Received: 13 April 2022 | Revised: 24 February 2023 | Accepted: 18 March 2023

DOI: 10.1002/bimj.202200112

Biometrical Journal

RESEARCH ARTICLE

# A comparison of the multilevel MIMIC model to the multilevel regression and mixed ANOVA model for the estimation and testing of a cross-level interaction effect: A simulation study

**Rob Kessels**[1] | **Mirjam Moerbeek**[2]

[1]Department of Biometrics, Netherlands Cancer Institute, Amsterdam, The Netherlands

[2]Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

**Correspondence**
Mirjam Moerbeek, Department of Methodology and Statistics, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands.
Email: m.moerbeek@uu.nl

## Abstract

When observing data on a patient-reported outcome measure in, for example, clinical trials, the variables observed are often correlated and intended to measure a latent variable. In addition, such data are also often characterized by a hierarchical structure, meaning that the outcome is repeatedly measured within patients. To analyze such data, it is important to use an appropriate statistical model, such as structural equation modeling (SEM). However, researchers may rely on simpler statistical models that are applied to an aggregated data structure. For example, correlated variables are combined into one sum score that approximates a latent variable. This may have implications when, for example, the sum score consists of indicators that relate differently to the latent variable being measured. This study compares three models that can be applied to analyze such data: the multilevel multiple indicators multiple causes (ML-MIMIC) model, a univariate multilevel model, and a mixed analysis of variance (ANOVA) model. The focus is on the estimation of a cross-level interaction effect that presents the difference over time on the patient-reported outcome between two treatment groups. The ML-MIMIC model is an SEM-type model that considers the relationship between the indicators and the latent variable in a multilevel setting, whereas the univariate multilevel and mixed ANOVA model rely on sum scores to approximate the latent variable. In addition, the mixed ANOVA model uses aggregated second-level means as outcome. This study showed that the ML-MIMIC model produced unbiased cross-level interaction effect estimates when the relationships between the indicators and the latent variable being measured varied across indicators. In contrast, under similar conditions, the univariate multilevel and mixed ANOVA model underestimated the cross-level interaction effect.

**KEYWORDS**
cross-level interaction, MIMIC, multilevel analysis, patient-reported outcomes, structural equation modeling, sum scoring

## 1 | INTRODUCTION

This simulation study compared three different statistical models on their performance in estimating a cross-level interaction effect in the presence of correlated data. These models are the multilevel multiple indicator multiple causes (ML-MIMIC) model, the univariate multilevel regression model, and the mixed analysis of variance (ANOVA) model. The ML-MIMIC model is a structural equation model (SEM) (Bollen, 1989) and SEMs form a flexible modeling framework to model complex (multivariate) correlated data in order to explain the relationship among a number of observed variables and latent variables. In general, SEMs consist of a measurement model part and a latent variable model part. The measurement model is a confirmatory factor analysis model where the covariance among observed variables is analyzed to investigate how well the observed variables measure a prespecified number of latent variables and where measurement error is taken into account (Brown, 2014). The latent variable model is a regression model where the effects of independent latent variables or independent observed variables are assessed on dependent latent variables or dependent observed variables. The ML-MIMIC model is an SEM that combines both model parts (Jöreskog & Goldberger, 1975; Muthén, 1989) in the presence of a hierarchical data structure (Mehta & Neale, 2005).

The ML-MIMIC model can be regarded as an alternative to a univariate multilevel regression model (Hox et al., 2018; Singer & Willett, 2003) (often referred to as mixed model in the medical literature (Brown & Prescott, 2015)) where the outcome represents the sum score of a number of correlated items from a questionnaire. When this outcome is measured at multiple occasions over time nested within patients, the univariate multilevel regression then models the effect over time within patients on this univariate outcome by including time as within-patient-level covariate. By adding covariates at the between-patient-level predicting the univariate outcome at the patient level, differences in this time effect can be assessed between patients by including a cross-level interaction effect. The latent variable model part of the ML-MIMIC model works in a similar way by including covariates at the within-patient and between-patient level predicting the latent variable outcome at both levels and a cross-level interaction can be added. However, in contrast to using the sum score to estimate the latent variable in a univariate multilevel regression, the measurement model part of the ML-MIMIC model models the actual relationships between the questionnaire items and the latent variable and it accounts for measurement error, thereby precluding the need for sum scoring a-priori. This will lead to less biased latent variable estimates as compared to sum scores (Bollen, 1989).

Summation of questionnaire items is a way of aggregating data. An additional aggregation step to (multivariate) correlated hierarchical data is the derivation of patient-level means of a univariate outcome out of multiple repeated observations. In such cases, a mixed ANOVA model can be used to analyze group differences on repeated patient-level means. The ML-MIMIC model, the univariate multilevel regression model, and the mixed ANOVA model can all be used to estimate the cross-level interaction effect, but there is a difference in the way the data are treated by applying aggregation. In this simulation study, we will investigate under what circumstances data aggregation becomes problematic in the presence of correlated variables (i.e., questionnaire items) that measure a latent variable in a multilevel context by comparing the three models. This simulation study is motivated by an applied example where data on a patient reported outcome measure were collected as part of a clinical trial.

The trial that served as the motivation was a double-blind, placebo-controlled, randomized clinical trial investigating the efficacy of the on-demand use of the combined administration of testosterone and sildenafil (T+S), compared to placebo, in American women diagnosed with hypoactive sexual desire disorder (HSDD; which currently is part of the diagnosis female sexual interest/arousal disorder [FSIAD]) caused by low sensitivity of the brain for sexual cues (Trial registration: ClinicalTrials.gov: ID: NCT01432665) (Tuiten et al., 2018). During the trial, patients were instructed to take the medication prior to an anticipated sexual event and report about the event using the validated Sexual Event Diary (SED) (Van Nes et al., 2017). The SED is a web-based questionnaire that patients had to fill out within 24 h following a sexual event. For each sexual event, patients had to indicate the level of pleasure, inhibition ("ability to let yourself go"), sexual desire, bodily arousal, and subjective arousal on five-point Likert scale items (1 = *not at all*, 5 = *totally*). These items intend to measure the latent variable sexual functioning, the outcome of interest (Kessels et al., 2019, 2021; Van Nes et al., 2017).

Furthermore, in this trial, all patients started with a 4-week baseline establishment (BLE) period where no medication was used by any patient. After the BLE period, patients were randomly assigned to an active drug (T+S) or placebo treatment and continued with an 8-week active treatment period (ATP) where patients used the drug they were randomly assigned to. These design elements, study period (BLE vs. ATP), and treatment group (placebo vs. T+S) were included as covariates (explanatory variables for the outcome sexual functioning), where study period was a within-patient-level covariate and treatment group a between-patient-level covariate. The cross-level interaction effect between study period and treatment group on the outcome sexual functioning was of primary interest.

In the previous work, we compared the application of the ML-MIMIC model, the univariate multilevel regression model, and the mixed ANOVA model on the clinical trial data. For the univariate multilevel regression model and mixed ANOVA model, the items were combined into a sum score to estimate the latent variable sexual functioning at each event, whereas in the ML-MIMIC model, the items were included as indicators in a multilevel latent variable model. For the mixed ANOVA model, the sum score per event was averaged over all events observed during the BLE period and all events during the ATP, thereby creating two patient-level mean scores. The three models were compared on the cross-level interaction effect between study period and treatment group on the latent construct sexual functioning (Kessels et al., 2019, 2021), and the results of this work indicated that the parameter estimates of the cross-level interaction effect across these three different models were very comparable. Part of this result could be explained by the fact that the relationship between the items and the latent variable were very much alike across items, as proven by the nearly identical factor loadings.

Although the sum score is a popular method for applied researchers and clinicians to approximate a latent variable (DiStefano et al., 2009), the use of sum scores should carefully be considered. By employing sum scoring, it is assumed that all items have equal weight in relation to the latent variable (McNeish & Wolf, 2020). When it turns out that items have varying weights, caution is warranted. When items relate differently to the latent variable, it means that two subjects having identical sum scores may have different latent variable scores. This can have implications when the sum score of a (patient reported) outcome is used as outcome in a clinical trial to assess treatment efficacy. For example, if the covariate represents treatment group membership and two patients from different groups have equal sum scores, but different latent variable scores, this difference will not be detected when using sum scores to approximate the clinical outcome. Also, the aggregation step that creates patient-level means out of multiple repeated observations may be problematic. One drawback is the potential loss in power, because aggregation leads to a loss of information.

The potential issues with aggregating the sexual event data have led to the motivation of proposing the ML-MIMIC model for analyzing such data (Kessels et al., 2021), because the ML-MIMIC model does not require aggregated data. ML-MIMIC models have been gaining popularity in educational and psychological research (Davidov et al., 2019; Jak et al., 2014; Roesch et al., 2010), but these models are less commonly used for analyzing patient reported outcomes in clinical trials. Examples where the ML-MIMIC model could offer a promising alternative for analyzing the data are clinical trials that study the impact of different treatments on health-related quality of life outcomes and other patient-reported outcomes, such as oncology trials (Osoba, 2002). For example, in Reck et al. (2018) and in Pompili et al. (2018), quality of life summary scales were compared between two different treatment groups in patients with advanced squamous non–small cell lung cancer. In both examples, the scales were derived by creating sum scores and univariate multilevel regressions were used to compare the scales between different groups of patients. However, the application of the ML-MIMIC model for analyzing patient-reported outcomes is just one of many potential use cases, which is why the results of this study can easily be translated to other research areas.

To our knowledge, there have been no studies where the performance of the ML-MIMIC model in estimating a cross-level interaction effect was compared to the univariate multilevel regression and mixed ANOVA. Cao et al. (2019) did examine the performance of the ML-MIMIC model in estimating a cross-level interaction effect using a simulation study. Although that study can be considered the first study to examine the cross-level interaction between two covariates on a latent factor in an SEM context, the study did not made a comparison between the ML-MIMIC model and other models. Furthermore, in the simulation study by Cao et al. (2019), the factor loadings were not varied across items, which will be of particular interest in this study as we expect that it will have an impact on the accuracy in estimating a latent variable using sum scoring or latent variable modeling. Finch and French (2011) conducted a simulation study to assess the performance of the MIMIC model in the presence of multilevel data. The authors compared the ML-MIMIC model to the standard MIMIC model that ignored the multilevel data structure. This study found that it is important to consider the multilevel data structure, but this simulation study was limited by the fact that it did not include a cross-level interaction effect. The current simulation study can be seen as an extension of the work by Cao et al. (2019) and Finch and French (2011).

## 2 | METHODS

In this section, we will first present the three different models along with their notations. The models will be presented considering the applied example introduced in Section 1. Then, we discuss the simulation study design, the conditions varied during the simulations, and how the models were fitted. Section 2 is ended by explaining the criteria that were used to evaluate the three models.

