg and Parides (2003), who applied several disaggregation

[4] In *linear* models, an exchangeable correlation structure in GEE is mathematically equivalent to including a random intercept in a multilevel model (e.g., Gardiner et al., 2009; Koper & Manseau, 2009).

methods (including UC, CWC, HB, MuCo) within the GEE framework using an exchangeable working correlation structure. They found that, in the linear case, GEE estimates of all methods closely resembled those from a GLMM, whereas under a logit link, this correspondence no longer held. This aligns with theoretical work showing that, under a logit link, GEE estimates are marginal and cannot recover cluster-specific (i.e., conditional) within-person effects (Neuhaus et al., 1991). More specifically, parameter estimates from GEE tend to attenuate (i.e., shrink toward zero) as the residual variance of the unmodeled random intercept increases (For an illustration of this, see wardeiling.shinyapps.io/GLMM_population-averaged-and-person-specific-interpretations; Begg & Parides, 2003; Neuhaus et al., 1991). This suggests that GEE-based disaggregation under a logit link only supports valid person-specific interpretations when the DGM includes no random effects—neither unexplained heterogeneity in the baseline outcome propensity ($\sigma_u = 0$) nor in the predictor-outcome relationships. This is implausible in most empirical settings.

In principle, disaggregation methods can be applied within the GEE framework in much the same way as in GLMMs. However, whereas GLMMs model within-cluster dependence via random effects (e.g., u_{0i}), GEEs account for this dependence through a working correlation matrix R. As a result, deriving the GEE-based formulations of each method are obtained by omitting u_{0i} from the corresponding GLMM equations. In the case of continuous outcomes (DGMs 1 and 2), the GEE analogues of the four disaggregation methods follow directly from removing u_{0i} from Equations 14, 17, 19, and 22. Likewise, for binary outcomes (DGMs 3 and 4), the GEE versions of the four disaggregation methods are derived by removing u_{0i} from the multilevel logistic model equations (Equations 23–26).

Simulation Study

This simulation study has two primary objectives. First, we examine whether disaggregation methods commonly used in MLM generalize to GLMMs when applied to (a) binary predictors and (b) binary outcomes estimated using a logit link. Second, we assess whether GEEs require explicit separation of within- and between-person effects—via disaggrega-

tion—when contextual effects are present. Specifically, we investigate the impact of different design factors, such as sample size, number of measurement occasions, and variance components, on estimation bias in the fixed within and contextual effects. Simulations and model estimation were conducted in `R` (version 4.4.2; R Core Team, 2024). Parallel computing was used for efficiency via the `doFuture` (version 1.0.2; Bengtsson, 2021) and `foreach` (version 1.5.2; Daniel et al., 2022) packages. The code used for simulating and analyzing these data are provided as online supplementary materials at github.com/wardeiling/multilevel-vs-gee-binary.

Data Generation

Data were simulated with the four DGMs. Across all simulation scenarios, certain parameters were held constant: the fixed intercept was set to $\gamma_{00} = 0$, the within-cluster effect to $\beta_1 = 1.5$, the within-person standard deviation (SD) of the continuous predictor (DGMs 1 and 3) to $\sigma_{X,\text{w}} = 1$, and the level 1 residual SD (only set for DGMs 1 and 2) to $\sigma_e = 1$. Non-zero values were chosen for the random intercept residual SD σ_u, reflecting the assumption that the cluster mean of X does not fully explain stable differences in Y. Between-cluster variability in X was captured by $\sigma_{X,\text{b}}$ (for continuous X; DGMs 1 and 3) or σ_Z (for binary X; DGMs 2 and 4). We systematically varied key design factors across the four DGMs, as summarized in Table 1, excluding scenarios with a contextual effect but no between-cluster SD, as they are conceptually incoherent. This resulted in $432 - 96 = 336$ unique simulation conditions (84 per DGM), each replicated 1,000 times.

Table 1

Summary of Parameters for Each of the Four Data Generating Models

Factor	Notation	Applies to DGMs	Values
Sample size	N	All	100, 200
Number of time points	T	All	5, 10, 20

Within-cluster SD in continuous X	$\sigma_{X,\text{w}}$	1, 2	1
Between-cluster SD in continuous X	$\sigma_{X,\text{b}}$	2, 4	0, 1, 3
SD in Z (represents between-cluster variability in binary X)	σ_Z	3, 4	0, 1, 3
Fixed intercept	γ_{00}	All	0
Contextual effect	γ_{01}	All	0, 1, 3
Within-cluster effect	β_1	All	1.5
Random intercept residual SD (level 2)	σ_u	All	1, 3
Residual SD (level 1)	σ_e	1, 2	1

Note. SD refers to standard deviation and DGM to data generating model.

Data Analysis

Each simulated dataset was analyzed using two estimation frameworks: GLMMs and GEEs. GLMMs were estimated using the `lmer` and `glmer` functions from the `lme4` package (version 1.1-36; Bates et al., 2015) with full maximum likelihood estimation for `lmer`, as the focus was on fixed effects rather than variance components. GEEs were fitted using the `geeglm` function from the `geepack` package (version 1.3.12; Halekoh et al., 2006), with independent, exchangeable, and AR(1) structures. For GEEs, the maximum number of iterations was increased from the default 25 to 50 to support convergence. In summary, estimation was carried out using one GLMM and three GEE configurations, yielding four strategies in total.

Each of these four modeling strategies was implemented with the four distinct methods (UC, CWC, HB and MuCo) outlined for the MLM (see Equations 14, 17, 19). The application of these methods to both the multilevel logistic model and GEEs was discussed in the preceding sections. Since methods HB and MuCo produced virtually identical results across most scenarios, we chose to present method HB only in the online supplementary materials (see wardeiling.github.io/multilevel-vs-gee-binary/supplementary_materials.html), rather than

in the main results and figures. With four estimation strategies and three methods, we arrive at 12 analysis strategies. The GLMM implementations are labeled as GLMM-UC, GLMM-CWC, and GLMM-MuCo, while the GEE implementations are denoted as GEE-UC, GEE-CWC, and GEE-MuCo, with the correlation structure name appended.

