Behavior Research Methods
https://doi.org/10.3758/s13428-023-02138-w

# Modeling variability in treatment effects for cluster randomized controlled trials using by-variable smooth functions in a generalized additive mixed model

**Sun-Joo Cho[1] · Kristopher J. Preacher[1] · Haley E. Yaremych[1] · Matthew Naveiras[1] · Douglas Fuchs[1] · Lynn S. Fuchs[1]**

Accepted: 25 April 2023


**Abstract**
Variability in treatment effects is common in intervention studies using cluster randomized controlled trial (C-RCT) designs. Such variability is often examined in multilevel modeling (MLM) to understand how treatment effects (TRT) differ based on the level of a covariate (COV), called TRT × COV. In detecting TRT × COV effects using MLM, relationships between covariates and outcomes are assumed to vary across clusters linearly. However, this linearity assumption may not hold in all applications and an incorrect assumption may lead to biased statistical inference about TRT × COV effects. In this study, we present generalized additive mixed model (GAMM) specifications in which cluster-specific functional relationships between covariates and outcomes can be modeled using by-variable smooth functions. In addition, the implementation for GAMM specifications is explained using the mgcv R package (Wood, 2021). The usefulness of the GAMM specifications is illustrated using intervention data from a C-RCT. Results of simulation studies showed that parameters and by-variable smooth functions were recovered well in various multilevel designs and the misspecification of the relationship between covariates and outcomes led to biased estimates of TRT × COV effects. Furthermore, this study evaluated the extent to which the GAMM can be treated as an alternative model to MLM in the presence of a linear relationship.

**Keywords** By-variable smooth function · Cluster randomized controlled trial · Functional covariate effects · Generalized additive mixed model · Nonlinear effects · Variability in treatment effects

## Introduction

### Study motivation: modeling variability in treatment effects

In the last two decades, the number of interventions to improve outcomes has increased. One popular study design for such interventions is a cluster randomized controlled trial (C-RCT). In C-RCT designs, clusters (e.g., schools, hospitals) are randomized to either the control or treatment group. As an example of the C-RCT design, D. Fuchs et al. (2021) evaluated the efficacy of an intervention called First-Grade Peer-Assisted Learning Strategies to improve students' scores of phonological awareness. In their study, teachers from 33 first-grade classrooms in eight elementary schools and their 491 students participated and the 33 teachers were assigned randomly to a control group or a treatment group. Hypotheses related to treatment effects are the primary target for inference, with the effectiveness of the intervention being assessed by determining whether a program, policy, or approach improves outcomes.

Researchers often report an average treatment effect (ATE), the average effect of a given intervention on the population of individuals or clusters. However, variability in treatment effects is commonly observed in intervention studies (Weiss et al., 2017). For example, Weiss et al. (2017) provided empirical evidence of cross-site treatment effect variation in 16 large multisite randomized trials. If variability in treatment effects is large and unexplained, knowing ATE will not tell us about how well an intervention works in particular settings. Examining variability in treatment effects is important to better understand when, how, why, and for whom interventions do or do not work (e.g., Raudenbush &

✉ Sun-Joo Cho
sj.cho@vanderbilt.edu

[1] Vanderbilt University, Nashville, TN, USA

Published online: 09 August 2023

Liu, 2000; Spybrook et al., 2016). For instance, Lawrence (2017) found that the effect of a vocabulary program was stronger for students with low initial vocabulary skills, not for all students in the study. Variability in treatment effects can inform the use of interventions for individuals (e.g., students) or clusters (e.g., schools, districts) by facilitating the targeting of resources toward individuals or clusters that are likely to benefit most from them.

Multilevel modeling (MLM) is a dominant analytic method to evaluate programs and interventions in the C-RCT design because the patterns of variability can be modeled within and across different levels of analysis (e.g., students in schools) due to the effects of covariates (e.g., variability in effects of student-level pretest scores across schools). In understanding variability in treatment effects, covariates in MLM have an important role to explain variability in treatment effects (e.g., Bloom & Spybrook, 2017; Tipton & Hedges, 2017). Specifically, an ATE that is conditional on preintervention factors (e.g., pretest scores, demographic information) can be examined in MLM to understand that treatment effects (TRT) differ based on the level of a covariate (COV), hereafter called TRT × COV. In multilevel settings, covariates may be measured at the individual level (level 1) or at the cluster level (level 2). They may also be categorical (e.g., student free lunch status) or continuous (e.g., student self-efficacy). In this study, we focus on interactions between continuous covariates measured at the individual level and a categorical TRT at the cluster level in two-level C-RCT designs.