### 2.1 | ML-MIMIC model

A MIMIC model describes the linear relation between observed variables and latent factors in a measurement model and a latent variable model. In an ML-MIMIC model with two-level data, these relationships are defined at the within-level and between-level, where the relationships at the within-level are allowed to vary across the between-level. Based on the sexual event data described in the previous section, consider the multivariate response vector $\boldsymbol{Y}_{ei}$ containing five observed item scores on event $e$ for patient $i$, intended to measure one latent variable at the within-patient and between-patient level. The two-level measurement model can then be written as:

$$\boldsymbol{Y}_{ei} = \boldsymbol{\mu} + \boldsymbol{\Lambda}_w \eta_{ei(w)} + \boldsymbol{\Lambda}_b \eta_{i(b)} + \boldsymbol{\epsilon}_{ei(w)} + \boldsymbol{\epsilon}_{i(b)}. \tag{1}$$

The observed scores $\boldsymbol{Y}_{ei}$ are predicted by a regression equation involving a vector of between-level intercepts $\boldsymbol{\mu}$, the within-patient factor loading matrix $\boldsymbol{\Lambda}_w$ multiplied by the within-patient latent variables $\eta_{ei(w)}$ for event $e$, the between-patient factor loading matrix $\boldsymbol{\Lambda}_b$ multiplied by the between-patient latent variable $\eta_{i(b)}$ for patient $i$ plus a within-patient vector of residual error terms $\boldsymbol{\epsilon}_{ei(w)}$, and a between-patient vector of residual error terms $\boldsymbol{\epsilon}_{i(b)}$. The residual error terms are multivariate normally distributed with means of zero and covariance matrices $\boldsymbol{\Theta}_w$ and $\boldsymbol{\Theta}_b$ for the within-patient residual terms and between-patient residual terms, respectively. The factor loading matrices reflects the pattern and magnitude of the relationship between the observed items and latent variables (factors) on both levels.

An important measurement assumption in ML-MIMIC models is the across-level invariance of factor loadings. Across-level invariance of factor loadings is a prerequisite to properly adopt the ML-MIMIC model described above, because it ensures that the same latent variable is measured at both levels. Mehta and Neale (2005) show that with equal factor loadings across levels, $\eta_{ei(w)} + \eta_{i(b)} = \eta_{ei}$, where $\eta_{ei}$ is a latent variable that is composed of within- and between-patient deviations. This means that the within-patient latent variable now has a random intercept at the between-patient level. Equal factor loadings across levels also indicates that latent factor variances are directly comparable, because invariant factor loadings equates the scale across levels of the common latent variable (Mehta & Neale, 2005). By comparing the latent factor variances, the proportion of variance located at the between-patient level can be calculated. This measure is also known as the intraclass correlation (ICC) of the factor, which is one of the conditions manipulated in this simulation study. Therefore, equal factor loadings across levels is also a prerequisite to manipulate the ICC of the common factor.

When there are covariates at the within-level and between-level, these covariates are included in the latent variable model part of the ML-MIMIC model with direct effects on the within-level and between-level latent variables. The regression coefficient of a within-level covariate on the within-level latent variable can be modeled as a random effect. This random effect indicates that the effect of that within-level covariate varies across second-level units at the between-level and this random effect can (partly) be explained by between-level covariates. If the effect of a between-level covariate on the random coefficient is significant, there exists a cross-level interaction effect, which means that the within-level covariate effect on the within-level latent variable depends on the value of the between-level covariate. An ML-MIMIC model with one latent variable at the within-level ($\eta_{ei(w)}$) and between-level ($\eta_{i(b)}$), one covariate at the within-level ($X_{ei(w)}$) and between-level ($X_{b(i)}$), a random slope, and a cross-level interaction effect is presented in Figure 1. The latent variable model part of this ML-MIMIC model can be written by the following set of equations:

$$\eta_{ei(w)} = \gamma_{i(w)} X_{ei(w)} + \zeta_{ei(w)}, \tag{2}$$

$$\eta_{i(b)} = \gamma_{(b)} X_{i(b)} + \zeta_{i(b)}, \tag{3}$$

$$\gamma_{i(w)} = \gamma_{10} + \gamma_{(c)} X_{i(b)} + \zeta_{\gamma_{i(w)}}, \tag{4}$$

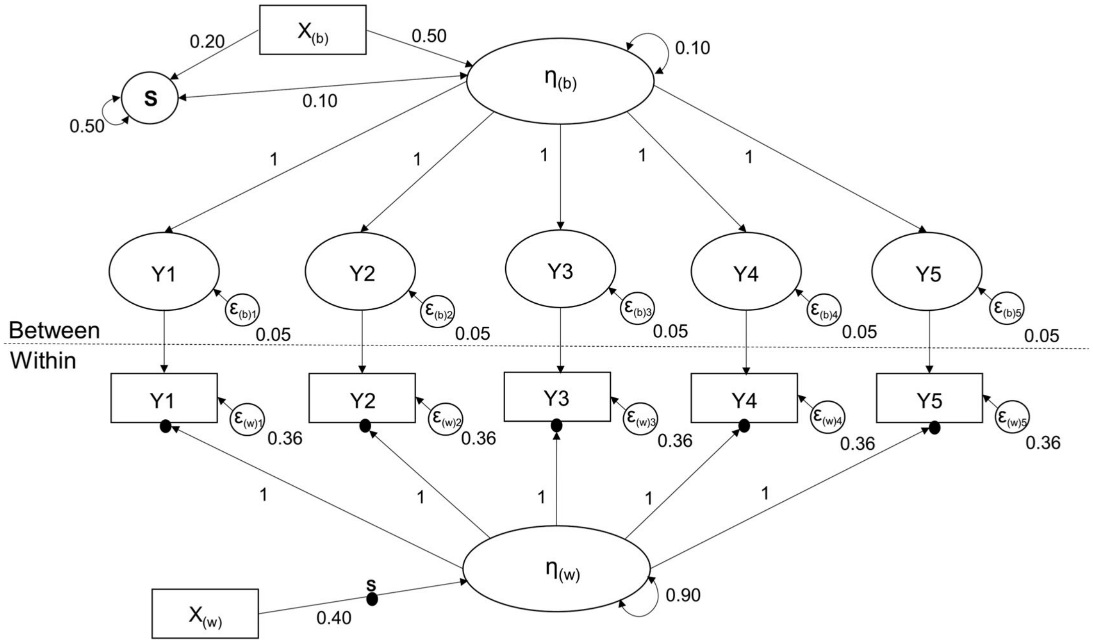

**FIGURE 1** ML-MIMIC population model from which the data were generated in one of the conditions.
Note: the bold S represents the random slope and other bold dots also indicate the parameter is random. The observed indicators are depicted as rectangles and latent variables are depicted as circles. The single arrows represent direct effects, whereas double arrows represent (co-)variances. The factor variances represent the variance under an unconditional model.

where $\gamma_{i(w)}$ and $\gamma_{(b)}$ represent the within-patient and between-patient effects of the covariates on within-level and between-level latent variable, respectively, and where $\gamma_{i(w)}$ represents the random slope of the within-level covariate $X_{ei(w)}$. This random slope is a function of the average slope $\gamma_{10}$, the effect of the between-level covariate $\gamma_{(c)}$, and the residual $\zeta_{\gamma_{i(w)}}$, implying that the variance of $\gamma_{i(w)}$ is not fully explained by the between-level covariate. The random slope residual ($\zeta_{\gamma_{i(w)}}$) and random intercept residual ($\zeta_{i(b)}$) are bivariate normally distributed with means of zero, and a 2 × 2 covariance matrix. The variance of the random slope reflects the degree to which the effect of study period varies across patients (after differences due to the between-level covariate between patients are considered). The residual term $\zeta_{ei(w)}$ is normally distributed with a mean of zero and within-level factor variance. The regression effect $\gamma_{(c)}$ represents the cross-level interaction effect between study period and treatment group. The combined model for the within-level latent factor, $\eta_{ei(w)}$, can be written as follows by substituting Equation (4) into Equation (2):

$$\eta_{ei(w)} = \gamma_{10}X_{ei(w)} + \gamma_{(c)}X_{ei(w)}X_{i(b)} + \zeta_{\gamma_{i(w)}}X_{ei(w)} + \zeta_{ei(w)}. \tag{5}$$

When comparing latent factor means across groups in SEMs, strong measurement invariance is a prerequisite to appropriately interpret latent mean differences (Meredith, 1993). This means that at least factor loadings and intercepts should be invariant across groups to ensure that any difference between groups is only attributable to the difference in latent factor means. A (ML-)MIMIC model estimates one model for the full combined sample of respondents (Muthén, 1989). This means that in a (ML-)MIMIC model, it is implicitly assumed that strong measurement invariance holds across groups located at the within-level and between-level. More specifically, factor loadings, intercepts, residual variances, and factor variances are implicitly assumed to be invariant across all groups when applying MIMIC models. In practice, this implicit measurement invariance assumption may be too strict and in a previous article (Kessels et al., 2021), we illustrated how to verify this assumption, inspired by the work of Kim et al. (2015). In the current simulation study, it was assumed that the measurement parameters were invariant across groups.

## 2.2 | Univariate multilevel model

In the univariate multilevel model, the multivariate response vectors $\boldsymbol{Y}_{ei}$ containing five observed scores on event $e$ for subject $i$ are combined into one univariate score by taking the sum of the five observed scores. This creates the observed univariate outcome $Y_{ei}$ measured at event $e$ for patient $i$. Considering the within-level covariate $X_{ei(w)}$ and the between-level covariate $X_{i(b)}$, the univariate multilevel regression model can be described as follows:

$$Y_{ei} = \gamma_{00} + \gamma_{10}X_{ei(w)} + \gamma_{(b)}X_{i(b)} + \gamma_{(c)}X_{ei(w)}X_{i(b)} + \zeta_{\gamma_{i(w)}}X_{ei(w)} + \zeta_{i(b)} + \zeta_{ei(w)}, \tag{6}$$

where $\gamma_{00}$ is the fixed intercept, $\gamma_{10}$ is the fixed main effect of the within-level covariate, $\gamma_{(b)}$ is the fixed main effect of the between-level covariate, and $\gamma_{(c)}$ is the fixed cross-level interaction effect. Furthermore, $\zeta_{i(b)}$ and $\zeta_{\gamma_{i(w)}}$ represent the residuals of the random intercept and random slope at the between-level, respectively, and $\zeta_{ei(w)}$ represents the within-level residual error term. Also, in the multilevel regression, the residuals of the random intercept and random slope are bivariate normally distributed with means of zero and a 2 × 2 covariance matrix. The within-level residual term is normally distributed with zero mean and variance $\sigma^2_{\zeta_{ei(w)}}$. The fixed regression coefficients $\gamma_{10}$ and $\gamma_{(c)}$ in Equation (6) correspond to $\gamma_{10}$ and $\gamma_{(c)}$ in Equation (5), respectively, whereas the fixed regression coefficient $\gamma_{(b)}$ in Equation (6) corresponds to $\gamma_{(b)}$ in Equation (3).