The primary estimates of interest were the within-cluster effect—estimated as β_1^* in UC, β_{w} in CWC and β_1 in MuCo—and the contextual effect γ_{01} (only estimated with MuCo). Model performance was evaluated in terms of estimation bias, computed as $\frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} \left(\hat{\theta}_j - \theta\right)$, where $\hat{\theta}_j$ is the estimated parameter value for the j^{th} replication, θ is the true parameter value and n_{sim} denotes the number of replications. When relevant, we also examined convergence rates and estimation efficiency (i.e., variability across replications). All warnings and errors encountered during model estimation were logged. Models failing due to convergence issues or singular fits were excluded from the analysis. To ensure robustness of the reported results, we excluded parameter estimates exceeding ± 100, which occurred sporadically in GEEs with binary outcomes and were likely indicative of convergence failures. These extreme values were most frequent in AR(1) implementations using methods GEE-CWC and GEE-MuCo (see the Appendix for details).

Results

To facilitate interpretation, we focus on four main scenarios, holding the following constant: $N = 200$, $\sigma_e = 1$, $\gamma_{00} = 0$, $\beta_{\text{w}} = 1.5$, $\gamma_{01} = 3$, $\sigma_{X,\text{w}} = 1$ and $\sigma_{X,\text{b}} = 3$. We varied $\sigma_u = \{1, 3\}$ to assess the impact of unexplained heterogeneity, especially on GEEs, which does not explicitly model random effects; and $T = \{5, 20\}$ to evaluate performance under limited versus sufficient repeated measures. Sample size N had a negligible impact on average bias and is not discussed further.

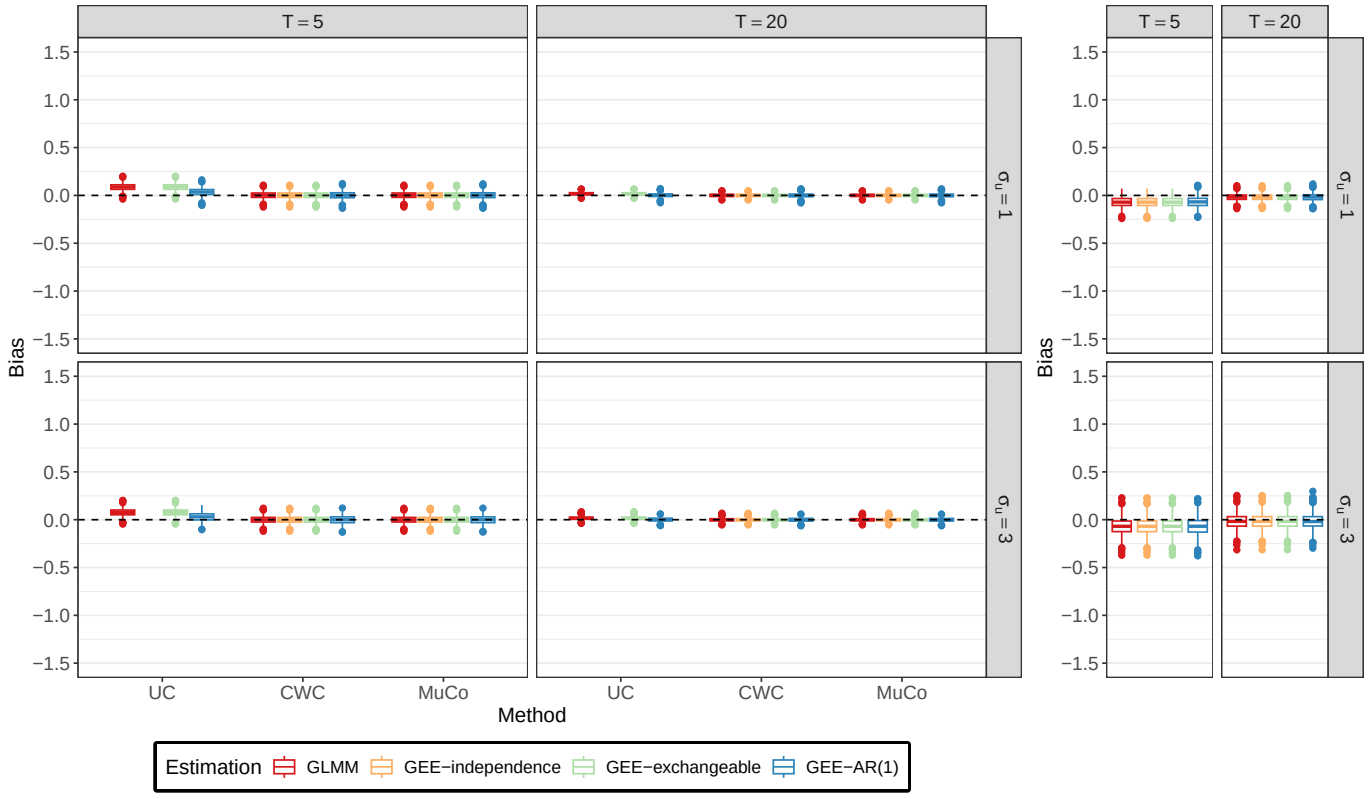
DGM 1: Continuous Predictor and Outcome. Figure 3 displays the bias in within-person and contextual effect estimates for benchmark DGM 1. As expected, Method UC yielded biased within-person estimates across all implementations, with bias and variability decreasing as T increased—consistent with the increasing dominance of within-person ef-

fect in the composite slope (as explained in the section on the MLM). GEEs with an exchangeable correlation (GEE-UC-exchangeable) closely mirrored the MLM estimates (GLMM-UC), which is unsurprising considering their equivalence under a linear model. The AR(1) structure yielded slightly lower bias (for some other scenarios this was reversed) and the independence structure yielded substantially more bias (out of bounds in Figure 3a). In contrast, disaggregation methods (Methods CWC and MuCo) resulted in tightly distributed unbiased within-person estimates across both frameworks among the correlation structures of GEEs. Contextual effect estimates (Method MuCo) were similarly unbiased and consistent across MLM and GEEs with a sufficient number of timepoints. However, an increase in unobserved heterogeneity resulted in lower precision across replications. In summary, all estimation frameworks and correlation structures can recover the within-person and contextual effects well in linear models when predictors and outcomes are continuous and a disaggregation method is applied.

