Behaviormetrika
https://doi.org/10.1007/s41237-020-00130-8

ORIGINAL PAPER

Small and negative correlations among clustered observations: limitations of the linear mixed effects model

Natalie M. Nielsen[1] · Wouter A. C. Smink[1,2] · Jean-Paul Fox[1]

Received: 12 February 2020 / Accepted: 31 December 2020

Abstract

The linear mixed effects model is an often used tool for the analysis of multilevel data. However, this model has an ill-understood shortcoming: it assumes that observations within clusters are always positively correlated. This assumption is not always true: individuals competing in a cluster for scarce resources are negatively correlated. Random effects in a mixed effects model can model a positive correlation among clustered observations but not a negative correlation. As negative clustering effects are largely unknown to the sheer majority of the research community, we conducted a simulation study to detail the bias that occurs when analysing negative clustering effects with the linear mixed effects model. We also demonstrate that ignoring a small negative correlation leads to deflated Type-I errors, invalid standard errors and confidence intervals in regression analysis. When negative clustering effects are ignored, mixed effects models incorrectly assume that observations are independently distributed. We highlight the importance of understanding these phenomena through analysis of the data from Lamers, Bohlmeijer, Korte, and Westerhof (2015). We conclude with a reflection on well-known multilevel modelling rules when dealing with negative dependencies in a cluster: negative clustering effects can, do and will occur and these effects cannot be ignored.

Keywords Negative clustering effects · Negative cluster correlation · Negative ICC · Covariance structure models · Linear mixed effects model

Communicated by Yasuo Miyazaki.

✉ Jean-Paul Fox
J.P.Fox@utwente.nl

[1] Department of Research Methodology, Measurement & Data Analysis, University of Twente, P.O. Box 217, 7500 A. E. Enschede, The Netherlands

[2] Department of Psychology, Health & Technology, University of Twente, P.O. Box 217, 7500 A. E. Enschede, The Netherlands

Published online: 24 January 2021

1 Introduction

The popularity of the linear mixed effects models (e.g., random effect models, multilevel models) is intuitively explained by the variety of different names under which the family of statistical models for clustered data are known. In clustered data (e.g., hierarchical data, multilevel data) observations are associated, and not independently observed (Dorman 2008). Dependencies among observations in clusters can be expressed as a correlation, where positively correlated observations share similar information (Kenny and Judd 1986), and are not as informative as independent observations (Galbraith et al. 2010). Only when accounting for the correlation between clustered observations, correct statistical inferences can be made. In the well-known multilevel modelling framework (McCulloch et al. 2008), this correlation is modelled by a random effect –also known as a latent variable– where clustered observations are positively correlated since they share the same random effect. The variance of the random effect then determines the strength of the correlation. As a result, the multilevel modelling frameworks restricts correlations to be positive, since a variance parameter cannot be negative.

However, negative correlations among clustered observations *can* and *do* occur (Kenny et al. 2002). For instance, when fixed resources are divided among group members, in non-random sampling when dissimilar groups are sampled by intention, or when there is competitive social interaction: when individuals compete for a scarce (and fixed) set of resources (e.g. litter mates are negatively correlated in terms of food, water and living space), and the speaking time of one individual is at the expense of another individual (Pryseley et al. 2011). When observations within clusters are negatively associated, observations within clusters are less alike than observations from different clusters (Kenny and Judd 1986). From a sampling perspective this is sometimes referred to as the situation where observations within a cluster are even less alike than under random assignment of observations to clusters (Molenberghs and Verbeke 2007; Verbeke and Molenberghs 2003; Molenberghs and Verbeke 2011). Negative intra-cluster correlations (ICC; see Table 1 for an overview of the often used abbreviations) can also be detected

Table 1 List with the often used abbreviations in the current article

Abbreviation	Full term
CI	Confidence interval
CR	Coverage rate
ICC	Intra-cluster correlation, intra-class correlation, intra-class correlation coefficient
LM	Linear model
LME	Linear mixed effects (model)
CSM	Covariance structure model
SE	Standard error
VIF	Variance inflation factor, also known as the design effect

in randomized experiments, when evaluating the effects of covariates that vary systematically within each cluster (Norton et al. 1996).

In general, it is well-known that ignoring a small positive clustering leads to the incorrect assumption that the observations are independently distributed. It is our aim to extend this knowledge with the current article: we will show that ignoring a positive *and negative* clustering leads to a violation of the independence assumption. In fact, any violation of the independence assumption (positive and negative) results in inaccurate Type-I errors, which increase the risk of accepting an incorrect hypothesis (Clarke 2008). Barcikowski (1981) quantified the effects of ignoring small positive correlations in clustered observations in a two-level study design (with a group and an individual level). He showed that, when having ten observations per group, even the ignorance of an ICC as small as .01 can lead to an inflation of making a Type-I error: a regression effect will be assumed to be significant with a significance level of 5% although the true significance level equalled 6%. Furthermore, Barcikowski showed that the Type-I error increased for increasing values of the ICC. For an ICC of .05 the Type-I error rate is .11 and for an ICC of .40 the Type-I error is .46. Moreover, by increasing the number of observations per group the Type-I error is even more inflated (the findings of Barcikowski are in line with those of many others, see for example Clarke 2008; Dorman 2008; Rosner and Grove 1999).

As negative clustering effects are largely unknown to the sheer majority of the research community, we conducted a simulation study to detail the bias that occurs when analysing negative clustering effects with the linear mixed effects model in similar fashion as Barcikowski (1981). Towards that end, we demonstrate that ignoring a small negative correlation leads to deflated Type-I errors, invalid standard errors and confidence intervals in regression analysis. We highlight the importance of understanding these phenomena through analysis of the data from Lamers et al. (2015). We conclude with an updated reflection on well-known multilevel modelling rules. In the remainder of this section, we discuss negative dependencies between observations in clustered data, show how the linear mixed effects model (LME) deals with negative clustering effects, and reflect on why the LME should include negative variance components. Note that the LME is used only to quantify bias when ignoring negative clustering effects. We stress that the LME is not designed to model negatively correlated observations and it should not be expected to perform properly to those cases. The covariance structure model (CSM) is introduced to deal with negative within-cluster correlations and this model is used in our real data example to examine negative clustering effects.

1.1 Type-I errors and positive and negative dependencies between observations

The inflation of the Type-I error under violated of the independence assumption can be explained by the variance inflation factor (VIF), also referred to as the design effect (Kish 1965). In case of cluster sampling, a design effect that is greater than one is known to indicate a positive within-cluster correlation, indicating that observations are not independent of each other. When VIF $>$ 1 the precision of cluster sample estimates are less than that of those based on a simple random sample with a similar size. The homogeneity in clustered observations leads to less information in comparison to an independent random sample. When ignoring a small positive ICC, the VIF is underestimated, which leads to an underestimation of the standard errors (i.e. overestimating the precision), and the corresponding confidence intervals (CIs) are too narrow, and effect sizes will then also be incorrect as they depend on standard error (SE) estimates (Hox et al. 2010; Kenny et al. 2002). When the CI of an estimate is too narrow, there is an increase in the probability to reject a correct null hypothesis, which corresponds to an inflation of the Type-I error.

Although a few researchers have reported about negative clustering effects (Kenny et al. 2002; Molenberghs and Verbeke 2007, 2011; Pryseley et al. 2011; Oliveira et al. 2017; Verbeke and Molenberghs 2003; El Leithy et al. 2016; Klotzke and Fox 2019a, b; Loeys and Molenberghs 2013), the effects of ignoring negatively clustered observations has hardly been recognized. Because negative clustering effects are not considered by the majority of the multilevel modelling community, these effects are not well understood. This is partly caused by the fact that the mixed effect models (to which we also refer as ‘LME’, see Table 1) can only describe positive correlations, and cannot handle negative correlations among clustered observations (Searle et al. 1992). In the next section, it is explained *why* negative clustering effects cannot be modeled with LME, and we reflect on the key principles of negative ICCs.

The LME cannot identify any negative correlation and will assume independently distributed observations. Researchers usually fix negative ICC estimates to zero and ignore any negative correlation within a cluster (Baldwin et al. 2008; Maas and Hox 2005). Furthermore, it is sometimes concluded that negative ICC estimates are caused by a small between-cluster variance (smaller than the within-cluster variance) and that such a small between-group variance can be ignored (Giberson et al. 2005; Krannitz et al. 2015; Langfred 2007). Other researchers relate negative ICC estimates to sampling error (cf. Eldridge et al. 2009), which can be ignored. Others –such as Baldwin et al. (2008), Norton et al. (1996), and Rosner and Grove (1999)– stated that the Type-I error will be deflated when fixing a negative ICC to zero.

1.2 The linear mixed effects model and negative dependencies

In this study, we consider two models: the LME and a covariance structure model (CSM, see Table 1). Both models can assess clustered data, where a one-way classification structured is considered. In the one-way classification, a common correlation is assumed among clustered observations, and observations from different clusters are assumed to be independently distributed.