## 2.3 | Mixed ANOVA model

The mixed ANOVA model was applied to an aggregated data structure. Here, we follow the procedure similar to previous work on this topic (Kessels et al., 2019, 2021), which was motivated by the real-world clinical trial example outlined in Section 1. This means that the univariate outcome $Y_{ei}$ defined in the previous section, which denotes the sum of the observed scores at event $e$ for patient $i$, was averaged over all events belonging to the same level of the within-level grouping covariate within a patient. Let the levels of the within-level covariate be denoted by $t = 1, 2$ when the within-level covariate presents a dichotomous grouping covariate. Then, for each patient, $Y_{ti}$ was calculated as the average sum score over the events observed at $t = 1$ and $t = 2$, resulting in two aggregated means for each patient. Furthermore, let $X_{ti(w)}$ be the aggregated within-level covariate and let $X_{i(b)}$ be the between-level covariate. Subsequently, we can write the mixed ANOVA model as a random intercept-only multilevel equation (Hox et al., 2018):

$$Y_{ti} = \gamma_{00} + \gamma_{10}X_{ti(w)} + \gamma_{(b)}X_{i(b)} + \gamma_{(c)}X_{ti(w)}X_{i(b)} + \zeta_{i(b)} + \zeta_{ti(w)}, \tag{7}$$

where the fixed regression coefficients have the same interpretation as the regular multilevel model in Equation (6), as well as the within-level residual term and the residual of the random intercept. The mixed ANOVA model does not have a random slope, indicating that the within-level covariate effect does not vary across second-level units. Model 7 is analogous to a mixed between-within-subject ANOVA, as a multilevel model with only a random intercept assumes compound symmetry (Hox et al., 2018). Compound symmetry requires that all population variances of the repeated measures are equal and that all population covariances of the repeated measures are equal, which is the same restriction present in repeated measures ANOVA (Maxwell & Delaney, 2004).

## 2.4 | Design of the simulation study

The purpose of this simulation study was to compare the ML-MIMIC model (Equations (1), (3), and (5)), the multilevel model using the sum scores per measurement as univariate outcome (Equation (6)) and the mixed ANOVA (Equation (7)) applied to the aggregated data on the performance and accuracy in estimating and testing the cross-level interaction effect. The three different models were compared on type I error rate, statistical power, standard error (SE) of the cross-level interaction effect, relative bias of the estimated interaction effect, relative SE bias of the estimated SE of the interaction effect, and mean square error (MSE) of the interaction effect.

All data were generated under a two-level ML-MIMIC model with one dichotomous covariate on the between-level, $X_{(b)}$, one dichotomous covariate on the within-level, $X_{(w)}$, their cross-level interaction effect, and a one-factor solution with five observed indicators specified at both levels as depicted in Figure 1. Furthermore, the population model that was used to generate the data included a random intercept, a random slope for the within-level main effect, and a covariance between the random intercept and random slope. Data were generated in two steps. First, two dichotomous covariates and their interaction were created, which were used along with the fixed main effects and fixed cross-level interaction

**TABLE 1** Overview of conditions varied in the simulation study and their values.

| Condition | Levels |
|---|---|
| Factor loadings ($\Lambda$[a]) | (1, 1, 1, 1, 1), (1, 0.95, 0.80, 0.90, 0.85) |
| | (1, 0.90, 0.60, 0.80, 0.70) |
| Sample size between-level ($n_2$) | 50, 100, 200 |
| Sample size within-level ($n_1$) | 10, 20 |
| Factor variances ($\tau^2, \pi^2$) | (0.10, 0.90), (0.25, 0.75) |
| Residual variances within-level ($\Theta_w$) | (0.36, 0.36, 0.36, 0.36, 0.36), (0.60, 0.19, 0.36, 0.42, 0.51) |
| Magnitude interaction effect ($\gamma_{(c)}$) | 0, 0.20, 0.40 |

*Note*: Item-level ICCs can be derived using factor variances and residual variances and ranged between 0.09 and 0.24.
aFactor loadings were equal across the within-level and the between-level.

effect to generate factor scores at both levels using Equations (5) and (3) for the within-level and between-level factor scores, respectively. At the between-level, two residual scores were generated that represented the residual factor scores ($\zeta_{i(b)}$) and residual slopes ($\zeta_{\gamma_{i(w)}}$). These were drawn from a bivariate normal distribution with zero means and covariance matrix with the between-level factor variance, hereafter denoted as $\tau^2$, and random slope variance on the main diagonal and the covariance between the between-level factor and random slope on the off-diagonal. Residual factor scores at the within-level, $\zeta_{ei(w)}$ in Equation (5), were drawn from a normal distribution with mean zero and standard deviation equal to the square root of the within-level factor variance, hereafter denoted as $\pi^2$. In the next step, the factor scores were used to generate the five observed indicator scores for each within-level unit using Equation (1). Residual scores for the observed indicators at the within-level ($\epsilon_{ei(w)}$ in Equation (1)) and indicator means at the between-level ($\epsilon_{i(b)}$ in Equation (1)) were drawn from a multivariate normal distribution with zero means and diagonal covariance matrix with the within-level or between-level residual variances at the main diagonal. The intercepts ($\mu$) at the between-level were set to zero.

Five conditions were varied in this simulation study: factor loadings, sample size at the within and between-level, ICC value of the factor by varying the factor variances, residual variance of the indicators at the within-level, and the magnitude of the cross-level interaction effect. These conditions and their levels are listed in Table 1.

Sum scores approximate factor scores when the unstandardized factor loadings are nearly identical across the indicators and in such a circumstance, sum scores may provide a valid alternative to factor scores (McNeish & Wolf, 2020). On the other hand, the more the factor loadings differ across the indicators, the more discrepancy is expected between an analysis on sum scores and an analysis on factor scores with factor scores expected to have less bias in such situations as compared to sum scores. To study the influence of this trade-off, three different conditions for the factor loadings were used. In the first condition, all loadings were set equal to 1. In the second condition, factor loadings were set to 1, 0.95, 0.80, 0.90, and 0.85 to create a condition with small differences in loadings between the indicators. In the third condition, larger differences in factor loadings were specified by setting the factor loadings to 1, 0.90, 0.60, 0.80, and 0.70. In all three conditions, across-level measurement invariance was assumed, that is, equality of within-level and between-level factor loadings.

Three conditions were specified for the sample size ($n_2$) at the between-level: 50, 100, 200. Similar values were used in previous simulation studies with multilevel (factor) models (Cao et al., 2019; Finch & French, 2011; Hox & Maas, 2001; Jak, 2019; Maas & Hox, 2005). At the within-level, two conditions for the sample size ($n_1$) were specified: 10, 20. These values represent the overall number of observations per second-level unit. The within-level sample sizes also match those used in previous research (Cao et al., 2019; Finch & French, 2011; Hox & Maas, 2001; Maas & Hox, 2005). The total sample size ($N$) is equal to $n_2 \times n_1$ creating a minimum total sample size of $N = 500$ and a maximum total sample size of $N = 4000$. Considering the motivating example, the sample size can be regarded as the number of patients in a clinical trial ($n_2$) with the number of events nested within each patient at the within-level sample size ($n_1$). Note that the sample size can just as well be interpreted as the number of clusters in a cluster randomized trial ($n_2$) with the number of patients nested within each cluster as the within-level sample size ($n_1$). Therefore, this simulation study offers insights for other applications in medical research as well.

The unconditional ICC of the common factor was also manipulated in this simulation study as previous studies showed that the ICC has an impact on statistical power and parameter estimates in ML-MIMIC models (Cao et al., 2019; Finch & French, 2011). The unconditional ICC is defined as: $\tau^2/(\tau^2 + \pi^2)$ where $\tau^2$ and $\pi^2$ are the factor variances at the between-level and within-level, respectively, under a model without any covariates. Consequently, the ICC can be manipulated by varying the factor variances. In one condition, the between-level factor variance was set to 0.10 and the within-level factor variance to 0.90, creating an ICC of 0.10 (small ICC). In a second condition, the variances were set to 0.25 and 0.75 for the between-level factor and within-level factor, respectively, creating an ICC of 0.25 (medium ICC). These ICC values represent the factor ICC unconditional on the covariate effects, which is analogous to the ICC value under an intercept-only multilevel regression model. In addition to the proportion of variance located at the between-level, the ICC is also interpreted as the expected correlation between two randomly drawn events from the same patient. The ICC values used in this simulation study are typical of psychological and educational data and correspond to ICC values used in previous simulation studies with a multilevel CFA model (Cao et al., 2019; Finch & French, 2011; Hox & Maas, 2001; Jak, 2019; Maas & Hox, 2005). Furthermore, ICC values ranging between 0.01 and 0.30 have been used in simulation studies for calculating optimal sample sizes in cluster randomized trials (Candel & Van Breukelen, 2015; Moerbeek, 2012).

Different ICC values were also used to investigate if the ICC had an effect on the power when comparing the univariate multilevel model with the mixed ANOVA model. In previous work, where we used empirical data to compare a univariate multilevel model to a mixed ANOVA model (Kessels et al., 2019), it was found that the *p*-value of the cross-level interaction effect of the multilevel model did not differ much from the *p*-value of the mixed ANOVA model. It was expected that the mixed ANOVA model had less power, because the mixed ANOVA was applied to an aggregated data set and aggregating data leads to a loss of information and therefore to a loss of power. However, the sum score of each event is likely measured with error and multiple sum scores within a patient are correlated. Then, by aggregating the sum scores into means, the resulting aggregated outcome will have less error, which could, in turn, lead to more power when applying the mixed ANOVA using the aggregated outcome. These two tendencies could work in opposite directions concerning the power. Since the ICC is also defined as the expected correlation between two randomly drawn scores of the same patient, this tendency was investigated by varying the ICC value of the univariate outcome (latent variable and sum score).

The within-level residual variances of the indicators were also varied in this simulation study. The within-level residual variances were all set to 0.36 in one condition, whereas in a second condition, these values were varied across indicators and set to 0.60, 0.19, 0.36, 0.42, and 0.51. The rationale behind these selected values is based on the following consideration. Sum scores assume that each indicator contributes an equal amount of information to the construct being measured, which is referred to as unit-weighting (McNeish & Wolf, 2020). McNeish and Wolf (2020) pointed out that unit-weighting can be specified in a factor model by constraining all standardized loadings to the same value. This can be obtained by setting all unstandardized factor loadings and residual variances to be equal across indicators. Consequently, a factor model with equal unstandardized factor loadings and residual variances creates factor scores that are perfectly correlated with sum scores. In the current simulation study, this perfect correlation was obtained in the conditions with equal factor loadings and equal residual variances. Conditions where the residual variances differed across indicators were included in addition to conditions with unequal factor loadings to explore the effect of both conditions on using the ML-MIMIC model or sum scores. The residual variances at the between-level were set to 0.05 for all conditions, because between-level residual variances need preferably to be close to zero in order to satisfy invariance of intercepts across between-level units (Jak et al., 2013).