DGM 2: Binary Predictor and Continuous Outcome. Figure 4 presents the results for DGM 2. Patterns were very similar to DGM 1, though with increased overall variability, impact of unexplained variability (σ_u) and more pronounced bias. The results for Method UC mirror those of DGM 1, except that in DGM 2 greater unexplained heterogeneity ($\sigma_u = 3$) reduced bias. This discrepancy stems from the bounded range of $\pi_{X,i} \in (0, 1)$ in contrast to the unbounded range of $\mu_{X,i} \in (-\infty, \infty)$ in DGM 1. The bounded scale compresses variance in $\pi_{X,i}$, inflating the ratio of unexplained-to-explained variance in β_{0i} (see Figure 2b). As before, disaggregation methods (CWC and MuCo) yielded unbiased within-person estimates across frameworks and GEEs correlation matrices. Contextual effect estimates were again similar across all four implementations using Method MuCo. However, unlike DGM 1, the contextual effect estimates were biased under a large number of timepoints ($T = 20$). Thus, when predictors are binary, disaggregation methods effectively recover the within-person effect across frameworks and GEEs correlation structures, but tend to yield similarly biased estimates of the contextual effect.

Figure 3

Bias for DGM 1 in Within-Person and Contextual Effect for Different Estimation Approaches



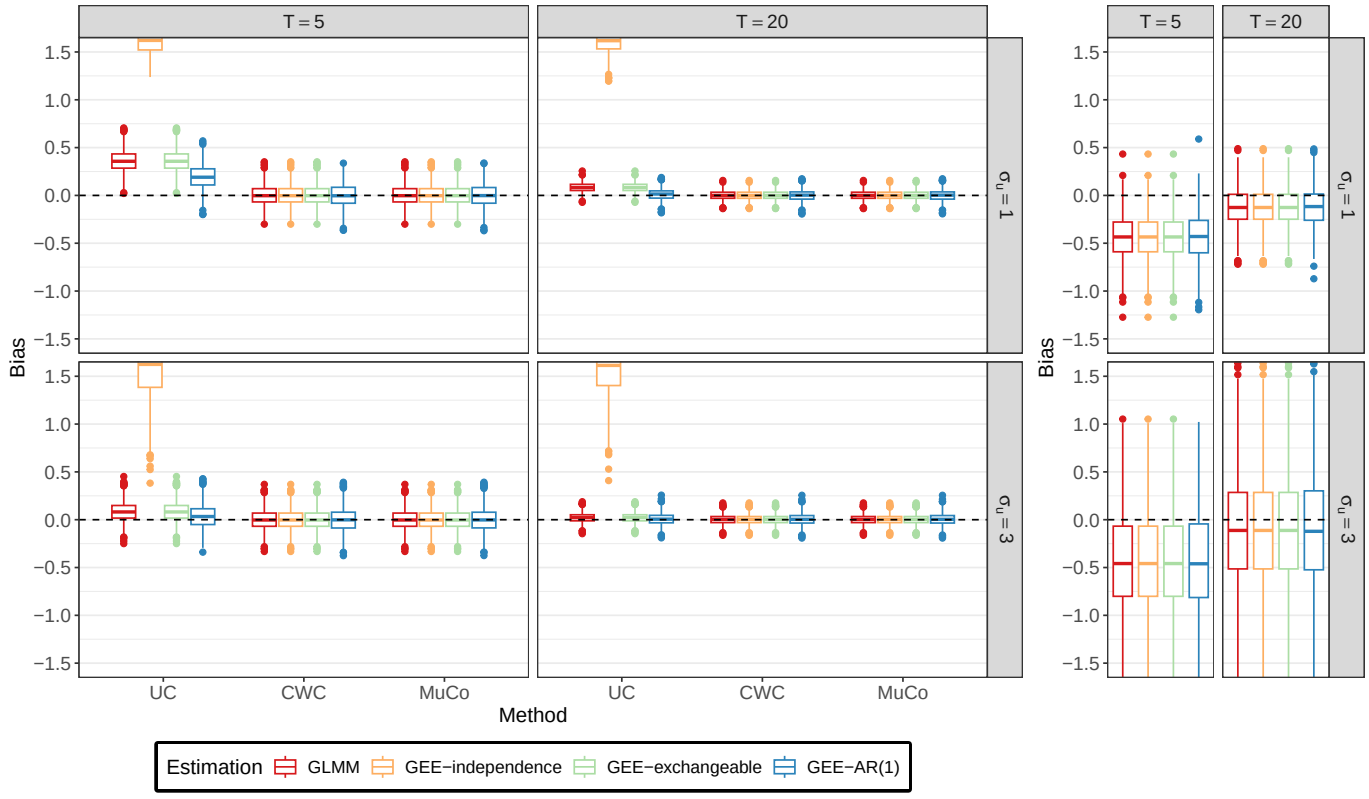
(a) *Within-Person Effect* β_1

(b) *Contextual Effect* γ_{01} *(Method MuCo)*

Note. The boxplot shows bias based on 1000 replications for data-generating model 1, which includes a continuous predictor and outcome. The boxes represent the interquartile range (IQR) from the 25th to 75th percentiles, with the median indicated by the horizontal line inside. Whiskers extend to 1.5 times the IQR, and replications outside this range are plotted as dots. UC = uncentered, CWC = centered within clusters, MuCo = Mundlak contextual, GLMM = generalized linear mixed model, GEE = generalized estimating equations. T represents the total number of time points, and σ_u denotes the random intercept residual variance. Y-axis breaks represent large intervals relative to bias but are consistent across data-generating models. In panel (a), GEE-UC with an independence correlation structure falls outside the plotted range, with estimates tightly clustered around 2.7 across all four scenarios.

Figure 4

Bias for DGM 2 in Within-Person and Contextual Effect for Different Estimation Approaches



(a) *Within-Person Effect* β_1

(b) *Contextual Effect* γ_{01} *(Method MuCo)*

Note. The boxplot shows bias based on 1000 replications for data-generating model 2, which includes a binary predictor and continuous outcome. The boxes represent the interquartile range (IQR) from the 25th to 75th percentiles, with the median indicated by the horizontal line inside. Whiskers extend to 1.5 times the IQR, and replications outside this range are plotted as dots. UC = uncentered, CWC = centered within clusters, MuCo = Mundlak contextual, GLMM = generalized linear mixed model, GEE = generalized estimating equations. T represents the total number of time points, and σ_u denotes the random intercept residual variance.