1.2.1 The linear mixed effects model

Without making an explicit distinction between a random variable and a realized value, the LME for the one-way classification is given by

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + e_{ij}, \tag{1}$$

referred to as the random intercept model, where the random effect is assumed to be normally distributed, $u_j \sim \mathcal{N}(0, \tau)$, and the error term is also assumed to follow a normal distribution $e_{ij} \sim \mathcal{N}(0, \sigma^2)$. A total of $j = 1, \dots, m$ clusters are assumed with each $i = 1, \dots, n$ observations, which leads to a balanced study design. The common intercept and regression parameter are referred to as β_0 and β_1, respectively. The outcome y_{ij} is assumed to be independently distributed given the random effect u_j.

It can be shown that the random effect u_j defines a variance-covariance structure for the data. The covariance between two clustered observations is equal to (suppressing the conditioning on $\mathbf{X}_j$)

$$\begin{aligned} cov\left(y_{ij}, y_{lj}\right) =& cov\left(E\left(y_{ij} \mid u_j\right), E\left(y_{lj} \mid u_j\right)\right) + E\left(cov\left(y_{ij}, y_{lj} \mid u_j\right)\right) \\ =& cov\left(\beta_0 + \beta_1 X_{ij} + u_j, \beta_0 + \beta_1 X_{lj} + u_j\right) + 0 \\ =& cov\left(u_j, u_j\right) = var\left(u_j\right) = \tau, \end{aligned} \tag{2}$$

and the variance of an observation equals

$$\begin{aligned} var\left(y_{ij}\right) =& var\left(E\left(y_{ij} \mid e_{ij}\right)\right) + E\left(var\left(y_{ij} \mid u_j\right)\right) \\ =& \sigma^2 + \tau. \end{aligned} \tag{3}$$

The dependence structure of the observations in the clusters $\mathbf{y}_j$ modelled by the random effect u_j is given by

$$var\left(\mathbf{y}_j\right) = \boldsymbol{\omega} = \begin{bmatrix} \sigma^2 + \tau & \tau & \dots & \tau \\ \tau & \sigma^2 + \tau & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \tau & \dots & \tau & \sigma^2 + \tau \end{bmatrix}. \tag{4}$$

Thus, $\boldsymbol{\omega} = \sigma^2 \mathbf{I}_n + \mathbf{J}_n \tau$ represents the dependence structure implied by the random effect u_j. The $\mathbf{J}_n$ is a matrix of dimension n with all elements equal to one and $\mathbf{I}_n$ is the identity matrix.

1.2.2 The covariance structure model

An alternative specification of the LME in Equation (1) can be given. In this approach, the covariance structure is modelled directly and not indirectly through the specification of a random effect. The distribution of clustered observations is assumed to be multivariate normally distributed with a covariance matrix $\boldsymbol{\omega}$,

$$\mathbf{y}_j = \beta_0 + \beta_1 \mathbf{X}_j + \mathbf{e}_j, \tag{5}$$

where the errors are multivariate normally distributed, $\mathbf{e}_j \sim \mathcal{N}(0, \boldsymbol{\omega})$. We refer to the model in Equation (5) as the CSM. The development and use of the covariance structure model has a long history, which is intertwined with the development of factor models. Classic works in covariance structure modelling can be found in that tradition (e.g., Bock and Bargmann 1966; Jöreskog 1969, 1971). Fox et al. (2017), Klotzke and Fox (2019a), and Klotzke and Fox (2019b) developed a novel Bayesian modelling framework in which they directly modelled the covariance structure of more complex dependence structures. In their *Bayesian covariance structure modelling* (BCSM) approach, dependencies among observations that are usually modelled through random effects are modelled directly through covariance parameters under the BCSM.

When comparing the modelling structure of the CSM (also referred to as BCSM; Klotzke and Fox 2019a, b) with that of the LME, it can be seen that the τ is restricted to be positive in the model in Equation (1), since it represents a *variance* parameter. However, in the model in Equation (5), the τ parameter can also be negative since it represents a *covariance* parameter. This makes the CSM more general than the LME, since the covariance parameters can be positive and negative, which allows for more flexibility in specifying complex dependence structures (cf. Klotzke and Fox 2019a, b).

1.3 The linear mixed effects model with negative variance components

There are some restrictions on the variance-covariance components in the CSM. From the definition of the error variance follows directly that the σ^2 is restricted to be greater than zero (i.e. $0 < \sigma^2 < \infty$). However, the covariance parameter τ is not necessarily restricted to be greater than zero. Under the CSM, the covariance matrix $\boldsymbol{\omega}$ needs to be positive definite, which that implies the restriction –for balanced designs– $n\tau + \sigma^2 > 0$. This important result follows from Rao (1973, p. 32), where the determinant of a compound symmetry covariance matrix is expressed as

$$\begin{aligned} det\left(\sigma^2\mathbf{I}_n + \tau\mathbf{J}_n\right) =& det\left(\sigma^2\mathbf{I}_n\right)\left(1 + \tau\mathbf{1}_n^t\mathbf{1}_n/\sigma^2\right) \\ =& \sigma^2\left(1 + n\tau/\sigma^2\right) = n\tau + \sigma^2, \end{aligned} \tag{6}$$

and the covariance matrix is positive definite if the determinant is greater than zero . Subsequently, τ needs to be greater than $-\sigma^2/n$. However, when modeling the covariance structure with the LME, the τ is restricted to be greater than zero, since it represents the random intercept variance. In the literature, it has been shown that the maximum likelihood estimate of the random effect variance can become negative (Kenny et al. 2002; Molenberghs and Verbeke 2007, 2011; Pryseley et al. 2011; Oliveira et al. 2017; Verbeke and Molenberghs 2003; El Leithy et al. 2016; Klotzke and Fox 2019a, b; Loeys and Molenberghs 2013). For the (one-way) LME (for balanced groups), two sums of squares are considered to estimate the covariance components τ and σ^2,

$$\begin{aligned} SS_A &= \sum_{j=1}^{m} n\left(\bar{y}_j - \bar{y}\right)^2, \\ SS_E &= \sum_{j=1}^{m}\sum_{i=1}^{n}\left(y_{ij} - \bar{y}_j\right)^2. \end{aligned} \tag{7}$$

Consider the sum of squares SS_A, which has as expected value $n\tau + \sigma^2$. It follows that, $\hat{\tau} = SS_A/(nm) - \sigma^2/n$, which leads to a negative estimate of τ if $\sigma^2 > SS_A/m$.

This scenario is often neglected or referred to as statistically incorrect, restricting τ to represent a positive covariance among clustered observations.

For $\tau > 0$, the ICC is often interpreted as the ratio of variance explained by the clustering of observations in comparison to the total variance in the data; $\rho = \tau/(\tau + \sigma^2)$ (Raudenbush and Bryk 2002; Snijders and Bosker 2012; Oliveira et al. 2017). However, the ICC can also be considered to quantify the degree of resemblance or average similarity of observations within a cluster, or as the 'average correlation' in each cluster (Kenny and Judd 1986; Kenny et al. 2002). Then, conceptually, a negative covariance ($\tau < 0$) represents a negative ICC. In that case, ρ becomes negative, and the ICC represents a negative association among clustered observations (i.e. observations within clusters are less alike than observations from different clusters). A negative ICC simply represents the opposite of a positive ICC: if an observation in a cluster is below the population mean, then it is more likely that another value in that cluster is above the population mean if the observations are negatively correlated (Kenny and Judd 1986; Kenny et al. 2002).

Even though negative clustering effects were discussed previously by others (cf. Oliveira et al. 2017; Pryseley et al. 2011), there still appears to be a lack of awareness about these effects. As the LME comes with the restriction that observations need to be positively clustered, several suggestions can be found in the literature that τ should be set to zero, when the ICC estimate becomes negative (see for example Baldwin et al. 2008; Maas and Hox 2005; Gibson et al. 2015; Krannitz et al. 2015; Langfred 2007; Eldridge et al. 2009). In the next section, we will discuss our simulation study which aims to not only show that fixing the ICC to 0 is –in fact– wrong, but also quantify the bias that arises when negative clustering effects are ignored.

2 Methods

The object of the simulation study was to quantify the estimation errors that are made in the statistical analysis, when ignoring a (small) negative or positive clustering in the data. While ignoring a small positive and negative ICC, the accuracy of the intercept and regression parameter estimates, the SEs, the 95% CIs, and the inflation and/or deflation of Type-I errors were assessed.

A large number of clusters can compensate for biases that occur due to ignoring small ICC values, where the number of observations per cluster affects the magnitude of the Type-I error (Barcikowski 1981; Dorman 2008). In this study, $m = 10$ clusters were considered to assess the bias, and the number of observations per cluster $n = \{10, 15, 30\}$ varied to examine the effects for small sample sizes. The error variance σ^2 was fixed to one, where τ took on values ranging from $-\sigma^2/n$ to 0, in incremental steps of .02 and from 0 to .20 in steps of .05. The step size for $\tau < 0$ were set to .005 for $n = 30$, so that at least two negative values of τ were used for data generation. For each condition a lower-bound was defined, $\tau_{Lb} = -\sigma^2/n$, which represented the lower bound on the allowable negative correlations in the clustered data. Only the CSM allowed generating data with negative correlations, where under the LME data could only be generated with positive correlations. The true value of the intercept and slope parameter were set to 0 and 0.1 ($\beta_0 = 0, \beta_1 = 0.1$), respectively, and those parameters were considered to assess the effects of ignoring the correlation in the data. An intercept-only model and an intercept-slope were used to simulate data. However, note that we could have evaluated the intercept estimates under the intercept-slope model, since the slope parameter estimates would not be affected by the considered dependence structures, and could not interfere with inferences about the intercept.