## Current issues

In detecting TRT × COV using MLM, *linear* relationships are often assumed between covariates and outcomes (i.e., the effects of covariates on outcomes are the same over the range of each covariate), which can vary across clusters (e.g., Raudenbush & Liu, 2000). However, this assumption is rarely tested in detecting treatment effects using MLM (Preacher & Sterba, 2019). When there are unmodeled relationships between covariates and outcomes, it is expected that the estimates and standard errors of the treatment effects are biased (e.g., Cho et al., 2022; Harrell, 2015). As a flexible alternative approach, the generalized additive mixed model (GAMM; Lin & Zhang, 1999) can be used; here, smooth functions can be used for covariates that are not known to predict an outcome in a linear or parametric way. The smooth functions for a covariate in GAMM represent *functional covariate effects*, which means that these covariate effects refer to the intrinsic structure of the data rather than to their explicit form (Ramsay & Silverman, 2005; p. 38). Cho et al. (2022) specified GAMM (with an identity link) to model functional covariate effects by levels of a TRT covariate using smooth functions, which are assumed to be the same across clusters (e.g., classes in which students are nested). Although Cho et al. (2022) showed that the specified GAMM fits well to the empirical data set in their study, it limits the applications of the specified GAMM to cases in which functional covariate effects are the same or similar across clusters. It has not been shown how to specify GAMM to detect functional covariate effects which vary across clusters (hereafter, called *cluster-specific functional covariate effects*) in C-RCT designs for unbiased statistical inference on the TRT × COV effects.

Functional data analysis (FDA; Ramsay & Silverman, 2002, 2005) is closely related to GAMM for modeling functional covariate effects (see examples for relations between FDA and GAMM in Wood (2017), pp. 390-397). FDA and its extensions to mixed-effects modeling (known as functional mixed-effects modeling, Guo (2002) have been applied mainly to longitudinal or time-series data in which functional covariates are time-related covariates. Examples in Ramsay and Silverman (2002, 2005) are for FDA with functional covariates such as age, years, days, and reaction time intervals. As another example, Fine et al. (2019) presented a functional mixed-effects model for longitudinal data to model complex nonlinear trajectories using person(cluster)-specific smooth functions. GAMM and functional mixed-effects models are general modeling frameworks, and a statistical package to estimate parameters of the models is presented for the general model. Therefore, it may not be straightforward to researchers how to apply the general models to detect level-specific (called the *unconflated* solution [e.g., Preacher & Sterba, 2019]) TRT × COV effects in cross-sectional data from C-RCT designs in the presence of cluster-specific functional relationships between non-time-related covariates and outcomes. In this study, GAMM was chosen over FDA because the smooth functions to model cluster-specific functional covariate effects of interest were developed within a GAMM framework (Wood, 2017).

## Study purposes

The first purpose of this study is to present GAMM specifications for detecting TRT × COV effects in C-RCT designs, controlling for cluster-specific functional relationships between continuous covariates and outcomes using a *by-variable smooth function*. The by-variable smooth function estimates a smooth function of a continuous covariate for each level of a categorical variable (i.e., control vs. treatment groups and clusters in C-RCT designs). For parameter estimation, we utilize the `gam` function in the `mgcv` package (Wood, 2021) in R (R Core Team, 2021) using penalized iteratively re-weighted least squares (PIRLS; Wood, 2017). Because the `mgcv` package was developed for a general specification of GAMM, the specificity of implementation for GAMM specifications in detecting TRT × COV effects in C-RCT designs and the description of selected estimation

methods are needed. Thus, the second purpose of the current study is to present and evaluate the estimation method of the specified GAMMs in using mgcv package. In this study, use of a smoother called the *by-variable smoother* in the mgcv package is explained to represent a by-variable smooth function in GAMM specifications for C-RCT designs, which has not been shown in the literature. The GAMM specifications and their parameter estimation methods are illustrated using an empirical data set from a C-RCT design. In addition, a simulation study is presented to investigate (a) the accuracy of GAMM parameter estimates and their standard errors in various multilevel designs, (b) consequences of modeling linear relationships between covariates and outcomes in the presence of cluster-specific functional relationships between them, and (c) the recovery of GAMM parameters in the presence of cluster-specific linear relationships between covariates and outcomes.