Finally, the magnitude of the cross-level interaction was manipulated as it has been shown that the power of a multilevel regression model to detect a cross-level interaction effect mainly depends on the effect size of the interaction (Mathieu et al., 2012). The magnitude of the cross-level interaction was set to 0, 0.2, and 0.4, which is consistent with cross-level interaction values in previous simulation studies that studied cross-level interaction effects in multilevel data (Cao et al., 2019; Mathieu et al., 2012). Conditions with an interaction effect of 0 were only used to assess type I error rates when there is no interaction present in the population model.

Figure 1 shows the population values for all parameters in the condition with all factor loadings equal to 1, factor variances of 0.10 and 0.90 at the between-level and within-level, respectively, equal within-level residual variances of the indicators, and an interaction effect of 0.20. The main effects of the covariates (0.40 and 0.50), the random slope variance (0.50), the covariance between the random slope and the random intercept (0.10), and the residual variances of the indicators at the between-level (0.05) were fixed across all simulation conditions, as can be seen in Figure 1. In sum, there were a total of $3 \times 3 \times 2 \times 2 \times 2 \times 3 = 216$ different conditions. For each condition, 1000 data sets were generated. All data were generated in R, version 4.0.4 (R Core Team, 2021).

Each generated data set was analyzed using the ML-MIMIC model, the multilevel model using the sum scores per measurement as univariate outcome, and the mixed ANOVA model using the patient-level means of the sum scores. The parameters of the ML-MIMIC model were estimated in `Mplus`, version 7.3 (Muthén & Muthén, 1998-2017) using robust maximum likelihood estimation. The R-package `MplusAutomation`, version 0.8 (Hallquist & Wiley, 2018) was used to load the parameter estimates directly into R. The latent factor scales of the ML-MIMIC model were identified by fixing the factor loading of the first indicator to 1.00. The remaining factor loadings, latent factor variances, random slope variance, residual variances, and the covariance between the between-level factor and random slope were estimated without any constraints. The difference in means between the groups was estimated by the regression coefficients.

For the univariate multilevel analysis, the sum score of the five generated indicators was derived for each measurement separately that served as the outcome variable in this analysis. For estimating the parameters of this multilevel model, full information maximum likelihood was employed. The multilevel model was fitted including a random slope for the within-level covariate and a covariance between the random slope and random intercept.

For the mixed ANOVA model, two average scores for each patient were derived: the average of the sum scores where $X_w = 0$ and the average of the sum scores where $X_w = 1$. This reduced the total sample size such that there were $n_2 \times 2$ observations when applying this model. The parameters of the mixed ANOVA model were estimated using a multilevel model with a random intercept only. To ensure that these parameter estimates are equivalent to ANOVA estimates, we employed restricted maximum likelihood estimation (as opposed to full information maximum likelihood estimation). With balanced groups (which always applied in this simulation study), restricted maximum likelihood estimates of a multilevel model with a random intercept only are equivalent to mixed ANOVA estimates (Hox et al., 2018). The multilevel model and the mixed ANOVA model were analyzed in `R`, version 4.0.4 (R Core Team, 2021), using the `lme4` package, version 1.1-26 (Bates et al., 2015) together with the `lmerTest` package, version 3.1-3 (Kuznetsova et al., 2017). By activating the `lmerTest` package before performing analyses using the `lme4` package, the output of `lme4` functions provides $p$-values for the $t$-tests of the fixed regression parameters based on the Satterthwaite (Satterthwaite, 1946) approximation for the calculation of the degrees-of-freedom.

## 2.5 | Criteria for evaluation

Criteria for evaluation included admissible solution rates (ASR), type I error rate, statistical power, SE, relative parameter bias, relative SE bias, and MSE. These criteria were derived for each model separately for each of the 216 conditions. For the ML-MIMIC model, a replication was classified inadmissible if the model estimation did not terminate normally, if there were any other error or warning messages, if no parameter estimates or SEs were produced or if the estimated parameters were not possible, such as negative SEs or negative variances. Furthermore, if the correlation between the random intercept or slope was larger than +1 or smaller than −1, it was also counted as an inadmissible solution. For the multilevel model and the mixed ANOVA model, a replication was classified inadmissible if the model estimation failed to converge, if a singular model was obtained, or if the estimated parameters were not possible, such as negative SEs, negative variances or a correlation coefficient between the random intercept and random slope larger than +1 or smaller than −1 (for the multilevel model only). The ASR was defined as the proportion of replications that produced admissible solutions during the first 1000 replications. A replication where for at least one of the models an inadmissible solution was obtained was discarded and not used in further analyses. The data generation process continued until 1000 replications were obtained for which all three models produced no inadmissible solutions. This ensured that all other evaluation criteria were evaluated considering 1000 admissible solutions.

Type I error rate was defined as the proportion of replications where the models falsely rejected the null hypothesis of no cross-level interaction effect at a two-sided test with $\alpha = 0.05$ under conditions where there was no cross-level interaction effect in the population. Statistical power was defined as the proportion of replications where the models correctly rejected the null hypothesis of no cross-level interaction effect at a two-sided test with $\alpha = 0.05$ under conditions where the population cross-level interaction effect was 0.2 or 0.4.

The relative bias, relative SE bias, and MSE were used to evaluate the accuracy of the interested cross-level interaction effect $\gamma_{(c)}$ in Equations (5)–(7). To be able to compare these criteria across the three models, the cross-level interaction effect estimates and their SEs of the univariate multilevel model and the mixed ANOVA model were rescaled to the same scale as the factor, by dividing them by five (number of indicators).

The relative bias of parameter $\gamma_{(c)}$ is defined as:

$$\text{Relative parameter bias} = \frac{\overline{\hat{\gamma}_{(c)}} - \gamma_{(c)}}{\gamma_{(c)}}, \tag{8}$$

where $\gamma_{(c)}$ is the population parameter value that is used to generate the data sets and $\overline{\hat{\gamma}_{(c)}}$ is its estimated average over the 1000 replicated data sets. The relative bias indicates how much the population parameter and the estimated parameter differ.

**TABLE 2** Type I error rates under equal within-level residual variances.

| | | $\Lambda = (1, 1, 1, 1, 1)$ | | | $\Lambda = (1, 0.95, 0.80, 0.90, 0.85)$ | | | $\Lambda = (1, 0.90, 0.60, 0.80, 0.70)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **ICC** | ***n2/n1*** | **MIMIC** | **MLM** | **MAOV** | **MIMIC** | **MLM** | **MAOV** | **MIMIC** | **MLM** | **MAOV** |
| 0.10 | 50/10 | 0.048 | 0.045 | 0.042 | 0.048 | 0.041 | 0.037 | 0.062 | 0.058 | 0.056 |
| | 50/20 | 0.066 | 0.061 | 0.059 | 0.042 | 0.039 | 0.033 | 0.057 | 0.052 | 0.050 |
| | 100/10 | 0.058 | 0.052 | 0.047 | 0.064 | 0.061 | 0.058 | 0.050 | 0.050 | 0.045 |
| | 100/20 | 0.047 | 0.045 | 0.043 | 0.047 | 0.046 | 0.045 | 0.044 | 0.046 | 0.043 |
| | 200/10 | 0.055 | 0.055 | 0.054 | 0.053 | 0.050 | 0.050 | 0.045 | 0.047 | 0.044 |
| | 200/20 | 0.051 | 0.049 | 0.048 | 0.058 | 0.058 | 0.058 | 0.041 | 0.041 | 0.039 |
| 0.25 | 50/10 | 0.056 | 0.049 | 0.047 | 0.075 | 0.068 | 0.063 | 0.065 | 0.058 | 0.053 |
| | 50/20 | 0.065 | 0.059 | 0.057 | 0.064 | 0.054 | 0.046 | 0.069 | 0.065 | 0.060 |
| | 100/10 | 0.048 | 0.045 | 0.044 | 0.054 | 0.053 | 0.052 | 0.047 | 0.045 | 0.044 |
| | 100/20 | 0.052 | 0.051 | 0.051 | 0.044 | 0.043 | 0.040 | 0.045 | 0.042 | 0.039 |
| | 200/10 | 0.051 | 0.051 | 0.050 | 0.052 | 0.052 | 0.051 | 0.050 | 0.049 | 0.048 |
| | 200/20 | 0.057 | 0.056 | 0.056 | 0.036 | 0.035 | 0.035 | 0.058 | 0.057 | 0.055 |

*Note*: $\gamma$ = interaction effect; ICC = intraclass correlation; $n2/n1$ = sample size; MIMIC = Multilevel MIMIC model; MLM = Multilevel model; MAOV = Mixed ANOVA model.

Shaded cells show type I error rates that differ significantly from the desired nominal alpha level of 0.05.

The relative SE bias of $\gamma_{(c)}$ indicates how much the average estimated SE $\overline{\widehat{se}(\widehat{\gamma}_{(c)})}$ deviates from the standard deviation $sd(\widehat{\gamma}_{(c)})$ over the 1000 estimates of $\widehat{\gamma}_{(c)}$:

$$\text{Relative SE bias} = \frac{\overline{\widehat{se}(\widehat{\gamma}_{(c)})} - sd(\widehat{\gamma}_{(c)})}{sd(\widehat{\gamma}_{(c)})}. \tag{9}$$

A positive value implies that on average the estimated SEs are too large; whereas a negative value is related to an underestimation of the SEs. Muthén and Muthén (2002) suggest that parameter bias should not exceed 10% and SE bias should preferably not exceed 5% for a parameter that is used to assess power.

Finally, the MSE was calculated and used as criterium for evaluation. The MSE is an overall measure of accuracy because it considers bias and the standard deviation over the 1000 replications:

$$MSE = (\overline{\widehat{\gamma}_{(c)}} - \gamma_{(c)})^2 + sd(\widehat{\gamma}_{(c)})^2. \tag{10}$$

## 3 | RESULTS

The main results with respect to the type I error rate, power, SE, and relative bias are presented in Tables 2–5. The ASR, relative SE bias, and MSE results are presented in the Supporting Information. Since the results under the equal within-level residual variance condition were similar to the results under the unequal within-level residual variance condition, the tables presented here only contain the results under the equal residual within-level variance condition. For a complete overview of all results, we refer to the Supporting Information. The results are discussed below per evaluation criterion.