A Monte Carlo simulation study was used to evaluate the appropriateness of the LME as an analysis tool for negative clustering and small positive clustering effects. Therefore, data were generated under the LME (see Equation (1)) which only generated data with positively correlated observations in clusters. Data with negatively or positively correlated observations were generated according to the CSM (Equation 5), with a covariance matrix displaying the dependence structure in a cluster. The LME was fitted to data generated under the LME and under the CSM, and the parameter estimates (REML) for the LME were obtained using the `lme4`-package in `R` (Bates et al. 2015; R Core Team 2020). A linear regression model (LM) was also fitted that ignored any correlation in the clustered observations (positive or negative). Parameter estimates for the LM were obtained using the `lm`-function in the `stats`-package in `R` (the `stats`-package is a core package in R Core Team 2020). The `lmerTest`-package was used to compute *p*-values of the fixed effects in the LME (Kuznetsova et al. 2017), where Satterthwaite's degrees of freedom method was used (Satterthwaite 1946). For each condition, a total of 1,000 data sets were generated, and reported per condition the average *p*-value, coverage rate (CR), SE, and bias of τ estimate. The bias was computed as the average over replications between the estimate and the true parameter value. The average SE estimate across replications was computed and reported as the SE. The CR was computed as the percentage of times the true parameter value was located in the 95% confidence interval over data replications. The average of the computed *p*-values across data replications was considered to be the average *p*-value. The ICC values were close to the τ values, since the error variance σ^2 was set to 1. Therefore, results were mostly reported for τ and only sparsely for the ICC.

3 Results

We use this section to first discuss the simulation study. We also include a real data example to illustrate that not accounting for negative clustering can lead to an increase of a Type-II error. For this example, we re-used data from Lamers et al. (2015), and Smink et al. (2019).

3.1 Coverage rate and Type-I error

The 95% CRs were estimated for the intercept-only condition ($\beta_0 = 0$, $\beta_1 = 0$), and the intercept-slope condition ($\beta_0 = 0$, $\beta_1 = .1$; see Equation 1). For the intercept-only condition, the results concerning the intercept are presented under the label

β_0. For the intercept-slope condition, the results concerning the slope are presented under the label β_1. In Table 2, results are given of data generated under the LME, which were analysed with the LME (model variant 1 in Table 2) and the LM (model variant 3), and results are given of data generated under the CSM which were analysed with the LME (model variant 2) and the LM (model variant 4). For data generated under CSM, (model 2 and 4) true values of τ were also allowed to be negative. For data generated under the LME (model 1 and 3), the τ was restricted to be greater than or equal to zero. Note that when $\tau = 0$, the u_j were equal to zero representing no variance across clusters (see Equation 1). The simulated data for $\tau = 0$ under the LME did not contain any clustering effects. For data generated under the LME, the CR was only computed for $\tau \geq 0$.

The reported CRs of the intercept in the intercept-only condition showed bias. When the true value of τ was smaller than zero, the CR was greater than .95, which means that the Type-I error was deflated. For all cluster sizes, the CR was at least .97 for very small negative ICCs ($\tau = .01$), with a maximum of one for more negative ICCs. A similar bias in CR estimates were found for the LME and the LM (model variant 2 and 4 in Table 2). Thus, for negative correlations the LME performed similar to the LM. Under the LME, the random effect variance τ was estimated to be (approximately) zero, which made the LME similar to the LM as model for analysis for $\tau \leq 0$. Note that a CR of one is an upper bound, but the width of the CIs still increased for more negative values of τ.

When the correlation was greater than zero, and the LM was used to analyse the data, the CRs were highly underestimated, where the underestimation was larger for $n = 30$ than for $n = 10$. When increasing the correlation, the CRs were more underestimated. LM ignored correlation within groups, and the model assumed more information in the data than there actually was observed. This led to an inflation of the Type-I error, since a significant result was more easily obtained with estimated CIs that were too narrow. This bias was not expected when the LME was used as model for analysis. However, in that case also an underestimation of the CRs was observed. For increasing value of the correlation a decrease in the CR was observed. This underestimation increased when increasing the cluster size, and was observed for data generated under CSM as well as for data generated under the LME. For medium cluster sizes ($n = \{15, 30\}$), and relatively high correlations within a cluster (.10-.20), the CRs were underestimated, although the LME was used as the model for analysis. For all cluster sizes used in this simulation study, the coverage was at least .97 for small negative ICCs ($\tau \leq -.01$), which means that the Type-I error was deflated. This effect was the same for the CSM variants. The LME cannot handle negative cluster correlation. When the true τ was negative, τ was estimated to be around zero under the LME, leading to similar results of the LME and LM.

For all model variants and all conditions, the CR of the slope parameter was (approximately) equal to 95%, representing the finding that the 95% CI covered the true value in 95% of the data sets. The estimated CRs for the slope parameter did not show any bias. Even for negative correlations among clustered observations were the CRs around the level of 95%. It was concluded that the ignorance of the correlation within groups did not have an influence on the Type-I error of

Table 2 95% coverage rates for the intercept (β_0) and the slope parameter (β_1) for different cluster sizes ($n = 10, 15, 30$) and various correlations among clustered observations (τ)

	(1)		(2)		(3)		(4)	
GEN	LME		CSM		LME		CSM	
ANA	LME		LME		LM		LM	
τ	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
$L_b = -1/10, n = 10$								
−0.09			1.00	0.93			1.00	0.94
−0.07			1.00	0.94			1.00	0.94
−0.05			1.00	0.95			1.00	0.95
−0.03			0.98	0.95			0.98	0.95
−0.01			0.98	0.95			0.97	0.96
0.00	0.96	0.94	0.97	0.95	0.95	0.94	0.95	0.95
0.05	0.93	0.94	0.94	0.94	0.89	0.94	0.90	0.95
0.10	0.94	0.95	0.94	0.95	0.86	0.96	0.86	0.95
0.15	0.94	0.95	0.93	0.94	0.81	0.95	0.80	0.95
0.20	0.94	0.94	0.94	0.94	0.76	0.95	0.80	0.95
$L_b = -1/15, n = 15$								
−0.06			1.00	0.96			1.00	0.96
−0.04			1.00	0.95			1.00	0.96
−0.02			0.98	0.94			0.98	0.94
0.00	0.95	0.95	0.96	0.94	0.94	0.95	0.95	0.94
0.05	0.95	0.95	0.94	0.95	0.88	0.95	0.88	0.95
0.10	0.93	0.96	0.93	0.95	0.81	0.97	0.80	0.95
0.15	0.93	0.94	0.92	0.95	0.75	0.94	0.76	0.96
0.20	0.91	0.94	0.93	0.96	0.70	0.94	0.72	0.95
$L_b = -1/30, n = 30$								
−0.03			1.00	0.95			1.00	0.96
−0.01			97	0.96			0.97	0.97
0.00	0.97	0.95	0.96	0.94	0.96	0.94	0.95	0.94
0.05	0.94	0.96	0.94	0.95	0.82	0.95	0.79	0.94
0.10	0.93	0.94	0.93	0.93	0.69	0.93	0.68	0.93
0.15	0.92	0.95	0.94	0.94	0.62	0.94	0.62	0.92
0.20	0.90	0.96	0.93	0.95	0.53	0.94	0.59	0.94

Note. The τ is restricted to a lower bound $L_b = -\sigma^2/n$. For data generated under the LME $\tau \geq 0$. LME is the linear mixed effects model, CSM is the covariance structure model, LM is the linear regression model. *GEN*, the model that *generated* the data; *ANA*, the model that *analysed* the data

the slope parameter. Adjusting the level of correlation in the observations did not affect the regression effect or its precision of the predictor variable. As the slope was unaffected by the ignorance of the correlation, the remainder of this results sections only concerns data generated under the only-intercept condition.

3.2 Standard error and *p*-value

Generally, when testing if the intercept is equal to zero, $\beta_0 = 0$ and the true value is zero, then *p*-values are expected to be uniformly distributed and centred at .50. Furthermore, 5% of the *p*-values are expected to be smaller than or equal to .05. When the *p*-value is biased upwards, less than 5% of the *p*-values take a value of at most .05, which means that the Type-I error is deflated (i.e., a decrease in the probability to reject a correct null hypothesis). For *p*-values that are biased downwards, more than 5% of the *p*-values take a value of at most .05, which means that the Type-I error is inflated (i.e., an increase in the probability to reject a correct null hypothesis).