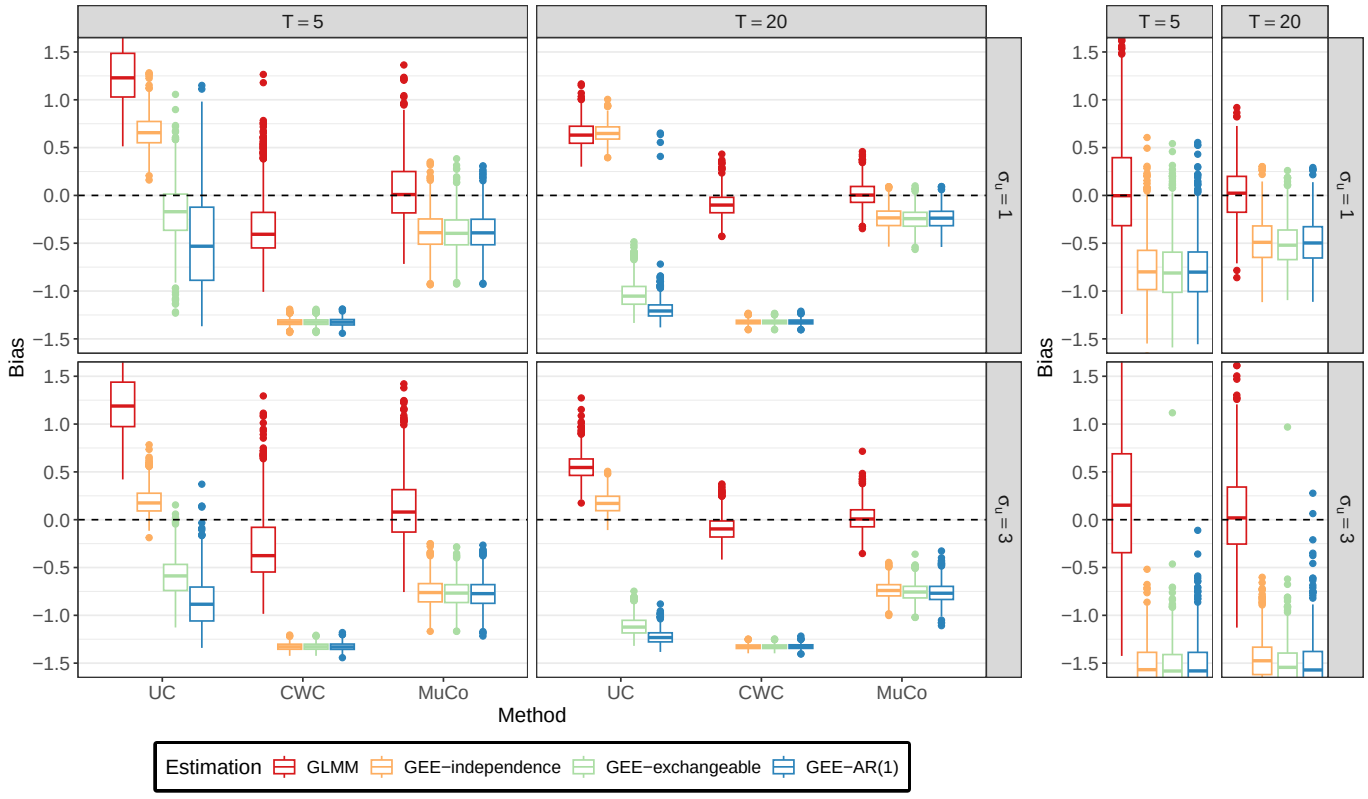
DGM 3: Continuous Predictor and Binary Outcome. Figure 5 shows the results for DGM 3. While higher T again reduced variability, patterns diverged from those in DGMs with continuous outcomes. Method UC showed increasing bias in GEEs models

as σ_u increased under exchangeable and AR(1) structures and an opposite trend under the independence structure. Notably, disaggregation performance varied across methods: CWC estimates were more biased than MuCo estimates across estimation approaches. More specifically, GLMM-CWC yielded biased within-person estimates even under optimal conditions ($T = 20$), whereas GLMM-MuCo yielded no bias under these circumstances. GEEs were biased for both disaggregation methods (GEE-CWC and GEE-MuCo), even under optimal conditions ($T = 20$ and $\sigma_u = 1$), though GEE-MuCo showed comparatively smaller bias. For contextual effects estimated by method MuCo, GLMM was clearly superior to GEEs across all conditions, especially when σ_u was large. As DGM 1, the contextual effects of GLMM where unbiased under a large number of timepoints, suggesting that scenarios with continuous predictors tend to be unbiased. These findings suggest that in logistic models, GEEs struggles to recover both within-person and contextual effects. Furthermore, within GLMM, Method MuCo is essential to avoid bias in the within-person effect.

DGM 4: Binary Predictor and Outcome. Figure 6 displays the results for DGM 4. Patterns were similar to DGM 3, reflecting the shared binary outcome. Method UC produced biased within-person estimates in all models, with the exception of GEE-independence under high unexplained heterogeneity ($\sigma_u = 3$)—a result that is difficult to explain. Unlike DGM 3, in the shown scenarios GLMM-CWC performed comparably well as GLMM-MuCo. However, across all scenarios, MuCo still outperformed CWC with a maximum bias of 0.07 and 0.13 respectively. As in DGM 3, all GEEs implementations with methods CWC and MuCo yielded biased results. However, unlike DGM 3, when unexplained heterogeneity was high, CWC and MuCo yielded similar estimates. For the contextual effect estimated by the MuCo method, GLMM consistently outperformed GEEs, where bias in GEEs increased as σ_u increased. As DGM 2, the contextual effects of GLMM where biased under a large number of timepoints, suggesting that scenarios with binary predictors tend to be biased. Overall, GLMM-MuCo was again the most consistent and robust method across conditions.