### 3.1 | ASR

The ASR was related to the ICC and sample size for the ML-MIMIC model and the univariate multilevel model. For these models, conditions with small sample sizes at the within-level and low ICC values produced relatively lower ASRs, as presented in Table S1. Especially for the multilevel model, the ASR was below 80% in the smallest sample size condition with a small ICC, as highlighted by the gray shaded cells in Table S1. This is not surprising as a relatively low ICC indicates that there is little variation at the between-level. Consequently, together with a small sample size, a multilevel model will then have difficulties estimating random slope variance increasing the likelihood that the model estimation will not

TABLE 3 Power estimates under equal within-level residual variances.

| | | | $\Lambda = (1, 1, 1, 1, 1)$ | | | $\Lambda = (1, 0.95, 0.80, 0.90, 0.85)$ | | | $\Lambda = (1, 0.90, 0.60, 0.80, 0.70)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\gamma$ | ICC | $n2/n1$ | MIMIC | MLM | MAOV | MIMIC | MLM | MAOV | MIMIC | MLM | MAOV |
| 0.20 | 0.10 | 50/10 | 0.121 | 0.115 | 0.109 | 0.125 | 0.120 | 0.111 | 0.111 | 0.103 | 0.095 |
| | | 50/20 | 0.146 | 0.139 | 0.134 | 0.146 | 0.138 | 0.127 | 0.142 | 0.133 | 0.127 |
| | | 100/10 | 0.185 | 0.181 | 0.169 | 0.187 | 0.186 | 0.181 | 0.199 | 0.199 | 0.193 |
| | | 100/20 | 0.233 | 0.230 | 0.222 | 0.223 | 0.215 | 0.205 | 0.207 | 0.198 | 0.195 |
| | | 200/10 | 0.324 | 0.319 | 0.316 | 0.354 | 0.354 | 0.349 | 0.294 | 0.293 | 0.288 |
| | | 200/20 | 0.392 | 0.391 | 0.385 | 0.390 | 0.389 | 0.385 | 0.382 | 0.380 | 0.377 |
| | 0.25 | 50/10 | 0.139 | 0.130 | 0.116 | 0.124 | 0.117 | 0.102 | 0.133 | 0.129 | 0.121 |
| | | 50/20 | 0.172 | 0.164 | 0.153 | 0.163 | 0.149 | 0.133 | 0.145 | 0.133 | 0.126 |
| | | 100/10 | 0.198 | 0.189 | 0.177 | 0.211 | 0.205 | 0.197 | 0.176 | 0.172 | 0.168 |
| | | 100/20 | 0.251 | 0.246 | 0.238 | 0.221 | 0.216 | 0.210 | 0.246 | 0.235 | 0.232 |
| | | 200/10 | 0.376 | 0.370 | 0.366 | 0.366 | 0.365 | 0.361 | 0.337 | 0.336 | 0.335 |
| | | 200/20 | 0.426 | 0.422 | 0.421 | 0.418 | 0.410 | 0.404 | 0.410 | 0.405 | 0.402 |
| 0.40 | 0.10 | 50/10 | 0.346 | 0.329 | 0.306 | 0.331 | 0.322 | 0.312 | 0.315 | 0.297 | 0.282 |
| | | 50/20 | 0.404 | 0.388 | 0.376 | 0.400 | 0.378 | 0.368 | 0.395 | 0.377 | 0.352 |
| | | 100/10 | 0.562 | 0.557 | 0.547 | 0.567 | 0.568 | 0.553 | 0.568 | 0.559 | 0.554 |
| | | 100/20 | 0.693 | 0.691 | 0.683 | 0.688 | 0.678 | 0.671 | 0.659 | 0.649 | 0.639 |
| | | 200/10 | 0.843 | 0.841 | 0.841 | 0.862 | 0.860 | 0.855 | 0.846 | 0.845 | 0.844 |
| | | 200/20 | 0.930 | 0.928 | 0.927 | 0.932 | 0.931 | 0.929 | 0.919 | 0.913 | 0.912 |
| | 0.25 | 50/10 | 0.360 | 0.345 | 0.337 | 0.360 | 0.344 | 0.329 | 0.355 | 0.344 | 0.334 |
| | | 50/20 | 0.445 | 0.427 | 0.414 | 0.419 | 0.401 | 0.388 | 0.428 | 0.411 | 0.398 |
| | | 100/10 | 0.605 | 0.602 | 0.596 | 0.610 | 0.606 | 0.595 | 0.606 | 0.595 | 0.589 |
| | | 100/20 | 0.705 | 0.696 | 0.687 | 0.715 | 0.709 | 0.703 | 0.688 | 0.678 | 0.673 |
| | | 200/10 | 0.890 | 0.886 | 0.884 | 0.871 | 0.870 | 0.868 | 0.863 | 0.857 | 0.855 |
| | | 200/20 | 0.943 | 0.942 | 0.939 | 0.945 | 0.943 | 0.941 | 0.944 | 0.944 | 0.942 |

*Note*: $\gamma$ = interaction effect; ICC = intraclass correlation; $n2/n1$ = sample size; MIMIC = Multilevel MIMIC model; MLM = Multilevel model; MAOV = Mixed ANOVA model.
Shaded cells show power estimates that exceed the desired 80%.

converge to a solution. However, as soon as the sample size at the within-level was 20, the ASR exceeded 90%. Overall, the ASR was often well above 90% and often 100% in the larger sample size conditions for the ML-MIMIC model and multilevel model. For the mixed ANOVA model, ASRs were all 100% across all simulation conditions, except for some conditions with smaller sample sizes. The same pattern was observed between the different factor loading conditions and between the within-level residual variance conditions.

## 3.2 | Type I error

The type I error rates are presented in Table 2 and Table S2. For the ML-MIMIC model, type I error rates were around 0.06 and ranging from 0.04 to 0.08. This is consistent with the results of Cao et al. (2019), where the same range in type I error rates was observed when testing cross-level interaction effects in ML-MIMIC models. When using the one-sample proportion test, it is possible to see whether the type I error rates differ significantly from the preferred nominal-level alpha level of 0.05. In the case of 1000 replications and a 5% significant level, the lower limit value of alpha is 0.03649 and the upper limit is 0.0635. The significant type I error values are highlighted by the gray shaded cells in Table 2. When looking at these results, it can be seen that the type I error rate under the ML-MIMIC model was more often significantly different from the nominal 5% alpha level compared to the univariate multilevel model and mixed ANOVA model. Type I error rates that are significantly larger than the nominal alpha level of 0.05 are undesirable, especially when a model is used in a clinical trial. For the ML-MIMIC model, 19% of the type I error rates were significantly larger than 0.05, usually with second-level sample sizes of 50. The remaining type I error rates of the ML-MIMIC model were reasonably well controlled.

**TABLE 4** Standard error estimates under equal within-level residual variances.

| | | | $\Lambda = (1, 1, 1, 1, 1)$ | | | $\Lambda = (1, 0.95, 0.80, 0.90, 0.85)$ | | | $\Lambda = (1, 0.90, 0.60, 0.80, 0.70)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\gamma$ | ICC | $n2/n1$ | MIMIC | MLM | MAOV | MIMIC | MLM | MAOV | MIMIC | MLM | MAOV |
| 0.20 | 0.10 | 50/10 | 0.266 | 0.266 | 0.271 | 0.267 | 0.240 | 0.244 | 0.268 | 0.214 | 0.219 |
| | | 50/20 | 0.230 | 0.230 | 0.235 | 0.232 | 0.209 | 0.213 | 0.233 | 0.186 | 0.190 |
| | | 100/10 | 0.188 | 0.188 | 0.190 | 0.188 | 0.169 | 0.171 | 0.190 | 0.152 | 0.153 |
| | | 100/20 | 0.165 | 0.165 | 0.166 | 0.165 | 0.148 | 0.150 | 0.165 | 0.133 | 0.134 |
| | | 200/10 | 0.132 | 0.132 | 0.133 | 0.133 | 0.120 | 0.121 | 0.134 | 0.107 | 0.108 |
| | | 200/20 | 0.117 | 0.117 | 0.118 | 0.118 | 0.106 | 0.106 | 0.117 | 0.094 | 0.094 |
| | 0.25 | 50/10 | 0.252 | 0.252 | 0.257 | 0.254 | 0.228 | 0.233 | 0.254 | 0.203 | 0.207 |
| | | 50/20 | 0.225 | 0.225 | 0.230 | 0.224 | 0.202 | 0.206 | 0.227 | 0.182 | 0.186 |
| | | 100/10 | 0.180 | 0.180 | 0.182 | 0.180 | 0.162 | 0.163 | 0.181 | 0.145 | 0.147 |
| | | 100/20 | 0.161 | 0.161 | 0.162 | 0.162 | 0.145 | 0.147 | 0.162 | 0.129 | 0.131 |
| | | 200/10 | 0.128 | 0.128 | 0.129 | 0.128 | 0.116 | 0.116 | 0.129 | 0.103 | 0.104 |
| | | 200/20 | 0.115 | 0.115 | 0.115 | 0.115 | 0.103 | 0.104 | 0.116 | 0.092 | 0.093 |
| 0.40 | 0.10 | 50/10 | 0.266 | 0.266 | 0.271 | 0.266 | 0.239 | 0.244 | 0.268 | 0.214 | 0.219 |
| | | 50/20 | 0.232 | 0.232 | 0.236 | 0.231 | 0.208 | 0.212 | 0.232 | 0.186 | 0.190 |
| | | 100/10 | 0.188 | 0.187 | 0.189 | 0.188 | 0.169 | 0.171 | 0.190 | 0.151 | 0.153 |
| | | 100/20 | 0.164 | 0.164 | 0.166 | 0.165 | 0.149 | 0.150 | 0.165 | 0.132 | 0.134 |
| | | 200/10 | 0.133 | 0.133 | 0.134 | 0.133 | 0.119 | 0.120 | 0.134 | 0.107 | 0.108 |
| | | 200/20 | 0.117 | 0.117 | 0.118 | 0.118 | 0.106 | 0.106 | 0.118 | 0.094 | 0.095 |
| | 0.25 | 50/10 | 0.253 | 0.252 | 0.257 | 0.254 | 0.228 | 0.233 | 0.255 | 0.204 | 0.208 |
| | | 50/20 | 0.225 | 0.225 | 0.230 | 0.227 | 0.204 | 0.208 | 0.226 | 0.181 | 0.184 |
| | | 100/10 | 0.180 | 0.180 | 0.182 | 0.181 | 0.162 | 0.164 | 0.181 | 0.145 | 0.146 |
| | | 100/20 | 0.162 | 0.162 | 0.163 | 0.161 | 0.145 | 0.146 | 0.162 | 0.129 | 0.130 |
| | | 200/10 | 0.128 | 0.128 | 0.129 | 0.129 | 0.116 | 0.116 | 0.129 | 0.103 | 0.104 |
| | | 200/20 | 0.114 | 0.114 | 0.115 | 0.115 | 0.103 | 0.104 | 0.115 | 0.092 | 0.092 |

*Note*: $\gamma$ = interaction effect; ICC = intraclass correlation; $n2/n1$ = sample size; MIMIC = Multilevel MIMIC model; MLM = Multilevel model; MAOV = Mixed ANOVA model.

For the univariate multilevel model, the type I error rates were slightly lower compared to the ML-MIMIC model, especially under small sample size conditions. As shown in Table 2, type I error rates for the multilevel models never exceeded 0.07. Also for the mixed ANOVA model, type I error rates were slightly lower than for the ML-MIMIC model and even slightly lower than for the multilevel model. These results reveal that the mixed ANOVA model has the best type I error control among the three models, but the differences between the models were minimal.