As can be seen in Table 3, for $\tau > 0$ and the LME or CSM was used to generate the data, the *p*-values were centred around .50 for different positive values of τ. This confirmed that LME can be used to control for the positive correlation in clustered data, and correct *p*-values were computed for the intercept. When the LM was used to analyse the data, for $\tau > 0$, the *p*-values were underestimated. With increasing τ that was ignored, the model assumed more and more information in the data than there actually was, which led to a higher percentage of significant *p*-values than the significance level of 5%. This led to an increase of the probability of rejecting a correct null hypothesis using a significance level of 5%.

For τ equal to zero, the *p*-value was always biased upwards for data analysed with the LME, since estimates of τ were upwardly biased when the τ was negative or close to 0, see Table 3. This was caused by the lower bound for τ under the LME. For negative τ, the *p*-values were always overestimated, when analysing the data with LME or LM. This bias was similar under both models, since the LME and LM cannot control for negative cluster correlation. It can also be seen that the level of overestimation increases quickly for a small increase in negative τ.

The bias of the *p*-value was larger for larger cluster sizes, which was caused by a smaller sampling error and more data information about the intercept. For a large number of observations per cluster, already a small negative correlation led to a large bias in *p*-value. For example, for $n = 30$ the correct *p*-value of .50 was increased by .06 for $\tau = -.01$; and for $\tau = -.03$ the *p*-value was overestimated by .27. For smaller cluster sizes the bias was smaller. However, for a smaller cluster size the lower bound for τ was also smaller, and for more negative correlations among clustered observations the bias in *p*-values again increased.

Similar findings were made for the SE of the estimated intercept. When data was generated under the LME or CSM, for τ greater than zero, results under the LME showed that the SE estimates were slowly increasing for increasing values of τ. The increasing correlation in the data led to an increase in the reduction of information about the intercept. When the LM was used as model for analysis, the SEs were underestimated (compared to those of the LME), since the correlation in the data was ignored. Both models cannot handle a negative correlation, which was ignored and the SE estimates showed a slight decrease for more negative correlations. The total variance in the observations was reduced for a negative τ (see Equation 3). This reduction in the total variance led to a reduction in the estimated SEs under LME and LM.

Table 3 Estimated SEs and p-values of the intercept (averaged across 1,000 replications) – which is set to 0 – for all four model variants, with varying number of observations per cluster (n)

	(1)	(2)	(3)	(4)
GEN	LME	CSM	LME	CSM
ANA	LME	LME	LM	LM
τ	SE (p)	SE (p)	SE (p)	SE (p)
$L_b = -1/10, n = 10$				
−0.09		0.095 (.80)		0.095 (.80)
−0.07		0.097 (.68)		0.097 (.68)
−0.05		0.098 (.61)		0.098 (.61)
−0.03		0.100 (.57)		0.098 (.56)
−0.01		0.105 (.53)		0.099 (.51)
0.00	0.107 (.52)	0.107 (.53)	0.100 (.50)	0.100 (.50)
0.05	0.123 (.50)	0.124 (.52)	0.102 (.43)	0.102 (.45)
0.10	0.138 (.50)	0.139 (.50)	0.104 (.41)	0.104 (.40)
0.15	0.157 (.51)	0.156 (.50)	0.107 (.38)	0.106 (.37)
0.20	0.170 (.49)	0.168 (.51)	0.108 (.35)	0.108 (.36)
$L_b = -1/15, n = 15$				
−0.06		0.079 (.76)		0.079 (.76)
−0.04		0.080 (.62)		0.080 (.62)
−0.02		0.083 (.57)		0.081 (.56)
0.00	0.087 (.52)	0.087 (.52)	0.081 (.49)	0.081 (.50)
0.05	0.108 (.52)	0.106 (.51)	0.084 (.43)	0.083 (.42)
0.10	0.127 (.50)	0.126 (.49)	0.085 (.37)	0.085 (.36)
0.15	0.144 (.49)	0.143 (.50)	0.087 (.33)	0.087 (.34)
0.20	0.157 (.49)	0.160 (.50)	0.088 (.32)	0.089 (.31)
$L_b = -1/30, n = 30$				
−0.03		0.057 (.77)		0.057 (.77)
−0.01		0.059 (.56)		0.057 (.55)
0.00	0.062 (.53)	0.062 (.51)	0.058 (.50)	0.058 (.49)
0.05	0.088 (.51)	0.089 (.50)	0.059 (.38)	0.059 (.37)
0.10	0.111 (.50)	0.112 (.48)	0.060 (.32)	0.060 (.29)
0.15	0.132 (.49)	0.131 (.49)	0.061 (.26)	0.061 (.25)
0.20	0.147 (.47)	0.147 (.50)	0.062 (.23)	0.063 (.25)

Note. The τ is restricted to a lower bound of $L_b = -\sigma^2/n$. For data generated under the LME $\tau \geq 0$. LME is the linear mixed effects model, CSM is the covariance structure model, LM is the linear regression model

In Fig. 1, it can be seen that for $\tau \geq .05$, correct estimates of the p-values (around .50) were obtained under the LME. For $0 \leq \tau < .05$, the correlation was (slightly) overestimated leading to an overestimation of the p-values under the LME. When accounting incorrectly for intra-cluster correlation the data contained more information about the intercept than captured by the LME. Specifically, when the correlation was zero, the LME slightly overestimated the p-value due to assuming that there was a small amount of correlation, while there was no correlation in the data. This was caused by the fact that the level of zero correlation is a lower bound under the LME. When the LM is used as model of analysis, the deflation and inflation of p-values is shown for increasing values of the correlation in each cluster. For $\tau > 0$, the underestimation of the p-values occurred when ignoring the correlation, which is shown under the LM. When the level of correlation is negative, there is a steep increase in the overestimation of the p-values for increasing negative values of the correlation under both the LME and LM. It can be concluded the LME performs as poorly as the LM, when it concerns a negative cluster correlation.

In Figure 2, the average SE estimates are shown for increasing values of the correlation, where the LM and the LME is used as model of analysis. For $\tau > .05$ the SE estimates under the LME were considered to be the baseline, as LME accounts for the positive correlation in the clusters. It can be seen that with the LM as model of analysis, the SEs were underestimated, which became more severe for increasing values of the correlation. This can be seen from the increase in difference in SE estimates under the LM and the LME. For $\tau = 0$, the LM was the correct model of analysis, and in that case the SE estimate under the LME was slightly higher showing that the LME incorrectly assumed more correlation in the data leading to an overestimation of the SE. For $\tau < 0$, the SE estimates under the LME were more similar to those under the LM. However, even for small negative correlations, the SE estimates under the LME were higher than those obtained under the LM. The LME incorrectly assumed a positive correlation among clustered observations. Furthermore, for negative correlations the total variance reduced leading to a slight decrease in estimated SEs under both models.

3.3 Estimation of the correlation τ

The values of τ were estimated under the LME, and the data were generated under LME and under CSM. The τ values for data generated under LME were restricted to be greater than or equal to zero, where under CSM the τ values were also negative. The reported estimates of τ in Table 4 were averaged values across the 1,000 data sets. It can be seen that the estimated τ values under LME were biased, when the true τ was less than .05 or negative. In specific, the τ estimates were biased upwards, when the true value of τ was equal to zero. As can be seen in Table 4, the bias was smaller for larger values of the cluster size, because with an increasing sample size there was less sampling error. This led to less variability in the data, which means that more information about the fixed effect was included in the data. For a true positive correlation of $\tau = .05$ still some bias was found in the estimates for small cluster sizes. When the true positive τ values were greater than or equal to .10, the estimates of τ were not biased. This effect was equal for all cluster sizes. However, for small negative τ, small positive correlations were estimated under the LME, when the cluster size was small, and a zero correlation was estimated, when the cluster size increased to 30 observations.

Table 4 Estimates of τ, averaged across 1,000 replication, for different observations per cluster and $m = 10$ clusters under the LME

	(1)	(2)
GEN	LME	CSM
ANA	LME	LME
τ	$\hat{\tau}$	$\hat{\tau}$
$L_b = -1/10, n = 10$		
−0.09		0.000
−0.07		0.000
−0.05		0.001
−0.03		0.005
−0.01		0.014
0.00	0.019	0.018
0.05	0.059	0.059
0.10	0.099	0.103
0.15	0.159	0.157
0.20	0.204	0.197
$L_b = -1/15, n = 15$		
−0.06		0.000
−0.04		0.000
−0.02		0.004
0.00	0.012	0.012
0.05	0.055	0.051
0.10	0.104	0.100
0.15	0.151	0.150
0.20	0.194	0.201
$L_b = -1/30, n = 30$		
−0.03		0.000
−0.01		0.002
0.00	0.006	0.007
0.05	0.049	0.050
0.10	0.097	0.099
0.15	0.150	0.150
0.20	0.196	0.195

Note. The τ is restricted to a lower bound of $L_b = -\sigma^2/n$. LME is the Linear Mixed Effects model, and CSM is the Covariance Structure model

4 Reformulating LMEs as SEMS

Rovine and Molenaar (2000) and Bauer (2003) showed that a general class of LMEs can be reformulated as structural equation models (SEMs). Therefore, estimation methods designed for SEMs can be used to fit LMEs. The general idea is straightforward, the LME is reformulated as a SEM by considering the marginal model for the data. The CSM in Equation (5), which is the reformulated LME of Equation (1),