The remainder of this paper is organized as follows. In Section 2, we present the GAMM specifications, and provide the estimation method and testing using the `mgcv` package. In Section 3, the empirical illustration is shown. In Section 4, the simulation studies are presented. In Section 5, a summary and a discussion are provided.

## Methods

In this section, GAMM is specified with a comparison to MLM, and its parameter estimation method and testing in the mgcv package are described.

### Model specifications

In the model specifications below, a continuous level-1 covariate ($x_{ij}$ where $i$ is an index for an individual and $j$ is an index for a cluster) is decomposed into a level-1 part of $(x_{ij} - x_{.j})$ and a level-2 part $(x_{.j})$ to model level-specific interaction TRT $\times$ COV effects by centering $x_{ij}$ at its cluster mean $x_{.j}$. An inferential goal using the specified models is to test the following hypothesis:
The level-2 part of $x_{ij}$ (COV; $x_{.j}$) moderates the effect of level-2 variable $z_j$ (TRT) on the outcome variable $y_{ij}$.

#### MLM specification with random effects

The following MLM specification utilizes notation and symbols from Raudenbush and Bryk (2002), and models *linear* relationships between covariates and outcomes. MLM with a random intercept $\beta_{0j}$ and a random slope $\beta_{1j}$ is written as follows for a two-level nested design:

$$\text{Level 1: } y_{ij} = \beta_{0j} + \beta_{1j}(x_{ij} - x_{.j}) + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}x_{.j} + \gamma_{02}z_j + \gamma_{03}x_{.j}z_j + u_{0j}$$
$$\text{and } \beta_{1j} = \gamma_{10} + \gamma_{11}z_j + u_{1j}$$

Reduced Form:

$$\begin{aligned} y_{ij} = {} & \gamma_{00} + \gamma_{10}(x_{ij} - x_{.j}) + \gamma_{01}x_{.j} + \gamma_{02}z_j + \gamma_{03}x_{.j}z_j \\ & + \gamma_{11}(x_{ij} - x_{.j})z_j + u_{0j} + (x_{ij} - x_{.j})u_{1j} + r_{ij}, \quad (1) \end{aligned}$$

where $\gamma_{00}$ is a fixed intercept; $\gamma_{10}$ is a fixed effect of the level-1 component $(x_{ij} - x_{.j})$ of COV $x_{ij}$ when $z_j = 0$; $\gamma_{01}$ is a fixed effect of the level-2 component $x_{.j}$ of COV $x_{ij}$ when $z_j = 0$; $\gamma_{02}$ is a fixed effect of the dummy coded level-2 TRT $z_j$ (a conditional treatment effect when $x_{.j} = 0$); $\gamma_{03}$ is a fixed linear interaction effect of the level-2 component of COV $(x_{.j})$ and the dummy coded level-2 TRT $z_j$; $\gamma_{11}$ is a fixed linear interaction effect of the level-1 component of COV $(x_{ij} - x_{.j})$ and the dummy coded level-2 TRT $z_j$; $u_{0j}$ is a random intercept; $u_{1j}$ is a random slope; and $r_{ij}$ is random error. It is assumed that the random effects, $[u_{0j}, u_{1j}]'$, follow a multivariate normal $(MVN)$ distribution, $[u_{0j}, u_{1j}]' \sim MVN([\gamma_{00}, \gamma_{10}]', \boldsymbol{\tau})$, where random intercept variance is indicated by $\tau_{00}$, random slope variance is $\tau_{11}$, and their covariance is $\tau_{01}$ in $\boldsymbol{\tau}$. It is also assumed that the random error, $r_{ij}$, follows a normal $(N)$ distribution, $r_{ij} \sim N(0, \sigma^2)$, where random error variance is indicated by $\sigma^2$.