Figure 5

Bias for DGM 3 in Within-Person and Contextual Effect for Different Estimation Approaches



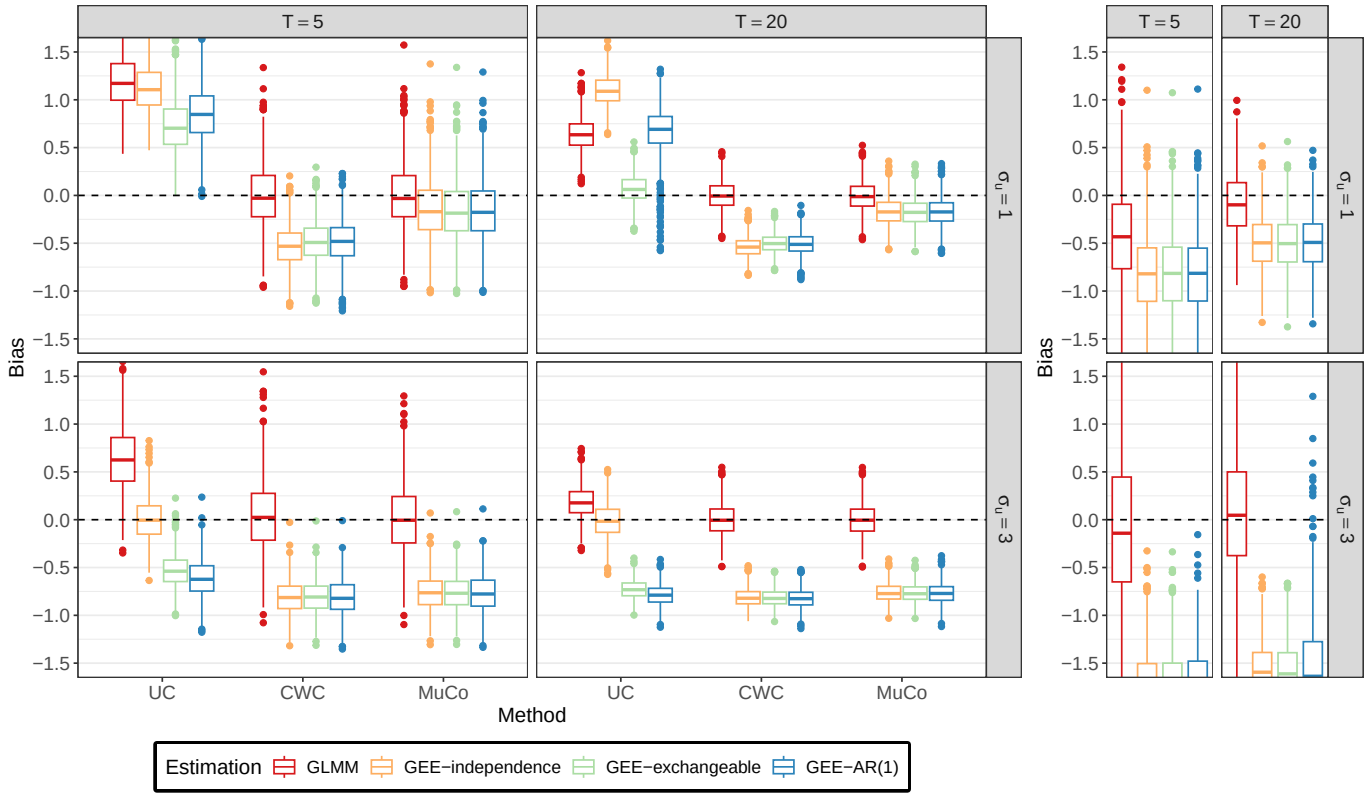
(a) *Within-Person Effect* β_1

(b) *Contextual Effect* γ_{01} *(Method MuCo)*

Note. The boxplot shows bias based on 1000 replications for data-generating model 3, which includes a continuous predictor and binary outcome. The boxes represent the interquartile range (IQR) from the 25th to 75th percentiles, with the median indicated by the horizontal line inside. Whiskers extend to 1.5 times the IQR, and replications outside this range are plotted as dots. UC = uncentered, CWC = centered within clusters, MuCo = Mundlak contextual, GLMM = generalized linear mixed model, GEE = generalized estimating equations. T represents the total number of time points, and σ_u denotes the random intercept residual variance.

Figure 6

Bias for DGM 4 in Within-Person and Contextual Effect for Different Estimation Approaches



(a) *Within-Person Effect* β_1

(b) *Contextual Effect* γ_{01} *(Method MuCo)*

Note. The boxplot shows bias based on 1000 replications for data-generating model 4, which includes a binary predictor and outcome. The boxes represent the interquartile range (IQR) from the 25th to 75th percentiles, with the median indicated by the horizontal line inside. Whiskers extend to 1.5 times the IQR, and replications outside this range are plotted as dots. UC = uncentered, CWC = centered within clusters, MuCo = Mundlak contextual, GLMM = generalized linear mixed model, GEE = generalized estimating equations. T represents the total number of time points, and σ_u denotes the random intercept residual variance.

Discussion

This study addressed an important gap in the literature on clustered longitudinal data analysis by extending the well-known within- and between-person effects debate beyond the MLM framework to GLMMs and GEEs with different variable types. First, we generalized the established distinction between within- and between-person slopes—traditionally discussed for continuous variables—to settings involving binary predictors and/or outcomes. Second, we evaluated how well each disaggregation method, implemented in GLMMs and GEEs, recovered the within-person effect and contextual effect (i.e., the difference between the between- and within-person slopes). To assess the role of variable scale, analyses were conducted across combinations of binary and continuous predictor and outcomes. Although prior work has examined binary predictors in MLMs and binary outcomes in GLMMs, and some studies have compared estimation frameworks across disciplines, no study has brought these strands together. This article fills that gap by offering a systematic evaluation of disaggregation methods across estimation strategies and data types.