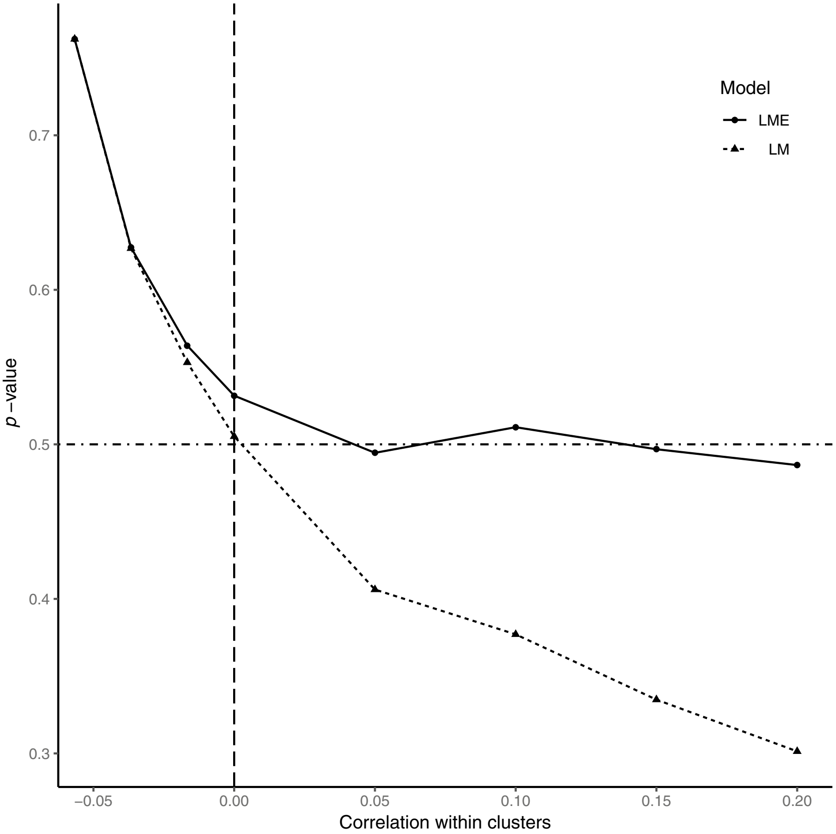

Fig. 1 Estimated p-values for the intercept (averaged across 1,000 replicated data sets) for different levels of correlation among $n = 10$ clustered observations and $m = 10$ clusters for the LME and the LM

represents the marginal distribution of the data with covariance matrix $\sigma^2\mathbf{I}_n + \tau\mathbf{1}_n\mathbf{1}_n^t$. To define the corresponding SEM, the factor loading matrix is represented by the design matrix of the random intercept, which is $\Lambda = \mathbf{1}_n$, the mean vector of the latent variable equals $\boldsymbol{\alpha} = (\beta_0, \beta_1)$, and the residual covariance matrix is given by $\boldsymbol{\Theta} = \sigma^2\mathbf{I}_n$. Although τ represents the variance of an endogenous latent variable in the SEM model, it enters the model-implied covariance matrix as a covariance parameter.

The maximum likelihood fitting method for SEM maximizes the likelihood of the observed covariance matrix, and minimizes the discrepancy between the implied covariance matrix and the observed covariance matrix. This estimation method supports the estimation of a negative τ, since it uses the implied covariance matrix for which τ is not restricted to be a variance parameter. The maximum likelihood parameters are obtained by minimizing the maximum likelihood fitting function (Ferron and Hess 2007). A numerical method (e.g., Newton-Raphson) is used to find the parameter values that minimizes the function. There are various software

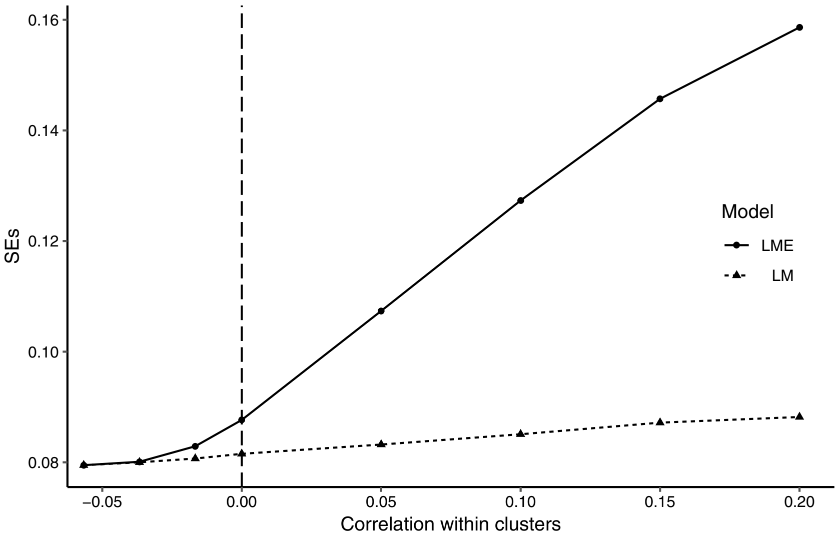

Fig. 2 Estimated standard errors (SEs) for the intercept (averaged across 1,000 replicated data sets) for different levels of correlation among $n = 10$ clustered observations and $m = 10$ clusters under the LME and the LM

packages that uses the maximum likelihood fitting function for estimation (e.g., Lavaan, Mplus).

Lavaan has been used to fit the SEM version of the random intercept model. However, for the considered conditions in the simulation study, the number of observations (number of clusters) was less than the number of variables (number of dependent plus independent variables). This led to estimation problems in Lavaan. When increasing the number of clustered observations, the number of variables rapidly increase the number of observed clusters. The sample size restriction is a serious limitation of the SEM estimation method. Wolf et al. (2013) showed that for a one-factor confirmatory factor analysis at least 30 to 190 independent cases are needed. This makes the SEM approach not useful for applications with a few clusters with many clustered observations. This will lead to a large covariance matrix, which needs to be estimated with only a few independent data cases.

4.1 Real data example

The data from Lamers et al. (2015) were used to study negative clustering effects. The dataset included 174 clients who were recruited through advertisements in Dutch newspapers and web-sites. Only participants who felt depressed and expressed interest in writing about their life were included. Clients with no missing data were used leading to a total of 90 clients, denoted as $i = 1, \ldots, n$, who were at random equally divided over $m = 5$ counselors, denoted as $j = 1, \ldots, m$ (leading to a balanced study design). The study had two treatment-arms: auto-biographic writing condition (AW), and expressive writing condition (EW). Clients were also randomly assigned to one of the treatments.

The AW condition was a life-review self-help intervention that consisted of homework assignments, divided over modules that had to be completed over the course of ten weeks. Clients communicated about their progress with trained counselors through a weekly e-mail interaction. The EW intervention was based on the method of expressive writing by Pennebaker (1997). The method consisted of daily writing about emotional experiences, for $15-30$ minutes on $3-4$ consecutive days during one week. The available data included the pre- and post-therapeutic measurements, denoted as $t=1,2$, of the Center for Epidemiologic Studies Depression Scale (CES-D) score. The CES-D is a brief self-report questionnaire to measure severity of depressive symptoms in the general population. Higher CES-D scores indicated more depressive symptoms (20 items, range $0-60$, $\alpha=0.78$). The random assignment of clients to treatment groups ensured that there was not a significant difference between the average pre-therapeutic measurements of both groups.

Interest was focused on examining the influence of the counselor on the health improvement of the clients. Therefore, a two-factor (LME) model was examined, with a (nested) factor client (clustering the pre- and post measurements), and clients were again clustered by counselors. The treatment indicator was included to examine differences in scores among the treatment groups. The LME represents this two-factor model, and is given by,

$$\begin{aligned}
Y_{ijt} &= \beta_{01} + \beta_{02} I(Post) + \beta_{10}\text{Treatment}_{ij} + \beta_{11}\text{Treatment}_{ij} I(Post) \\
&\qquad + \text{client}_{ij} + \text{counselor}_j + e_{ijt} \\
e_{ijt} &\sim N(0,\sigma^2) \\
\text{client}_{ij} &\sim N(0,\delta) \\
\text{counselor}_j &\sim N(0,\tau).
\end{aligned} \tag{8}$$

The treatment variable (Treatment_{ij}) equaled one when client i of counselor j was assigned to the EW intervention and zero when assigned to the AW intervention. The variable $I(Post)$ is the indicator for the post measurements and $\text{Treatment}_{ij} I(Post)$ the treatment indicator for the post measurements. Parameter β_{01} is the intercept for the pre measurements, and β_{02} the intercept for the post measurements. A pre-existing difference between intervention groups is represented by the parameter β_{10}, which represents a deviation from the intercept β_{01} for clients in the EW intervention. A difference at the post measurement is represented by parameter β_{11}, which represents a deviation from the intercept β_{02} for clients in the EW intervention. The fit of this LME model led to estimation problems. The estimated covariance matrix was singular, and the variance component of the latent variable counselor was estimated to be zero (or less than zero, but `LME4` does not report information about negative variance component estimates; Smink et al. 2019). It was concluded that the model with counselor as a latent variable was singular, and the model parameters could not be estimated.