## 3.3 | Power

In almost all simulation conditions, the power of the ML-MIMIC model was larger than the power of the multilevel model and the mixed ANOVA model, as presented in Table 3 and Table S3. When comparing the multilevel model with the mixed ANOVA model only, it shows that the power of the multilevel model was almost always larger than the power of the mixed ANOVA model with a few exceptions where the power of both models was the same. The differences in power between the models became smaller with increasing sample sizes. However, overall, the differences between the power estimates of the three models were small. The small power advantage of the ML-MIMIC model over the other two methods and the power advantage of the multilevel model over the mixed ANOVA model could be explained by base differences in type I error rates, where we found the same pattern across the models. However, whether the power difference is purely due to differences in type I error rates, is hard to evaluate.

For all three models, the power reached 80% under the simulation conditions with a cross-level interaction effect of 0.40, a between-level sample size of 200 and a within-level sample size of 10, regardless of the factor loading values, ICC, and within-level residual variances, as highlighted by the gray shaded cells in Table 3 and Table S3.

**TABLE 5** Relative Bias under equal within-level residual variances.

| | | | $\Lambda = (1, 1, 1, 1, 1)$ | | | $\Lambda = (1, 0.95, 0.80, 0.90, 0.85)$ | | | $\Lambda = (1, 0.90, 0.60, 0.80, 0.70)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\gamma$ | ICC | $n2/n1$ | MIMIC | MLM | MAOV | MIMIC | MLM | MAOV | MIMIC | MLM | MAOV |
| 0.20 | 0.10 | 50/10 | 0.028 | 0.029 | 0.029 | 0.011 | −0.089 | −0.089 | −0.000 | −0.200 | −0.200 |
| | | 50/20 | 0.028 | 0.027 | 0.027 | 0.022 | −0.079 | −0.079 | −0.019 | −0.215 | −0.215 |
| | | 100/10 | −0.070 | −0.071 | −0.071 | 0.023 | −0.080 | −0.080 | 0.032 | −0.174 | −0.174 |
| | | 100/20 | 0.007 | 0.008 | 0.008 | −0.020 | −0.119 | −0.119 | −0.030 | −0.223 | −0.223 |
| | | 200/10 | 0.016 | 0.017 | 0.017 | 0.031 | −0.073 | −0.073 | −0.008 | −0.207 | −0.207 |
| | | 200/20 | −0.028 | −0.028 | −0.028 | 0.023 | −0.079 | −0.079 | −0.011 | −0.208 | −0.208 |
| | 0.25 | 50/10 | 0.045 | 0.046 | 0.046 | −0.037 | −0.134 | −0.134 | −0.030 | −0.223 | −0.223 |
| | | 50/20 | 0.023 | 0.023 | 0.023 | −0.014 | −0.112 | −0.112 | −0.027 | −0.220 | −0.220 |
| | | 100/10 | −0.028 | −0.029 | −0.029 | −0.010 | −0.108 | −0.108 | −0.042 | −0.233 | −0.233 |
| | | 100/20 | 0.013 | 0.014 | 0.014 | −0.015 | −0.114 | −0.114 | 0.027 | −0.178 | −0.178 |
| | | 200/10 | 0.045 | 0.046 | 0.046 | 0.014 | −0.087 | −0.087 | 0.008 | −0.193 | −0.193 |
| | | 200/20 | 0.005 | 0.005 | 0.005 | −0.005 | −0.105 | −0.105 | −0.019 | −0.215 | −0.215 |
| 0.40 | 0.10 | 50/10 | 0.026 | 0.027 | 0.027 | 0.035 | −0.070 | −0.070 | −0.001 | −0.200 | −0.200 |
| | | 50/20 | 0.009 | 0.009 | 0.009 | −0.012 | −0.110 | −0.110 | −0.019 | −0.214 | −0.214 |
| | | 100/10 | 0.000 | 0.000 | 0.000 | 0.018 | −0.084 | −0.084 | 0.005 | −0.196 | −0.196 |
| | | 100/20 | 0.017 | 0.017 | 0.017 | −0.002 | −0.102 | −0.102 | −0.025 | −0.220 | −0.220 |
| | | 200/10 | 0.004 | 0.004 | 0.004 | 0.000 | −0.100 | −0.100 | 0.005 | −0.196 | −0.196 |
| | | 200/20 | 0.004 | 0.004 | 0.004 | 0.014 | −0.087 | −0.087 | −0.008 | −0.206 | −0.206 |
| | 0.25 | 50/10 | 0.004 | 0.004 | 0.004 | 0.028 | −0.075 | −0.075 | 0.012 | −0.190 | −0.190 |
| | | 50/20 | 0.013 | 0.013 | 0.013 | −0.006 | −0.105 | −0.105 | −0.014 | −0.211 | −0.211 |
| | | 100/10 | 0.017 | 0.017 | 0.017 | 0.011 | −0.091 | −0.091 | 0.009 | −0.193 | −0.193 |
| | | 100/20 | −0.002 | −0.001 | −0.001 | 0.016 | −0.086 | −0.086 | 0.004 | −0.197 | −0.197 |
| | | 200/10 | 0.012 | 0.012 | 0.012 | 0.007 | −0.093 | −0.093 | −0.004 | −0.203 | −0.203 |
| | | 200/20 | −0.006 | −0.006 | −0.006 | −0.000 | −0.100 | −0.100 | 0.019 | −0.185 | −0.185 |

*Note*: $\gamma$ = interaction effect; ICC = intraclass correlation; $n2/n1$ = sample size; MIMIC = Multilevel MIMIC model; MLM = Multilevel model; MAOV = Mixed ANOVA model.
Shaded cells show relative bias estimates that exceed 10%.

With respect to the power, the magnitude of the cross-level interaction effect and the sample size were the most important factors. All other simulation conditions had little or no impact on the power. The power was often slightly larger under conditions with an ICC of 0.25 compared to an ICC of 0.10 and this pattern was observed across all three models. The ICC did not have an influence on the difference in power between the three models, which implies that the difference in power between the multilevel model and mixed ANOVA model was not affected by the correlation between scores within the same patient. The factor loading conditions did not have an impact on the power, nor on the difference in power between the three models. Also with respect to the within-level variance conditions, no clear pattern could be observed.

## 3.4 | Standard error

All SEs ranged from 0.10 (rounded) to 0.30 for all three models as seen in Table 4 and Table S4. The SEs varied as a function of the sample size with SEs decreasing when the sample size increased. A similar relationship was observed between the SEs and the ICC with larger ICCs having somewhat smaller SEs, but this relationship was weaker compared to the relationship between sample size and SEs. The within-level residual variance conditions turned out to have no impact on the estimated SEs.

In the conditions with equal factor loadings, the SE estimates between the ML-MIMIC model and the multilevel model were almost equal, whereas the SE estimates of the mixed ANOVA model were somewhat larger. However, the differences in SE estimates between the mixed ANOVA model, on the one hand, and the ML-MIMIC model and multilevel model, on the other hand, became smaller with increasing sample sizes. In the conditions with unequal factor loadings, the SE estimates of the multilevel model and mixed ANOVA model were smaller compared to SEs of the same models under the equal factor loading condition. Under unequal factor loading conditions, SEs of the mixed ANOVA model were always larger than the SEs of the multilevel model (again, differences between the models decreased with increasing sample sizes). SEs of the multilevel model and mixed ANOVA model became even smaller, the more the factor loadings varied across the indicators. Furthermore, in the unequal factor loading conditions, the SEs of the multilevel model and mixed ANOVA model were smaller compared to the SEs of the ML-MIMIC model. In fact, the SEs of the ML-MIMIC model were all approximately equal in size over the loading conditions, which shows that the various factor loading conditions only had an impact on the SE estimates of the multilevel and mixed ANOVA model.

### 3.5 | Relative bias

Under all conditions, the relative parameter bias of the ML-MIMIC model was negligible, with all parameter bias values far below 10% (see Table 5 and and Table S5). None of the simulation conditions had a serious impact on the relative parameter bias of the ML-MIMIC model.

For the multilevel model and mixed ANOVA model, the relative parameter bias only varied as a function of the factor loading values. Under the condition with equal loadings, the relative parameter bias values were all acceptable and below 10%. In fact, under the equal loading condition, the parameter bias across the three models were very comparable. However, the more the factor loadings varied across indicators, the larger the relative parameter bias of the multilevel model and mixed ANOVA model. Under the condition with factor loadings varying with a magnitude of 0.05, the relative parameter bias of the multilevel model and mixed ANOVA model was around 10% or over 10%, as highlighted by the gray shaded cells in Table 5 and Table S5. The relative parameter bias of both models was already around 20% when factor loadings varied with a magnitude of 0.10. Under both varying loading conditions, the population cross-level interaction effect was (severely) underestimated when applying the multilevel model or mixed ANOVA model, resulting in negative bias. These results show that even in situations where the true factor loadings varied only minimally, relative parameter bias can already exceed 10% when applying models that rely on sum scores. Under varying loading conditions, the ML-MIMIC model clearly outperformed the multilevel model and mixed ANOVA model. Varying within-level residual variances seemed to have no impact on the difference in relative parameter bias between the models, which suggests that only the true factor loadings values affect the (relative) bias when using sum scores.

The relative parameter bias values between the multilevel model and mixed ANOVA model did not differ, meaning that none of the simulation conditions had an impact on the difference between the two models that rely on sum scores.

### 3.6 | Relative standard error bias

The relative SE bias results are presented in Table S6. Overall, the relative SE bias was well controlled below 5% for all three models. A few exceptions (7 out of 216 conditions) were observed for the ML-MIMIC model and multilevel model where the SE bias exceeded the 5%, primarily under smaller sample size conditions within the unequal within-level residual variance condition. For the ML-MIMIC model and multilevel model, the largest relative SE bias values were 0.08 and 0.085, respectively, and these were observed under the condition with unequal residual variances, a cross-level interaction effect of 0.40, an ICC of 0.25, a sample size of 50/20, and factor loadings varying with a magnitude of 0.05 (see Table S6). For the mixed ANOVA model, only in two occasions, the relative SE bias was above 5%. No clear pattern across the different simulation conditions with respect to the relative SE bias was observed.

### 3.7 | Mean squared error

Under the condition with equal factor loadings, the MSE across the different models was equal, as shown in Table S7. For the multilevel model and the mixed ANOVA model, the MSE decreased with increasing variation across the factor loadings, whereas for the ML-MIMIC model, the various factor loading conditions had no impact on the MSE. These results indicate that under varying factor loading conditions, the multilevel model and mixed ANOVA model obtain more accurate parameter estimates compared to the ML-MIMIC model. This observation implies that while the multilevel model and mixed ANOVA model underestimate the cross-level interaction effect under the condition with varying factor loadings, this underestimation occurs with less variation over the replications. On the other hand, the ML-MIMIC model shows more variation over the estimated cross-level interaction effects, but the average estimated cross-level interaction effect over the replications lies close to the population value.