In Equation 1, focal parameters are level-2 fixed effects for TRT $\times$ COV in C-RCTs, $\gamma_{02} + \gamma_{03}x_{.j}$. In addition, it is expected that the standard errors of fixed-effect estimates $\widehat{\gamma}_{02}$ and $\widehat{\gamma}_{03}$ are underestimated when the random effects $[u_{0j}, u_{1j}]'$ are ignored as controlling parameters in the presence of the linear relationship between covariates and outcomes (see Longford, 1993, pp. 53–56 for technical details). This bias leads to overestimating the statistical significance of the fixed effects. It is expected that the functional relationships between covariates and outcomes can be another source of bias for $\gamma_{02} + \gamma_{03}x_{.j}$ when there are cluster-specific functions relating them.

#### GAMM specification with by-variable smooth functions

GAMM can be specified to model *functional* relationships between covariates and outcomes as follows:

$$\begin{aligned} y_{ij} = {} & \gamma_{00} + \gamma_{02}z_j + f_1(x_{.j})(z_j = 0) + f_1(x_{.j})(z_j = 1) \\ & + f_2(x_{ij} - x_{.j})(z_j = 0) + f_2(x_{ij} - x_{.j})(z_j = 1) \\ & + f_3(x_{ij} - x_{.j})(Cluster_j = j) + u_{0j} + r_{ij}, \quad (2) \end{aligned}$$

where $\gamma_{02}$ is the mean of all smooth functions when $z_j$ ($z_j = 0$ for a control group; $z_j = 1$ for a treatment group) is specified as a factor in `R`; $f_1(x_{.j})(z_j = 0)$ is the smooth function of a level-2 component $x_{.j}$ of COV $x_{ij}$ where

$z_j = 0$ (i.e., functional level-2 effect for a control group); $f_1(x_{.j})(z_j = 1)$ is the smooth function of a level-2 component $x_{.j}$ of COV $x_{ij}$ where $z_j = 1$ (i.e., functional level-2 effect for a treatment group); $f_2(x_{ij} - x_{.j})(z_j = 0)$ is the smooth function of a level-1 component $x_{ij} - x_{.j}$ of COV $x_{ij}$ where $z_j = 0$ (i.e., functional level-1 effect for a control group); $f_2(x_{ij} - x_{.j})(z_j = 1)$ is the smooth function of a level-1 component $x_{ij} - x_{.j}$ of COV $x_{ij}$ where $z_j = 1$ (i.e., functional level-1 effect for a treatment group); and $f_3(x_{ij} - x_{.j})(Cluster_j = j)$ is a cluster-specific by-variable smooth function to model functional covariate effects over $(x_{ij} - x_{.j})$ for each cluster.[1] Cho et al. (2022) did not consider $f_3(x_{ij} - x_{.j})(Cluster_j = j)$ in detecting TRT × COV effects. Here, the functional level-1 and level-2 effects mean that the effects change at a different rate as a function of changes in $x_{ij} - x_{.j}$ and $x_{.j}$, respectively.

In Equation 2, focal parameters are level-2 functional effects for TRT × COV in C-RCTs, $\gamma_{02} + \{f_1(x_{.j})(z_j = 1) - f_1(x_{.j})(z_j = 0)\}$. The two by-variable smooth functions of the level-1 COV, $f_2(x_{ij} - x_{.j})z_j$ and $f_3(x_{ij} - x_{.j})Cluster_j$, are controlling terms to describe the functional relationship between $(x_{ij} - x_{.j})$ and $y_{ij}$ for each cluster adequately so as to have unbiased estimates for the focal parameters.