The model can be rewritten as a multilevel SEM to allow for negative within-cluster correlations, by assuming that the pre- and post-measurements of each client are multivariate normally distributed. The multilevel structure is defined by clients (factor variable) who are nested within counselors (factor variable). Then, the ML-SEM, referred to as M1, is represented by

$$\begin{aligned}
&\mathbf{Y}_{ij} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \text{Treatment}_{ij} + \boldsymbol{\Lambda}\boldsymbol{\eta}_{ij} + \mathbf{e}_{ij} \\
&\boldsymbol{\eta}_{ij} = \text{client}_{ij} + \text{counselor}_j \\
&\text{client}_{ij} \sim N(0, \delta) \\
&\text{counselor}_j \sim N(0, \tau). \\
&\mathbf{e}_{ij} \sim N(0, \sigma^2 \mathbf{I}_2),
\end{aligned} \tag{9}$$

where the vector of loadings, $\boldsymbol{\Lambda}$, contains only ones. The $\boldsymbol{\beta}_0$ represents the intercepts for the pre- and post measurement, and the $\boldsymbol{\beta}_1$ the (within-counselor) treatment effects for the pre- and post measurements, as described in Equation (8).

Model M1 was fitted in Lavaan (version 0.6-5), and it reported convergence of the estimation algorithm with a warning that some latent variable variance estimates were negative. The fit indices did not indicate a misfit of the model; CFI=1 and RMSEA=0. In Table 5, the parameter estimates and standard errors are reported. It follows that on average the CES-D decreased from around 21.7 to 17.6 for those in the AW condition. Clients in the EW condition scored approximately 1.5 points lower on the post measurement and 0.1 lower on the pre measurement than those in the AW condition. However, the (within-counselor) treatment effect on the post measurements was not significantly different from zero. The pre-existing difference in scores between conditions was expected to be around zero due to the random assignment of clients to treatments.

The covariance among scores of clients of the same counselor was estimated to be around -1.156, which differed significantly from zero. The negative correlation made it impossible to investigate a homogenous treatment effect of counselors. The negative correlation among observations clustered by the counselor showed that scores differed substantially among clients who were treated by the same counselor. We argue that the counselor differentiated clients in their treatment, where some clients benefitted much more from the treatment than other clients.

It was investigated if the negative correlation of clustered scores by counselors differed across treatment levels. Therefore, a multivariate distribution was assumed for the client scores in the EW and AW condition, where for each condition an intercept (pre- and post measurement), residual variance, client and counselor factor variance was defined. In this model M2, in contrast to model M1, intercepts were defined for the pre and post scores for the AW and EW condition, instead of defining deviations from the general intercepts for the EW condition.

The parameter estimates of the multilevel SEM are given in Table 5 under the label model M2. It can be seen that on average the clients in the EW condition scored lower on the pre- and post measurement. When examining the intra-cluster correlation (ICC) of the factor variables, it follows that for the AW condition the ICC is around -8.1% (-1.088/(-1.088+14.463)), and for the EW condition around -9.5% (-.852/(-.852+9.865)). However, the differentiation in EW and AW counselor-specific cluster effects led to non-significant cluster (co)variance estimates. Note that

Table 5 Real data: estimated ML-SEM parameters

Model	Parameter	Estimate	Std. Error
M1	*Fixed*		
	Intercept (Pre (AW))	21.690	0.771
	Intercept (Post (AW))	17.645	0.721
	Treatment (Pre (EW))	−0.113	1.281
	Treatment (Post (EW))	−1.468	1.281
	Random		
	Residual (W)	21.382	3.187
	Client (W)	16.002	4.396
	Counselor (B)	−1.156	0.309
M2	*Fixed*		
	Intercept (Pre (AW)	21.800	0.832
	Intercept (Post (AW))	17.625	0.832
	Intercept (Pre (EW)	20.500	0.715
	Intercept (Post (EW))	15.675	0.715
	Random		
	Residual (AW) (W)	21.922	4.902
	Residual (EW) (W)	17.422	3.896
	Client (AW) (W)	14.463	6.553
	Client (EW) (W)	9.865	4.894
	Counselor (AW) (B)	−1.088	1.525
	Counselor (EW) (B)	−0.852	1.083

the negative ICC shows that the counselor explained variability in client scores, but it led to more dissimilarity between clients. The treatment effect of a counselor varied across clients, where some clients benefitted from the treatment and others did not. A positive ICC would indicate a homogenous treatment effect, where all clients benefit from the treatment, but some more than others. A more negative ICC leads to a greater distinction between clients treated by the same counselor who benefit and who do not benefit from the treatment. Clients in the EW condition were more likely to differentiate in the effect of their treatment than those in the AW condition but the differentiation in clustering between EW and AW clients was not significant.

Data was retrieved under a pretest-posttest randomized experiment with two treatment arms. The randomization procedure (i.e. clients were randomly assigned to counselors and treatment groups) ensured that it was unlikely that another factor(s), besides the counselor, was responsible for the negative ICCs. Therefore, it was concluded that counselors induced the negative correlation among client scores. It was argued that a differentiation in the treatment of the clients by the counselor could have caused this. An individualized treatment by counselors could have worked for some clients and not for other clients, which led to negative clustering effects. The differentiation in treatment effects between clients (most likely) led to a non-significant main difference between the AW and EW condition.

The clustered scores by counselors correlate negatively, which leads to a negative variance estimate of the factor variable counselor. A negative variance estimate is often referred to as a “Heywood case”. There are many causes given in the literature for getting negative variance estimates: outliers, nonconvergence, underidentification, structurally misspecified models, missing data, and sampling error. In this study, it can be argued that a model misspecification was the cause for the negative variance estimate. For instance, mediator variable(s) could be missing, representing an indirect effect of the counselor on the client scores. The relationship between counselor and clients could depend on a moderator variable, which was not included in the model. Furthermore, a possible mediation effect could be moderated by another variable. However, mediator and/or moderator variables were not measured during the experiment. Given the randomization procedure, there were no theoretical reasons that it was necessary to control for other (confounding) variables at the client level.

5 Discussion

The purpose of this study was to confirm that estimates of the ICC were biased upwards for true small positive and negative values. The results confirm that the ignorance of (small) positive correlation within groups leads to an increase of Type-I error, *p*-values that are biased downwards and CIs that are too narrow. The results for ignoring true positive correlation within clusters validate earlier research. The inflation of Type-I errors of this study correspond to those reported by Barcikowski (1981). Regarding the different conditions, when ignoring a positive correlation within groups, the results of this study also confirm earlier research. As reported by Barcikowski (1981) and Dorman (2008), the smallest bias in Type-I error was found for a smaller number of observations per group. When the number of observations per group increased, the inflation of the Type-I error also increased. Furthermore, when increasing the positive correlation, the Type-I error was increasing as well. Furthermore, the results show that, in general, only the intercept parameter and not the slope parameter was affected by ignoring a common correlation within clusters. Effects of slope variables are not affected, when ignoring a common correlation the distortion only concerns the intercept parameters.

Regarding the estimates of the correlation, the results show the negative effects of restricting the correlation among clustered observations to be positive. Already for small positive correlation ($\tau < .05$) the correlation was overestimated. The bias was larger when the correlation was very small positive or negative. However, for a positive correlation within groups larger than .05, the LME model gave correct estimates, which validates the accuracy of the LME for clustered data with positively correlated observations.

Next to ignoring a positive correlation, the ignorance of a (small) negative correlation within groups was considered. The results showed opposite effects compared to ignoring positive correlation within groups: a deflation of Type-I errors, *p*-values that are biased upwards and overestimated SEs (i.e. CIs that are too wide). The deflation of the Type-I error, when ignoring a negative correlation has been mentioned by other researchers (Barcikowski 1981; Rosner and Grove 1999), but without quantifying the negative effects of ignoring the common negative correlation. Current findings indicate that when clustered observations are negatively correlated, the data is more informative than it would be under independent sampling. This study adds that smaller biases of the Type-I error and the SEs occur, when increasing the number of observations in a cluster. The increase in sample size leads to smaller sampling error and more accurate estimates. In addition, the bias of the Type-I error is examined, where a deflation of the Type-I error by 2% can already occur for a true negative ICC of $-.01$. Furthermore, the results indicate that *p*-values were biased upwards when negative correlation within groups was ignored. For a larger number of observations per cluster, even the ignorance of a very small negative correlation within groups can lead to substantial bias of the *p*-value.

With respect to the results of the current study, it is recommended to be aware of the fact that a negative correlation within groups can occur. The advice is to be very cautious with clustered data that possibly contains a common negative correlation among observations within groups. Researchers need to be aware of the fact that the maximum likelihood estimate of the between-cluster variance, representing the covariance among clustered observations (see Equation 2) can become negative. Common software packages for multilevel and mixed effects models, such as the `lme4`-package, restrict the random-effect variance estimate to be positive (although SAS PROC MIXED allows for negative variance estimates). This means that the correlation among clustered observations is restricted to be positive even when the true correlation is negative. This study showed that this can lead to a deflation of Type-I errors, *p*-values that are biased upwards and SEs that are overestimated. In addition, a large number of observations per cluster does not compensate for bias. It is important to notice that with 30 observations per cluster a very small true negative correlation of $-.008$ can lead to an inflation of the *p*-value with 5%. It is also noted that for a smaller number of observations per cluster, the correlation can become more negative than with a larger number of observations. Thus, the bias of the *p*-value is smaller when the true negative correlation between observations is small. However, in that case the correlation can become more negative which can lead to more bias.