Furthermore, the MSE decreased with increasing sample sizes and the MSE also seemed a little smaller under the larger ICC condition, which is line with the patterns observed for the statistical power and SE. The within-level residual variance conditions had no influence on the MSE. These results applied to all three models.

## 4 DISCUSSION

This simulation study compared the performance on estimating a cross-level interaction effect between three different statistical models in the presence of correlated variables (i.e., questionnaire items) that measure a latent variable in a multilevel context. The models compared in this simulation study varied in complexity and this complexity was partly related to the way the data were treated. The ML-MIMIC model was regarded the most complex model and used the data without conducting any aggregation steps. On the contrary, prior to applying the univariate multilevel model, the observed correlated indicators were first combined into one sum score that served as univariate outcome. For employing the mixed ANOVA model, which was the most simplified statistical model of this simulation study, the data were further aggregated by calculating two mean scores per patient representing the average scores of the two within-level covariate categories. The main goal of this study was to examine under which conditions more simplified models may become problematic when estimating a cross-level interaction effect.

The main result of this study was that the ML-MIMIC model produced unbiased cross-level interaction effect estimates under all conditions compared to the univariate multilevel model and the mixed ANOVA model. Especially when observed indicators had varying weights in relation to the latent variable, the multilevel model and the mixed ANOVA model severely underestimated the cross-level interaction effect. This result confirmed our expectation and is consistent with the arguments listed in McNeish and Wolf (2020) where the use of sum scores is discouraged as even small differences in the relationships between the indicators and the latent variable can already lead to biased results. This study showed that when unstandardized factor loadings across indicators are fairly close to each other, bias in estimating a cross-level interaction effect can already exceed 10%. This finding may have serious implications for applied researchers, because the use of sum scoring might be justified by the result of a factor analysis. Many researchers would interpret factor loading values of 1, 0.95, 0.80, 0.90, 0.85 or 1, 0.90, 0.60, 0.80, 0.70 as sufficient evidence that sum scoring is allowed, but our results revealed that even in these situations, sum scoring can be problematic. It is therefore recommended to always apply factor analysis prior to composing and using sum scores in order to verify whether the indicators have equal weight (McNeish & Wolf, 2020). However, alternative methods can imply sum scores as well, such as latent class analysis or psychometric network models, and it has been argued that in certain situations, sum scores can be justified based on a theoretical model (Edelsbrunner, 2022). Also, a reason to use sum scores is that they can more easily be compared across studies than if different weighting is used across these studies (Widaman & Revelle, 2022). Nonetheless, to ensure scores are as precise as possible, which is desired when comparing two different treatments in a clinical trial, the ML-MIMIC model should be preferred.

An additional benefit of the ML-MIMIC model over the other two models is that the ML-MIMIC model generally had larger statistical power in detecting a cross-level interaction effect. Models that obtain more power are attractive to clinicians because this can lead to smaller required sample sizes. However, the differences in power between the three models were small, which implies that the choice of which model to use based on power alone seems irrelevant. In fact, this can be a reason not to pursue employing the more complicated ML-MIMIC model.

The interpretation of the results with respect to the power cannot be dissociated with the results on the SE and relative parameter bias. The simulation results revealed that the multilevel model and mixed ANOVA model had lower SEs, especially in the conditions with unequal and slightly smaller factor loadings. This is consistent with previous research, suggesting that regression coefficients between observed variables are estimated with more precision (and hence lower SEs) compared to regression coefficients between observed variables and latent variables (Ledgerwood & Shrout, 2011; Savalei, 2014; Wang & Rhemtulla, 2021). Particularly, the use of RML for estimating the parameters of the ML-MIMIC model has contributed to the observed lower precision when applying the ML-MIMIC model (Savalei, 2014). This result was also noted when considering the MSE, where the multilevel model and mixed ANOVA model turned out to be more accurate compared to the ML-MIMIC model under the varying factor loading conditions.

Intuitively, smaller SEs and better accuracy would lead to more power. A recent simulation study indeed showed that a multiple regression on observed sum scores had larger power compared to a latent variable model (Wang & Rhemtulla, 2021). However, in that study, factor loadings were not varied across observed indicators. In contrast, our simulation study particularly examined the effect of varying factor loadings on parameter bias and we discovered that under varying factor loading conditions, the multilevel model and mixed ANOVA underestimated the cross-level interaction effect. Underestimation indicates that the effect size is smaller, which has a negative impact on power. Consequently, the smaller SEs and larger negative bias work in opposite directions concerning the power for the multilevel and mixed ANOVA model. Overall, although the ML-MIMIC model produced larger SEs, it did not induce bias in estimating the cross-level interaction, which, in the end, resulted in slightly higher power in detecting this interaction effect. Therefore, when considering the power, SEs, and relative bias results, the ML-MIMIC model is preferred. These conclusions should be interpreted considering the current simulation conditions, because it is unclear whether the multilevel model and mixed ANOVA will always underestimate the cross-level interaction effect.

With respect to the type I error rates, our simulation study showed that the ML-MIMIC model had more often a type I error rate that was significantly larger than the nominal alpha level of 0.05 compared to the univariate multilevel model and mixed ANOVA model, especially in the smaller sample size conditions. Inflated type I error rates for the ML-MIMIC model under small sample sizes were also found in previous studies (Cao et al., 2019; Finch & French, 2011). This finding calls for some restraint when recommending the ML-MIMIC model for the analysis of clinical trials with small sample sizes, because a decent type I error control rate is necessary when considering an analytic tool to be suitable for clinical trials. For future research, other simulation conditions should be considered to further study the type I error rates in ML-MIMIC models, such as different random slope variances (see further on) or different estimation methods. For the univariate multilevel and mixed ANOVA model, the type I error rates were well under control. A possible explanation for this finding could be the use of the Satterthwaite approximation for calculating the degrees of freedom, as it has been shown that the Satterthwaite approximation resulted in improved type I error rates compared to likelihood ratio tests and Wald-type tests with chi-square approximation when testing intercepts and slopes in mixed effects linear models under small sample size conditions (Kuznetsova et al., 2017; Manor & Zucker, 2004).

Regarding the impact of the simulation conditions on the performance of the models and the comparison between the models, it can be concluded that the within and between-level sample sizes and the magnitude of the cross-level interaction were most relevant for the performance of the models on their own, which is in line with previous research (Cao et al., 2019; Mathieu et al., 2012; Wang & Rhemtulla, 2021). Larger sample sizes lead to larger ASRs, larger power, smaller SEs, and smaller MSEs. The magnitude of the interaction effect was highly associated with power. The power under a cross-level interaction effect of 0.20 never exceeded 80%, which does not correspond to the work by Cao et al. (2019), where a cross-level interaction effect of 0.20 already lead to power of 80% under the larger sample size conditions. The reason for this discrepancy is because in the current simulation study, random slope variation was incorporated during the data-generation process and this random slope variance was set at a relatively large value (0.50) compared to the study by Cao et al. (2019). A larger random slope variation results in larger SEs and therefore lower power. In fact, Cao et al. did not incorporate random slope variance during the data-generation process, but rather restricted the random slope variation to a specified value of 0.10 during estimation of the ML-MIMIC model. Important to note is that when there clearly exists random slope variation, this should be included during model fitting as otherwise the SEs will be downward biased (Bell et al., 2019). However, for power calculations during the design stage, providing an estimate of the random slope variation is often not straightforward, but setting this artificially at a low value might lead to less required patients, which may result in underpowered studies when the actual random slope variation is larger. The effect of varying random slope variances was not part of this research, but manipulating random slope variation would be an interesting feature to include in future studies that investigate the performance of ML-MIMIC models. In addition to varying the random slope variance, future studies on the performance of ML-MIMIC models in comparison to univariate regression models could also study the impact of varying the covariance between the random intercept and random slope.

The impact of the ICC on the estimation of the cross-level interaction effect was small and did not result in noteworthy findings, except that a small ICC had lower ASRs under the smallest sample size condition, especially for the multilevel model. Overall, the negligible impact of the ICC on the estimation of the cross-level interaction effect was in line with Cao et al. (2019). In fact, Cao et al. (2019) found that the ICC only had a large impact on the SE and power when a covariate interaction was located at the between-level, where a larger SE and lower power was observed with increasing ICCs. Clinicians who want to adopt the ML-MIMIC model should be aware of this result when they want to test an interaction effect between two dichotomous covariates located at the patient level.

When considering the impact of the simulation conditions on the comparison between the models, it was already emphasized that the varying factor loading conditions had a major impact on the differences in relative bias between the three models. In contrast, the different within-level residual variance conditions had no impact on the relative bias, which suggests that a justified application of sum scoring is primarily affected by the factor loading values. This finding adds valuable information to the work of McNeish and Wolf (2020), because it implies that mainly unstandardized factor loadings, rather than standardized factor loadings, should be studied when planning to use sum scores in this specific ML-MIMIC context. However, to gain a better understanding of the effect of varying residual variances in multilevel data, it would be interesting to consider other (simulation) settings. For example, including conditions with varying between-level residual variances as well (in addition to varying within-level residual variances), looking at more complicated models with more latent variables, or focusing on main effects of covariates on the within- and between-level would all be relevant topics for future work.

The differences between the univariate multilevel model and the mixed ANOVA model were small. Especially with respect to the relative bias, SE, and MSE, the differences were negligible. This corresponds with another study where no difference in the parameter estimate and SE was found between a multilevel model and a mixed ANOVA model that relies on summary measures using aggregated data when applied to multicenter intervention studies (Moerbeek et al., 2003). However, the multilevel model has several advantages over the mixed ANOVA model. A multilevel model can handle unbalanced measurements, it has the ability to accommodate missing data (Krueger & Tian, 2004), and in a multilevel model, it is possible to examine varying drug effects across each patient by estimating a random effect (Kessels et al., 2019). These advantages make the multilevel model way more flexible compared to the mixed ANOVA model. In addition, we found that the multilevel model had slightly larger power in detecting a cross-level interaction effect compared to the mixed ANOVA model, which provides another argument for preferring multilevel modeling. The ICC had very little impact on the power difference between the multilevel model and the mixed ANOVA model, which suggests that the power was not affected by the correlation between scores within a patient. Therefore, this finding is not in line with our expectations that aggregating data could lead to reduced measurement error due to correlated within-patient measurements and therefore to larger power. However, because our study was not fully focused on studying this trade-off, the simulation conditions were rather limited to investigate this more deeply. For example, different ICC values that range from very low to very high could be interesting to incorporate in the simulation design as well as conditions with unbalanced data. Moreover, the data were simulated under a latent variable model that is not an ideal setting for comparing models that cannot deal with latent variables appropriately.