Main Findings

The simulation results demonstrate that the effectiveness of disaggregation methods depends on both the estimation framework and the type of variables involved. As expected, when contextual effects were present, the uncentered predictor approach consistently failed to recover the within-person effect, both in the GLMM and across all three GEE variants. This failure was particularly pronounced for scenarios involving (a) a small number of time points and (b) binary predictors or outcomes, where bias and variability were substantially larger than in continuous-variable settings. Furthermore, with continuous outcomes (DGMs 1 and 2), all three disaggregation methods (CWC, HB, MuCo) performed comparably well across the GLMM and GEE variants. However, this pattern did not generalize to binary outcomes (DGMs 3 and 4). None of the GEE variants consistently recovered the within-person or contextual effects. In the GLMM context, the hybrid or Mundlak methods produced robust estimates, whereas the CWC approach was more susceptible to bias—especially with

continuous predictors. When comparing results across predictor types, we found that GLMM estimates of the contextual effect were more biased for binary predictors (DGMs 2 and 4) than for continuous predictors (DGMs 1 and 3). Taken together, these results point to consistent differences in performance across methods, designs and variable types.

These findings yield several insights. First, although the person-mean centered-only approach (Method CWC) is often treated as the default or even "gold standard" for estimating within-person effects in multilevel modeling (e.g., Enders & Tofighi, 2007; Hamaker & Muthén, 2020; Raudenbush, 2009), our results indicate that this strategy may yield biased estimates when applied to binary outcomes. This is particularly concerning given the widespread use of this method in studies with continuous outcomes, where researchers may be inclined to extend the same approach to non-continuous settings. Our findings suggest that such a transfer is not without risk. To mitigate bias, we found that it is crucial to include the person-mean of the predictor as a predictor of the random intercept, as is done in Mundlak's contextual and the hybrid method. This discrepancy must stem from differences in the statistical properties of the identity and logit link functions. Specifically, Neuhaus and Jewell (1993, p. 807) demonstrated that omitting relevant predictors in logistic models can lead to substantial bias in the estimated coefficients, even when the omitted variables are independent from those included. In the CWC formulation, the person-mean is omitted from the model, thereby violating this condition and potentially introducing bias.

Second, our findings contribute to the ongoing debate on the interpretability of GEEs by clarifying a key point of contention. In line with earlier work (Begg & Parides, 2003), we demonstrate that for continuous outcomes, GEEs—when paired with a disaggregation approach—yield estimates of within-person and contextual effects that closely mirror those from GLMMs. This equivalence is not incidental but follows directly from statistical theory: With an identity link and normally distributed outcomes, marginal (GEE) and conditional (GLMM) estimators coincide (Neuhaus et al., 1991; Zeger et al., 1988). Yet, in psychological treatments, GEEs are often characterized as producing population-averaged estimates

that preclude person-specific interpretations (e.g., Ballinger, 2004; Bauer & Sterba, 2011; McNeish et al., 2017). However, it should be nuanced that with continuous outcomes, individual-level interpretations remain valid, allowing for the estimation of within-person parameters.

The picture changes substantially for binary outcomes. Our simulations show that GEEs, even when combined with a disaggregation approach, cannot recover the true within-person and contextual effects. The degree of bias increases with the residual variance of the random intercept, aligning with prior work demonstrating that non-linear link functions—such as the logistic link—break the equivalence between GLMM and GEE estimates (Neuhaus et al., 1991; Zeger et al., 1988). This divergence arises because the random intercept, which captures outcome heterogeneity due to unobserved covariates, is explicitly modeled in GLMMs but left unaccounted for in GEEs (For an illustration of the impact of the random intercept residual variance on this discrepancy, see wardeiling.shinyapps.io/GLMM_population-averaged-and-person-specific-interpretations). As a consequence, GEE estimates are attenuated toward zero (Neuhaus et al., 1991; Zeger et al., 1988). This is visible in our results, where the GEE estimates fell consistently below the zero-bias line. In the context of positive true effects, this implies an underestimation of the effect magnitude. Accordingly, in the context of binary outcomes, Neuhaus et al. (1991) pointed out that GEEs "cannot provide estimates of changes within individuals over time; these are often quantities of central interest in longitudinal studies" (p. 33). For research questions that focus on within-person processes, as is often the case in psychology, such attenuation renders the logit-linked GEE framework ill-suited. In these contexts, conditional models such as GLMMs are generally more appropriate (Allison, 1999, p. 78).

Third, our findings highlight limitations in estimating contextual effects when relying on observed person-means within the GLMM framework. Across most simulation conditions, we observed small but consistent bias in the contextual estimates. One plausible contributor is Lüdtke's bias: the sample-based person-mean is an imperfect proxy for the true latent

mean, especially when the number of observations per person is limited (Lüdtke et al., 2008). In line with prior work on the limitations of disaggregation using observed means (Asparouhov & Muthén, 2019), the magnitude of bias increased under stronger contextual effects, reduced between-person variance, and smaller cluster sizes. Notably, this bias was substantially more pronounced for binary predictors (DGMs 2 and 4) than for continuous ones (DGMs 1 and 3), underscoring the additional complexities introduced when modeling non-continuous predictors. While Lüdtke's bias offers a partial explanation, other sources likely contribute as well. As discussed in the section on the MLM, binary predictors pose unique challenges such as truncated distributions and violations of independence (see also Asparouhov & Muthén, 2019). In the DGMs, the binary predictor was conceptualized as the observed expression of a latent continuous construct. If this is the case, we recommend the use of latent centering approaches of multilevel structural equation modeling (e.g., Asparouhov & Muthén, 2019), which provide more robust estimates, particularly in the presence of binary predictors.

Finally, an unanticipated but important finding concerns the instability of GEE estimation when disaggregation methods are applied. Although GEEs are frequently commended for their robustness to misspecification of the working correlation structure (e.g., Ballinger, 2004; Zeger et al., 1988), our simulations revealed instances of extreme and erratic estimates under several conditions. These estimation anomalies were specific to binary outcomes and most prevalent when the AR(1) working correlation structure was used—rarely occurring under an exchangeable structure and absent under independence (see the Appendix for details). This suggests that the combination of a logit link with complex correlation structures, particularly when misspecified, may exacerbate numerical instability in GEEs.

Limitations and Future Directions

This study has several limitations that point to important directions for future research. First, our simulations examined the recoverability of within-person effects under DGMs that satisfy the core assumptions of GLMMs and GEEs (see McNeish et al., 2017).