The reformulation of an LME to SEM for continuous outcomes makes it possible to use SEM software (e.g., Lavaan, Mplus). The maximum likelihood fitting method can be used to estimate the parameters of the implied covariance matrix in which the factor variance is represented as a covariance parameter. This makes it possible to examine negative within-cluster correlations as shown in the real data example. The situation is different for categorical outcomes, where numerical integration is performed to integrate out factor variables. In general, nonlinear and generalized LME models cannot be parameterized as SEMs (Bauer 2003). Another issue is that the sample size requirements are higher for SEMs than LMEs. For instance, the sample size conditions considered in our simulation study led to estimation issues for corresponding SEMs. Baird and Maxwell (2016) showed that unconstrained variance estimation for random effects can improve the convergence of the estimation method, and unconstrained estimation can also provide more insight about sources of misfit. However, negative variance estimates can indicate negative within-cluster correlation, which needs to be taken into account.

When the data has negative correlations SEM software can be used to analyse the clustered data, but more research is needed to improve the modeling of negative within-cluster correlations. For instance, bias in the estimated precision of the slope parameter can occur when ignoring correlation among clustered observations. Consider a random intercept-slope model,

$$\begin{aligned} \mathbf{y}_j =& \beta_0 + \beta_1 \mathbf{X}_j + u_{0j} + u_{1j}\mathbf{X}_j + \mathbf{e}_j, \\ u_{0j} \sim& \mathcal{N}(0, \tau_0), \\ u_{1j} \sim& \mathcal{N}(0, \tau_1), \\ \mathbf{e}_j \sim& \mathcal{N}(0, \sigma^2 \mathbf{I}_n), \end{aligned} \tag{10}$$

where both random effects are normally distributed. Subsequently, the implied covariance matrix for cluster j is represented by $\boldsymbol{\omega}_j = \sigma^2\mathbf{I}_n + \tau_0\mathbf{J}_n + \tau_1\mathbf{X}_j\mathbf{X}_j^t$. The estimated precision of the intercept contained bias, when ignoring the common correlation τ_0 in cluster j. In the same way, the estimated precision of the slope parameter will contain bias, when ignoring the common correlation τ_1 modified by the outer product of $\mathbf{X}_j$. However, more research is needed to quantify bias in the precision of the common regression effect, when ignoring cluster correlation implied by a cluster-specific regression effect.

Statistical testing (negative) within-cluster correlations is very important to understand the dependence structure of the clustered data. This is usually complex, since this means testing the variance parameter of a random effect at the lower-bound of the parameter space. This task is easier, when testing the corresponding covariance parameter of the implied covariance matrix. Then, hypotheses concerning a positive or a negative dependence structure can be examined. Mulder and Fox (2019) developed a Bayes factor to evaluate restrictions on intra-class correlation coefficients defined under a random intercept model. In their proposed method, the intra-class correlation is defined from an implied covariance matrix and is allowed to be negative as long as the covariance matrix is positive definite. They also show that a hypothesis concerning a negative intra-class correlation can be evaluated. This test method has the potential to be applicable to test the dependence structure of an implied covariance structure of an LME, even when the within-cluster correlation is negative.

5.1 What can be learned from our study?

We conclude with a discussion of what we found to be the most important lessons (that can be learned based on our study), and we give an updated overview of well-known multilevel modelling (golden) rules by including the findings concerning negative clustering effects. This reflection highlights the importance of understanding negative clustering effects.

5.1.1 Lesson 1: Negative correlations among clustered observations can –and do– occur in practice

Although this has been addressed in the introduction, it is important to stress that –although simulated data were used in the current study– negative correlations can and do occur in practice. Several practical examples were discussed, when the correlation τ is negative. Pryseley et al. (2011) and Kenny et al. (2002) mention several examples: when individuals compare for a scarce (and fixed) set of resources, the speaking time of one individual is at the expense of another individual (i.e. '*one's pain is the other's gain*'). Litter mates are also negatively correlated in terms of food, water and living space. In addition to that, it has been shown that *Bayesian covariance structure modelling* (BCSM; Klotzke and Fox 2019a, b) can be used to simulate these negative correlations in clusters. In doing so, a simulation study was developed where observations are negatively clustered and data were simulated that others, most notably (Kenny et al. 2002; Oliveira et al. 2017), and Pryseley et al. (2011) described.

5.1.2 Lesson 2: The ICC has multiple interpretations

It is important to be aware of the multiple interpretations of the ICC, because viewing it as the proportion of explained variance due to clustering restricts the ICC to only positive values (Oliveira et al. 2017). It is good to stress that the ICC remains a measure of correlation, and correlations can become negative. A negative τ leads to a negative ICC, thus both values can become negative. Researchers need to be aware that negative correlation between observations in clustered data may occur, and that negatively correlated clustered data are not (always) a modelling error.

5.1.3 Lesson 3: The LME ignores negative associations in clusters

The LME (e.g., multilevel model, random effects model) is an inappropriate tool for analysing clustered data with negative correlations between observations within clusters. The LME is a well-established tool to model the dependence structure of the clustered data using random effects. However, this is only possible for positively correlated observations, but –as shown– the random effects cannot model negative correlations, and consequently will ignore the negative dependence between observations. When the LME is used for analysing data with negative correlations within clusters, the model assumes the observations to be independently distributed. Therefore, the LME is an inappropriate tool for analysing clustered data with negative correlations within clusters.

5.1.4 Lesson 4: Do not fix the ICC or τ to 0

Fixing a negative ICC to zero forces the statistical model to ignore the clustered structure in the data. Then, the observations are assumed to be independently distributed while they are negatively correlated. It is shown that if a common negative correlation in clusters is ignored, the Type-I errors are deflated, the SEs are too large, and the p-values are also overestimated. Fixing the ICC to 0 forces the statistical model to assume that data are independently distributed. Researchers should always account for a non-zero ICC, positive or negative. Huang (2018) stated that "literally too many to list" suggested that it is not necessary to account for relatively small (positive) ICC values. Given the fact that ignoring a negative correlation can also have a large impact, it is stressed that small positive or negative ICCs cannot be ignored. For instance, when having ten observations in each group, ignoring an ICC of only $-.029$ ($\tau = -.03$) leads to an deflation of the Type-I error of around 6%. The occurrence of a Type-I errors quickly increases: an ICC of $-.047$ ($\tau = -.05$) deflates the Type-I error with 11%.

5.1.5 Lesson 5: An increase in cluster size increases the bias

When the number of observations in each cluster n tends to get larger, the bias of the Type-I error and the SEs increases. The results indicate that p-values were biased upwards (downwards) when negative (positive) correlation within groups was ignored, and the bias increased with an increasing cluster size. When assuming more clustered observations to be independently distributed, while they are correlated, the bias will increase, since an incorrect assumption is made about a larger data set.

5.1.6 Lesson 6: Negatively correlated clustered observations cannot be ignored

It was shown that ignoring a small negative ICC leads to serious bias. It can also be argued that a cluster sample with negatively correlated observations contains more information than a simple random sample of the same size. The variance inflation factor (VIF; i.e. design effect; Kish 1965) can be used to explain the impact of the ICC (ρ). In the random selection of equal clusters, the VIF can be expressed as a function of ρ;

$$\mathrm{VIF} = 1 + (n - 1)\rho, \tag{11}$$

where n is the number of observations in a cluster. The VIF represents the ratio of the actual sample variance to the variance of a simple random sample (i.e. independently distributed observations). In cluster sampling, when the ICC is positive, the VIF is greater than one. There is a loss in precision (i.e., increase in sample variance) due to the homogeneity of observations within each cluster. A simple random sample of the same size contains more information than a cluster sample with positively correlated observations. However, when the observations are negatively correlated, leading to a negative ICC, the VIF will be smaller than one, but greater than zero given that the ICC is greater than $-1/(n-1)$. In that case, the cluster sample leads to a gain in precision in comparison to simple random sampling. The cluster sample, with negatively correlated data provide more information than a simple random sample of equal size.

5.2 Conclusion

Just as small (positive) clustering effects mandate adjustment, negative clustering effects also require the attention of researchers. It is shown that ignoring negative clustering leads to deflated Type-I errors, an overestimation of the SEs, and overestimated *p*-values. The LME is an inappropriate tool for the analysis of negative clustering in data and can only handle positive clustering. Although it seems obvious, the LME should not be used when there is no clustering.