Since our simulation code was used to simulate power, it can also be administered by applied researchers and statisticians to calculate power and required sample sizes during the design stage when planning to use the ML-MIMIC model for testing a cross-level interaction effect. To perform a power analysis, one needs to consider the factor loading values, main and interaction effects, factor and indicator variances, and the random slope variance. However, specifying these parameter values during the design stage can be quite challenging, especially when information on parameter values of previous studies or expert knowledge is lacking. We therefore advice researchers to perform sensitivity analyses to study the effect of multiple parameter values on the power and to assess under which conditions the model obtains well-controlled type I error rates. As outlined above, the magnitude of the random slope variance has a negative impact on power that could lead to underpowered studies when the random slope variance turns out to be larger than hypothesized. Although manipulating the random slope variance was not part of our study, a nice feature of our simulation code is that the random slope variance can be adapted. This allows researchers to study multiple scenarios and to avoid underpowered studies when one selects the largest realistic random slope variance. Furthermore, our study revealed that the magnitude of the cross-level interaction effect was the only condition, besides sample size, that was highly related with power. This indicates that the power for the cross-level interaction effect test is less affected by factor loading values, the ICC, and within-level residual variances, but that does not necessarily guarantee that this will also be the case in other situations.

When planning to use the ML-MIMIC model in their studies, researchers should be aware about the design used in this simulation study in comparison to their own. In our simulation study, there were two dichotomous covariates at the within-patient and between-patient level. Like regular multilevel regression models, the ML-MIMIC model can easily be extended with more categorical and continuous covariates at both levels. Considering the applied example introduced at the beginning of this paper, the covariate at the within-patient level (study period) models a change in time and each time point is a collection of multiple events. This implies that the covariate study period is not treated any different than we would treat, for example, the variable gender of patients in a cluster randomized trial where patients (level 1) are nested within hospitals (level 2). For longitudinal studies, where, for example, a patient reported outcome measure is repeatedly measured over time within patients and where each time point represents one observation of the patient reported outcome, time can be modeled in an ML-MIMIC model as a within-patient covariate with multiple categories representing the time points (e.g., by using dummy variables). However, this specific type of design was not examined in this simulation study, so generalizing the results to those kind of longitudinal studies should be made with caution.

Furthermore, when applying the ML-MIMIC model in practice, it is important to verify the measurement invariance assumptions. In ML-MIMIC models, invariance of factor loadings and intercepts across groups is implicitly assumed (as was the case in our simulations), but it is advised to test these strict assumptions (Kessels et al., 2021; Kim et al., 2015), in order to ensure that unbiased covariate effects are obtained. Research has shown that in situations where the measurement invariance assumption is partly violated (some but not all intercepts and/or factor loadings are invariant), covariate effects are still unbiased as long as the model is corrected for the invariant parameters (Hsiao & Lai, 2018). Researchers should be aware that accuracy of covariate effects is affected by the measurement invariance status (Hsiao & Lai, 2018). In addition, researchers should also consider the issue of response shift, the phenomenon where patients alter the way they conceptualize and evaluate items over time (Carlier et al., 2019). Response shift might therefore affect measurement invariance and impact the interpretation of constructs over time. The presence of response shift should therefore also be investigated in ML-MIMIC models and it has been shown that response shift plays a role in studies using patient reported outcomes, such as quality of life outcomes in cancer studies (Ilie et al., 2019). However, how to deal with these possible measurement invariance violations is beyond the scope of this study and we refer for this to other work (Carlier et al., 2019; Kessels et al., 2021; Kim et al., 2015).

In sum, this research showed that when a cross-level interaction effect is of focal interest to, for example, assess the difference between two treatments on a latent variable measured at multiple occasions, the ML-MIMIC model produced unbiased cross-level interaction effects under all conditions. In contrast, models that rely on sum scores produced severely negatively biased estimates under varying factor loading conditions. All models had comparable power. Since clinicians and applied researchers are generally only interested in power and bias, the ML-MIMIC model is the preferred model, but considering the type I error rates, caution is advised with small sample sizes.

## CONFLICT OF INTEREST STATEMENT

The authors have declared no conflicts of interest.

## DATA AVAILABILITY STATEMENT

Computer code used to conduct the simulation study is available on github at https://github.com/ebkesselsr/MLMIMIC. This repository also contains an R-program to perform a power analysis using the ML-MIMIC model.

## ORCID

*Rob Kessels* https://orcid.org/0000-0002-2479-3872
*Mirjam Moerbeek* https://orcid.org/0000-0001-5537-1237

## REFERENCES

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: Making an informed choice. *Quality & Quantity*, *53*, 1051–1074.

Bollen, K. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Brown, H., & Prescott, R. (2015). *Applied mixed models in medicine* (3rd ed.). John Wiley & Sons.

Candel, M., & Van Breukelen, G. (2015). Sample size calculation for treatment effects in randomized trials with fixed cluster sizes and heterogeneous intraclass correlations and variances. *Statistical Methods in Medical Research*, *24*, 557–573.

Cao, C., Kim, E. S., Chen, Y., Ferron, J., & Stark, S. (2019). Exploring the test of covariate moderation effects in multilevel MIMIC models. *Educational and Psychological Measurement*, *79*, 512–544.

Carlier, I. V. E., van Eeden, W. A., de Jong, K., Giltay, E. J., van Noorden, M. S., van der Feltz-Cornelis, C., Zitman, F. G., Kelderman, H., & van Hemert, A. M. (2019). Testing for response shift in treatment evaluation of change in self-reported psychopathology amongst secondary psychiatric care outpatients. *International Journal of Methods in Psychiatric Research*, *28*, e1785.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2019). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*, 558–575.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, *14*, 1–11.

Finch, W. H., & French, B. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*, 229–252.

Edelsbrunner, P. (2022). A model and its fit lie in the eye of the beholder: Long live the sum score. *Frontiers in Psychology*, *13*, 986767.

Hallquist, M., & Wiley, J. (2018). MplusAutomation: an R-package for facilitating large-scale latent variable analyses in M-plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 621–638.

Hox, J., & Maas, C. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 157–174.

Hox, J., Moerbeek, M., & Van De Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Taylor & Francis Group.

Hsiao, Y. Y., & Lai, M. H. (2018). The impact of partial measurement invariance on testing moderation for single and multi-level data. *Frontiers in Psychology*, *9*, 740.

Ilie, G., Bradfield, J., Moodie, L., Lawen, T., Ilie, A., Lawen, Z., Blackman, C., Gainer, R., & Rutledge, R. D. H. (2019). The role of response-shift in studies assessing quality of life outcomes among cancer patients: A systematic review. *Frontiers in Psychology*, *9*, 783.

Jak, S. (2019). Cross-level invariance in multilevel factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 607–622.

Jak, S., Oort, F., & Dolan, C. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 265–282.

Jak, S., Oort, F., & Dolan, C. (2013). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 31–39.

Jöreskog, K., & Goldberger, A. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631–639.

Kessels, R., Bloemers, J., Tuiten, A., & Van Der Heijden, P. G. M. (2019). Multilevel analyses of on-demand medication data, with an application to the treatment of Female Sexual Interest/Arousal Disorder. *PLoS One*, *14*, e0221063.

Kessels, R., Moerbeek, M., Bloemers, J., & van der Heijden, P. G. (2021). A multilevel structural equation model for assessing a drug effect on a patient-reported outcome measure in on-demand medication data. *Biometrical Journal*, *63*, 1652–1672.

Kim, E. S., Yoon, M., Wen, Y., Luo, W., & Kwok, O. (2015). Within-level group factorial invariance with multilevel data: Multilevel factor mixture and multilevel MIMIC models. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 603–616.

Krueger, C., & Tian, L. (2004). A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological Research for Nursing*, *6*, 151–157.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.

Ledgerwood, A., & Shrout, P. E. (2011). The trade-off between accuracy and precision in latent variable models of mediation processes. *Journal of Personality and Social Psychology*, *101*, 1174–1188.

Maas, C., & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 86–92.

Manor, O., & Zucker, D. M. (2004). Small sample inference for the fixed effects in the mixed linear model. *Computational Statistics & Data Analysis*, *46*, 801–817.

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, *97*, 951–966.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (Vol. 1). Psychology Press.

McNeish, D., & Wolf, M. (2020). Thinking twice about sum scores. *Behavioral Research Methods*, *52*, 1–19.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Moerbeek, M. (2012). Sample size issues for cluster randomized trials with discrete-time survival endpoints. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *8*, 146–158.

Moerbeek, M., van Breukelen, G. J., & Berger, M. P. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, *56*, 341–350.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (7th ed.). Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 599–620.

Osoba, D. (2011). Health-related quality of life and cancer clinical trials. *Therapeutic Advances in Medical Oncology*, *3*, 57–71.

Pompili, C., Koller, M., Velikova, G., Franks, K., Absolom, K., Callister, M., Robson, J., Imperatori, A., & Brunelli, A. (2018). EORTC QLQ-C30 summary score reliably detects changes in QoL three months after anatomic lung resection for non-small cell lung cancer (NSCLC). *Lung Cancer*, *123*, 149–154.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. URL https://www.R-project.org/

Reck, M., Taylor, F., Penrod, J. R., DeRosa, M., Morrissey, L., Dastani, H., Orsini, L., & Gralla, R. J. (2018). Impact of nivolumab versus docetaxel on health-related quality of life and symptoms in patients with advanced squamous non–small cell lung cancer: Results from the CheckMate 017 study. *Journal of Thoracic Oncology*, *13*, 194–204.

Roesch, S., Aldridge, A., Stocking, S., Villodas, F., Leung, Q., Bartley, C. E., & Black, L. J. (2010). Multilevel factor analysis and structural equation modeling of daily diary coping data: Modeling trait and state variation. *Multivariate Behavioral Research*, *45*, 767–789.

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

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 149–160.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press Inc.

Tuiten, A., van Rooij, K., Bloemers, J., Eisenegger, C., van Honk, J., Kessels, R., Kingsberg, S., Derogatis, L. R., de Leede, L., Gerritsen, J., Koppeschaar, H. P. F., Olivier, B., Everaerd, W., Frijlink, H. W., Höhle, D., de Lange, R. P. J., Böcker, K. B. E., & Pfaus, J. G. (2018). Efficacy and safety of on-demand use of 2 treatments designed for different etiologies of female sexual interest/arousal disorder: 3 randomized clinical trials. *Journal of Sexual Medicine*, *15*, 201–216.

Van Nes, Y., Bloemers, J., van der Heijden, P., Van Rooij, K., Gerritsen, J., Kessels, R., DeRogatis, L., & Tuiten, A. (2017). The Sexual Event Diary (SED): Development and validation of a standardized questionnaire for assessing female sexual functioning during discrete sexual events. *Journal of Sexual Medicine*, *14*, 1438–1450.

Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, *4*, 1–17.

Widaman, K. F., & Revelle, W. (2022). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, *55*, 1–19.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Kessels, R., & Moerbeek, M. (2023). A comparison of the multilevel MIMIC model to the multilevel regression and mixed ANOVA model for the estimation and testing of a cross-level interaction effect: A simulation study. *Biometrical Journal*, *65,* 2200112. https://doi.org/10.1002/bimj.202200112