In practice, however, these assumptions are often violated, potentially leading to biased estimates. Within biomedical applications, GEEs are sometimes preferred for their robustness to distributional assumptions about random effects (Hubbard et al., 2010), yet it remains unclear whether this robustness generalizes to the retrieval of within-person effects. Future studies should explicitly examine how violations—such as non-normal random effects or misspecified variance structures—impact estimation, and whether the equivalence of GEE and GLMM estimates for continuous outcomes persists under such conditions.

Second, while our focus was on random intercept models, the inclusion of random slopes is a central strength of the multilevel framework (Bell & Jones, 2015). In longitudinal data, researchers are often interested in capturing within-person effects while allowing for individual differences in the strength of these effects (e.g., Geschwind et al., 2011). Assuming a homogeneous within-person association across individuals can be overly restrictive, as many psychological processes plausibly vary from person to person. The inclusion of a random slope would alter the application of disaggregation methods in GLMM estimation, as the equivalence between the hybrid model and Mundlak's contextual model is lost under this specification (Snijders & Bosker, 2011). Additionally, since the GEE framework does not account for random slope heterogeneity, it remains an open question whether disaggregation-based GEEs can still recover within-person effects when such heterogeneity is present. Therefore, future research should extend the DGMs to include random slopes, allowing for a more comprehensive evaluation of GLMM methods and the suitability of GEE for investigating within-person effects in such contexts.

Third, we restricted our simulations to binary predictors and outcomes. Although prior work has begun to explore centering strategies for multi-categorical predictors (e.g., Yaremych et al., 2023), applications involving multi-categorical outcomes remain largely unexamined. Expanding the current framework to accommodate these outcome types represents a critical step for broadening the generalizability of disaggregation methods.

Conclusion

Within the multilevel modeling literature, the importance of distinguishing within- from between-cluster effects has been stressed repeatedly and by many (e.g., Enders & Tofighi, 2007; Kreft et al., 1995; Raudenbush & Bryk, 2002), to the point that it has become standard practice among psychological researchers. However, the applicability of disaggregation methods to multilevel models with non-continuous predictors and outcomes, as well as GEEs, has not been systematically examined This study offers the first comprehensive account of how the within- and between-person effects debate extends to binary predictors, binary outcomes, and the GEEs framework. Based on our findings, we recommend using GLMMs with Mundlak's contextual or the hybrid approach when estimating within-person effects regardless of variable type.

References

Allison, P. D. (1999). *A Logistic Regression Using SAS: Theory and Application* (1st ed.). Cary, NC: SAS Institute Inc.

Anderson, D. A., & Aitkin, M. (1985). Variance Component Models with Binary Response: Interviewer Variability. *Journal of the Royal Statistical Society: Series B (Methodological)*, *47*(2), 203–210. https://doi.org/10.1111/j.2517-6161.1985.tb01346.x

Asparouhov, T., & Muthén, B. (2019). Latent Variable Centering of Predictors and Mediators in Multilevel and Time-Series Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 119–142. https://doi.org/10.1080/10705511.2018.1511375

Austin, P. C., & Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. *Statistics in Medicine*, *36*(20), 3257–3277. https://doi.org/10.1002/sim.7336

Baer, R. A., & Sauer, S. E. (2011). Relationships between depressive rumination, anger rumination, and borderline personality features. *Personality Disorders: Theory, Research, and Treatment*, *2*(2), 142–150. https://doi.org/10.1037/a0019478

Ballinger, G. A. (2004). Using Generalized Estimating Equations for Longitudinal Data Analysis. *Organizational Research Methods*, *7*(2), 127–150. https://doi.org/10.1177/1094428104263672

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

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, *16*(4), 373–390. https://doi.org/10.1037/a0025813

Begg, M. D., & Parides, M. K. (2003). Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine*, *22*(16), 2591–2602. https://doi.org/10.1002/sim.1524

Bell, A., & Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, *3*(1), 133–153. https://doi.org/10.1017/psrm.2014.7

Bengtsson, H. (2021). A unifying framework for parallel and distributed processing in R using futures. *The R Journal*, *13*(2), 208–227. https://doi.org/10.32614/RJ-2021-048

Berlin, J. A., Kimmel, S. E., Ten Have, T. R., & Sammel, M. D. (1999). An Empirical Comparison of Several Clustered Data Approaches Under Confounding Due to Cluster Effects in the Analysis of Complications of Coronary Angioplasty. *Biometrics*, *55*(2), 470–476. https://doi.org/10.1111/j.0006-341X.1999.00470.x

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*. Guilford Press.

Borders, A., Earleywine, M., & Jajodia, A. (2010). Could mindfulness decrease anger, hostility, and aggression by decreasing rumination? *Aggressive Behavior*, *36*(1), 28–44. https://doi.org/10.1002/ab.20327

Breslow, N. E., & Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, *88*(421), 9–25. https://doi.org/10.1080/01621459.1993.10594284

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*(4), 822–848. https://doi.org/10.1037/0022-3514.84.4.822

Curran, P. J., & Bauer, D. J. (2011). The Disaggregation of Within-Person and Between-Person Effects in Longitudinal Models of Change. *Annual Review of Psychology*, *62*(Volume 62, 2011), 583–619. https://doi.org/10.1146/annurev.psych.093008.100356

Daniel, F., Ooi, H., Calaway, R., Microsoft, & Weston, S. (2022). Foreach: Provides Foreach Looping Construct. Retrieved April 9, 2025, from https://cran.r-project.org/web/packages/foreach/index.html

Diggle, P., Heagerty, P., Liang, K.-Y., & Zeger, S. (2002, June). *Analysis of Longitudinal Data*. OUP Oxford.

Eisenlohr-Moul, T. A., Peters, J. R., Pond, R. S., & DeWall, C. N. (2016). Both Trait and State Mindfulness Predict Lower Aggressiveness via Anger Rumination: A Multilevel Mediation Analysis. *Mindfulness*, *7*(3), 713–726. https://doi.org/10.1007/s12671-016-0508-x

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Gardiner, J. C., Luo, Z., & Roman, L. A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine*, *28*(2), 221–239. https://doi.org/10.1002/sim.3478

Geschwind, N., Peeters, F., Drukker, M., van Os, J., & Wichers, M. (2011). Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *79*(5), 618–628. https://doi.org/10.1037/a0024595

Halekoh, U., Højsgaard, S., & Yan, J. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, *15/2*, 1–11.