Acknowledgements The author(s) received no financial support for the research, authorship, and/or publication ofthis article.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Baird R, Maxwell SE (2016) Performance of time-varying predictors in multilevel models under an assumption of fixed or random effects. Psychol Methods 21(2):175–188. https://doi.org/10.1037/met0000070

Baldwin SA, Stice E, Rohde P (2008) Statistical analysis of group-administered intervention data: reanalysis of two randomized trials. Psychother Res 18(4):365–376. https://doi.org/10.1080/10503300701796992

Barcikowski RS (1981) Statistical power with group mean as the unit of analysis. J Educ Stat 6(3):267–285. https://doi.org/10.2307/1164877

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using flme4g. J Stat Softw 67(1):1–48. https://doi.org/10.18637/jss.v067.i01

Bauer DJ (2003) Estimating multilevel linear models as structural equation models. J Educ Behav Stat 28(2):135–167. https://doi.org/10.3102/10769986028002135

Bock RD, Bargmann RE (1966) Analysis of covariance structures. Psychometrika 31(4):507–534. https://doi.org/10.1007/bf02289521

Clarke P (2008) When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. J Epidemiol Community Health 62(8):752–758. https://doi.org/10.1136/jech.2007.060798

Dorman JP (2008) The effect of clustering on statistical tests: an illustration using classroom environment data. Educ Psychol 28(5):583–595. https://doi.org/10.1080/01443410801954201

El Leithy HA, AbdelWahed ZA, Abdallah MS (2016) On non-negative estimation of variance components in mixed linear models. J Adv Res 7(1):59–68. https://doi.org/10.1016/J.JARE.2015.02.001

Eldridge SM, Ukoumunne OC, Carlin JB (2009) The intra-cluster correlation coefficient in cluster randomized trials: a review of definitions. Int Stat Rev 77(3):378–394. https://doi.org/10.1111/j.1751-5823.2009.00092.x

Ferron JM, Hess MR (2007) Estimation in SEM: a concrete example. J Educ Behav Stat 32(1):110–120. https://doi.org/10.1037/met00000701

Fox J-P, Mulder J, Sinharay S (2017) Bayes factor covariance testing in item response models. Psychometrika 82(4):979–1006. https://doi.org/10.1007/s11336-017-9577-6

Galbraith S, Daniel JA, Vissel B (2010) A study of clustered data and approaches to its analysis. Journal of Neuroscience 30(32):10601–10608. https://doi.org/10.1037/met00000703

Giberson TR, Resick CJ, Dickson MW (2005) Embedding leader characteristics: an examination of homogeneity of personality and values in organizations. J Appl Psychol 90(5):1002. https://doi.org/10.1037/met00000704

Gibson J, Malandrakis N, Romero F, Atkins DC, Narayanan S (2015) Predicting Therapist Empathy in Motivational Interviews using Language Features Inspired by Psycholinguistic Norms. In: Sixteenth annual conference of the international speech communication association. Dresden

Hox JJ, Maas CJ, Brinkhuis MJ (2010) The effect of estimation method and sample size in multilevel structural equation modeling. Statistica Neerlandica 64(2):157–170. https://doi.org/10.1037/met00000705

Huang FL (2018) Multilevel modeling myths. School Psychol Q 33(3):492–499. https://doi.org/10.1037/met00000706

Jöreskog KG (1969) A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 34(2):183–202. https://doi.org/10.1037/met00000707

Jöreskog KG (1971) Simultaneous factor analysis in several populations. Psychometrika 36(4):409–426. https://doi.org/10.1037/met00000708

Kenny DA, Judd CM (1986) Consequences of violating the independence assumption in analysis of variance. Psychol Bull 99(3):422–431. https://doi.org/10.1037/0033-2909.99.3.4229

Kenny DA, Mannetti L, Pierro A, Livi S, Kashy DA (2002) The statistical analysis of data from small groups. J Personal Soc Psychol 83(1):126–137. https://doi.org/10.1080/1050330070179699 20

Kish L (1965) Survey sampling. John Wiley and Sons Inc, New York

Klotzke K, Fox J-P (2019) Bayesian covariance structure modelling of responses and process data. Front Psychol 10:1675. https://doi.org/10.3389/fpsyg.2019.01675

Klotzke K, Fox J-P (2019) Modeling dependence structures for response times in a Bayesian framework. Psychometrika 84(3):649–672. https://doi.org/10.1007/s11336-019-09671-8

Krannitz MA, Grandey AA, Liu S, Almeida DA (2015) Workplace surface acting and marital partner discontent: anxiety and exhaustion spillover mechanisms. J Occup Health Psychol 20(3):314. https://doi.org/10.1037/a00387633

Kuznetsova A, Brockhoff PB, Christensen RHB (2017) flmerTestg Package: Tests in Linear Mixed Effects Models. https://doi.org/10.18637/jss.v082.i13

Lamers SMA, Bohlmeijer ET, Korte J, Westerhof GJ (2015) The efficacy of life-review as online-guided self-help for adults: a randomized trial. J Gerontol Ser B Psychol Sci Soc Sci 70(1):24–34. https://doi.org/10.1080/105033007017969924

Langfred CW (2007) The downside of self-management: a longitudinal study of the effects of conflict on trust, autonomy, and task interdependence in self-managing teams. Acad Manag J 50(4):885–900. https://doi.org/10.1080/105033007017969925

Loeys T, Molenberghs G (2013) Modeling actor and partner effects in dyadic data when outcomes are categorical. Psychol Methods 18(2):220–236. https://doi.org/10.1080/105033007017969926

Maas CJ, Hox JJ (2005) Sufficient sample sizes for multilevel modeling. Methodology 1(3):86–92. https://doi.org/10.1080/105033007017969927

McCulloch CE, Searle SR, Neuhaus JM (2008) Generalized, Linear, and Mixed Models Generalized, Linear, and Mixed Models, 2nd edn. John Wiley & Sons Ltd, Hoboken. https://doi.org/10.1198/tech.2003.s13

Molenberghs G, Verbeke G (2007) Likelihood ratio, score, and Wald tests in a constrained parameter space. Am Stat 61(1):22–27. https://doi.org/10.1198/016214505000000024

Molenberghs G, Verbeke G (2011) A note on a hierarchical interpretation for negative variance components. Stat Model 11(5):389–408. https://doi.org/10.2307/11648770

Mulder J, Fox J-P (2019) Bayes factor testing of multiple intraclass correlations. Bayesian Anal 14(2):521–552. https://doi.org/10.2307/11648771

Norton EC, Bieler GS, Ennett ST, Zarkin GA (1996) Analysis of prevention program effectiveness with clustered data using generalized estimating equations. J Consult Clin Psychol 64(5):919. https://doi.org/10.2307/11648772

Oliveira IRC, Demétrio CGB, Dias CTS, Molenberghs G, Verbeke G (2017) Negative variance components for non-negative hierarchical data with correlation, over-, and/or underdispersion. J Appl Stat 44(6):1047–1063. https://doi.org/10.2307/11648773

Pennebaker JW (1997) Writing about emotional experiences as a therapeutic process. Psychol Sci 8:162–166. https://doi.org/10.2307/11648774

Pryseley A, Tchonlafi C, Verbeke G, Molenberghs G (2011) Estimating negative variance components from Gaussian and non-Gaussian data: a mixed models approach. Comput Stat Data Anal 55(2):1071–1085. https://doi.org/10.2307/11648775

R Core Team (2020) R: A Language and Environment for Statistical Computing. Vienna, Austria. https://doi.org/10.1007/978-3-540-74686-7

Rao CR (1973) Linear Statistical Inference and its Applications, 2nd edn. John Wiley & Sons Inc, New York. https://doi.org/10.2307/11648776

Raudenbush SW, Bryk AS (2002) Hierarchical linear models: applications and data analysis methods, 2nd edn. Sage Publications, Los Angeles

Rosner B, Grove D (1999) Use of the Mann-Whitney U-test for clustered data. Statistics in medicine 18(11):1387–1400. https://doi.org/10.2307/11648777

Rovine MJ, Molenaar PCM (2000) Multivariate behavioral a structural modeling approach to a multilevel random coefficients model. Multivar Behav Res 35(1):51–88. https://doi.org/10.2307/11648778

Satterthwaite FE (1946) An approximate distribution of estimates of variance components. Biom Bull 2(6):110–114. https://doi.org/10.2307/11648779

Searle SR, Casella G, McCulloch CE (1992) Variance components, 3rd edn. John Wiley & Sons, New York. https://doi.org/10.1375/twin.14.1.250

Smink WAC, Fox J-P, Sang TK, E., Sools, A. M., Westerhof, G. J., & Veldkamp, B. P. (2019) Understanding terapeutic change process research through multilevel modelling and text mining. Front Psychol 10:1186. https://doi.org/10.3389/fpsyg.2019.01186

Snijders TA, Bosker RJ (2012) Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, 2nd edn. SAGE Publications Ltd., London

Verbeke G, Molenberghs G (2003) The use of score tests for inference on variance components. Biometrics 59(2):254–262. https://doi.org/10.18637/jss.v067.i011

Wolf EJ, Harrington KM, Clark SL, Miller MW (2013) Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety Erika. Educ Psychol Meas 76(6):913–934. https://doi.org/10.18637/jss.v067.i012 **arXiv:NIHMS150003**

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.