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

Hardin, J. W., & Hilbe, J. M. (2012). *Generalized Estimating Equations*. CRC Press.

Hoffman, L. (2015). *Longitudinal Analysis: Modeling Within-Person Fluctuation and Change*. Routledge.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Laan, M. V. d., Lippman, S. A., Jewell, N., Bruckner, T., & Satariano, W. A. (2010). To GEE or Not to GEE: Comparing Population Average and Mixed Models for Estimating the Associations Between Neighborhood

Risk Factors and Health. *Epidemiology*, *21*(4), 467. https://doi.org/10.1097/EDE.0b013e3181caeb90

Kashdan, T. B., Goodman, F. R., Mallard, T. T., & DeWall, C. N. (2016). What Triggers Anger in Everyday Life? Links to the Intensity, Control, and Regulation of These Emotions, and Personality Traits. *Journal of Personality*, *84*(6), 737–749. https://doi.org/10.1111/jopy.12214

Kiken, L. G., Garland, E. L., Bluth, K., Palsson, O. S., & Gaylord, S. A. (2015). From a state to a trait: Trajectories of state mindfulness in meditation during intervention predict changes in trait mindfulness. *Personality and Individual Differences*, *81*, 41–46. https://doi.org/10.1016/j.paid.2014.12.044

Koper, N., & Manseau, M. (2009). Generalized estimating equations and generalized linear mixed-effects models for modelling resource selection. *Journal of Applied Ecology*, *46*(3), 590–599. https://doi.org/10.1111/j.1365-2664.2009.01642.x

Kreft, I. G., de Leeuw, J., & Aiken, L. S. (1995). The Effect of Different Forms of Centering in Hierarchical Linear Models. *Multivariate Behavioral Research*, *30*(1), 1–21. https://doi.org/10.1207/s15327906mbr3001_1

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229. https://doi.org/10.1037/a0012869

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140. https://doi.org/10.1037/met0000078

Merlo, J., Chaix, B., Ohlsson, H., Beckman, A., Johnell, K., Hjerpe, P., Råstam, L., & Larsen, K. (2006). A brief conceptual tutorial of multilevel analysis in social epidemiology:

Using measures of clustering in multilevel logistic regression to investigate contextual phenomena. *Journal of Epidemiology & Community Health*, *60*(4), 290–297. https://doi.org/10.1136/jech.2004.029454

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica*, *46*(1), 69–85. https://doi.org/10.2307/1913646

Muth, C., Bales, K. L., Hinde, K., Maninger, N., Mendoza, S. P., & Ferrer, E. (2016). Alternative Models for Small Samples in Psychological Research: Applying Linear Mixed Effects Models and Generalized Estimating Equations to Repeated Measures Data. *Educational and Psychological Measurement*, *76*(1), 64–87. https://doi.org/10.1177/0013164415580432

Neuhaus, J. M., & Kalbfleisch, J. D. (1998). Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data. *Biometrics*, *54*(2), 638–645. https://doi.org/10.2307/3109770

Neuhaus, J. M., Kalbfleisch, J. D., & Hauck, W. W. (1991). A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data. *International Statistical Review / Revue Internationale de Statistique*, *59*(1), 25–35. https://doi.org/10.2307/1403572

Neuhaus, J. M., & Jewell, N. P. (1993). A Geometric Approach to Assess Bias Due to Omitted Covariates in Generalized Linear Models. *Biometrika*, *80*(4), 807–815. https://doi.org/10.2307/2336872

R Core Team. (2024). *R: A language and environment for statistical computing* (manual). R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raudenbush, S. W. (2009). Adaptive Centering with Random Effects: An Alternative to the Fixed Effects Model for Studying Time-Varying Treatments in School Settings. *Education Finance and Policy*, *4*(4), 468–491. https://doi.org/10.1162/edfp.2009.4.4.468

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd). SAGE.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*, 351–357. https://doi.org/10.2307/2087176

Schunck, R., & Perales, F. (2017). Within- and Between-cluster Effects in Generalized Linear Mixed Models: A Discussion of Approaches and the Xthybrid command. *The Stata Journal*, *17*(1), 89–115. https://doi.org/10.1177/1536867X1701700106

Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE.

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-Effects Models for Serial Observations with Binary Response. *Biometrics*, *40*(4), 961–971. https://doi.org/10.2307/2531147

Stroup, W. W. (2012). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press.

Yan, J., Aseltine, R. H., & Harel, O. (2013). Comparing Regression Coefficients Between Nested Linear Models for Clustered Data With Generalized Estimating Equations. *Journal of Educational and Behavioral Statistics*, *38*(2), 172–189. https://doi.org/10.3102/1076998611432175

Yaremych, H. E., Preacher, K. J., & Hedeker, D. (2023). Centering categorical predictors in multilevel models: Best practices and interpretation. *Psychological Methods*, *28*(3), 613–630. https://doi.org/10.1037/met0000434

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, *42*(1), 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*(4), 1049–1060. https://doi.org/10.2307/2531734

Appendix

Handling of Extreme Parameter Estimates in GEEs

Diagnostic evaluation of extreme parameter estimates is reported in the supplemental materials at wardeiling.github.io/multilevel-vs-gee-binary/supplementary_materials.html. This revealed that, when modeling binary outcomes, the `geeglm` function (Halekoh et al., 2006) occasionally produced bias estimates of implausibly large magnitude (e.g., 10^{12} or higher), likely reflecting non-convergence. Such extreme values consistently exceeded $\pm 10^{11}$, while the vast majority of estimates fell within a narrow range of ± 2.5 units around the mean bias. Notably, the number of extreme estimates remained unchanged when using a threshold of ± 20, ± 100 or $\pm 10^{11}$, indicating that non-convergence resulted in extreme rather than moderate deviations from the true parameter values. Across 76 simulation scenarios, we observed at least one replication with an estimate exceeding $\pm 10^{11}$. Extreme estimates were specific to binary outcomes and most prevalent when the AR(1) working correlation structure was paired with a disaggregation method (methods CWC, HB and MuCo)—rarely occurring under an exchangeable structure and absent under independence. In one scenario, approximately half of the replications produced extreme values. To maintain the integrity of the reported results, we excluded all estimates exceeding ± 100 from the final analyses.