The difference among the three kinds of by-variable smooth functions in Eq. 2 is that $f_1(x_{.j})z_j$ and $f_2(x_{ij} - x_{.j})z_j$ have two levels of a factor ($z_j = 0$ for a control group; $z_j = 1$ for a treatment group), while $f_2(x_{ij} - x_{.j})Cluster_j$ has multiple levels of a factor. As a general form of $f_1(x_{.j})z_j$, $f_2(x_{ij} - x_{.j})z_j$, and $f_3(x_{ij} - x_{.j})Cluster_j$, a by-variable smooth function ($f_h(x_h)factor_j$, where $h$ is an index for a smooth function) creates a smooth function of a covariate ($f_h(x_h)$) for each factor level ($factor_j$) in GAMM. The smooth function $f_h(x_h)factor_j$ is specified as a weighted sum of a set of basis functions over the covariate $x_h$ for each factor level:

$$f_h(x_h)factor_j = \sum_{k=1}^{K} \delta_{hk.j} b_{hk.j}(x_h)factor_j, \tag{3}$$

where $j$ is an index for a factor level ($j = 1, \ldots, J$), $k$ is an index for a basis function ($k = 1, \ldots, K$), $\delta_{hk.j}$ is a basis coefficient for each factor level (each factor level for $z_j$ and $Cluster_j$ in Eq. 2), and $b_{hk.j}(x_h)$ is the $k$th basis function for each factor level in a smooth function. Note that we use the subscript $j$ as an index for both cluster and factor because factor in the current study is at the cluster level in C-RCTs.

[1] We use a notation of a by-variable smooth function as $f(x)factor$ following notations in GAMM (e.g., Wood, 2017, p. 326), which means that a smooth function $f_h$ (where $h$ is an index for a smooth function) is conditional on a factor.

### Comparisons between MLM and GAMM

Table 1 shows the comparable terms of MLM (Equation 1) and GAMM (Equation 2). Because $\gamma_{03}$ is a fixed linear interaction effect in the focal parameters of MLM ($\gamma_{02} + \gamma_{03}x_{.j}$), it is expected that the differences between a control group and a treatment group (TRT effects) change at a constant rate as a function of changes in $x_{.j}$. In contrast, the differences between the two groups (TRT effects) can change at different rates as a function of changes in $x_{.j}$ in the focal parameter of GAMM ($\gamma_{02} + \{f_1(x_{.j})(z_j = 1) - f_1(x_{.j})(z_j = 0)\}$) due to smooth functions.

In addition, Fig. 1 illustrates differences between random effects in MLM and by-variable smooth functions in GAMM with 10 clusters as an example. As shown in Fig. 1(a), random effects in MLM ($u_{0j}$ and $u_{1j}$ in Equation 1) can be specified to model *linear* relationships between covariates and outcomes that vary across clusters. Figure 1(b) and (c) present two different types of *functional* relationships between covariates and outcomes across clusters, which can be specified using a smooth function or a by-variable smooth function in GAMM: a global smooth function as depicted in Fig. 1(b) can be modeled, assuming that the functional relationship is the same across clusters; and varying smooth functions having different wiggliness across clusters can be modeled as presented in Fig. 1(c). Cho et al. (2022) presented GAMM specifications for the case depicted in Fig. 1(b). This study presents the new case depicted in Fig. 1(c) in GAMM in testing TRT × COV.

## Estimation and testing

### Estimation of by-variable smooth functions and random effects

A thin plate regression spline (TPRS; Wood, 2017, 5.5.1) is recommended for factor-smooth interactions as a default in the mgcv package (Wood, 2021); thus, we used TPRS

**Table 1** Comparability between MLM and GAMM Specifications

| MLM (Eq. 1) | GAMM (Eq. 2) |
|---|---|
| $\gamma_{00}$ | $\gamma_{00}$ |
| $\gamma_{10}(x_{ij} - x_{.j})$ | $f_2(x_{ij} - x_{.j})(z_j = 0)$ |
| $\gamma_{01}x_{.j}$ | $f_1(x_{.j})(z_j = 0)$ |
| $\gamma_{02}z_j$ | $\gamma_{02}z_j$ |
| $\gamma_{03}x_{.j}z_j$ | $f_1(x_{.j})(z_j = 1)$ |
| $\gamma_{11}(x_{ij} - x_{.j})z_j$ | $f_2(x_{ij} - x_{.j})(z_j = 1)$ |
| $u_{0j}$ | $u_{0j}$ |
| $(x_{ij} - x_{.j})u_{1j}$ | $f_3(x_{ij} - x_{.j})(Cluster_j = j)$ |
| $r_{ij}$ | $r_{ij}$ |

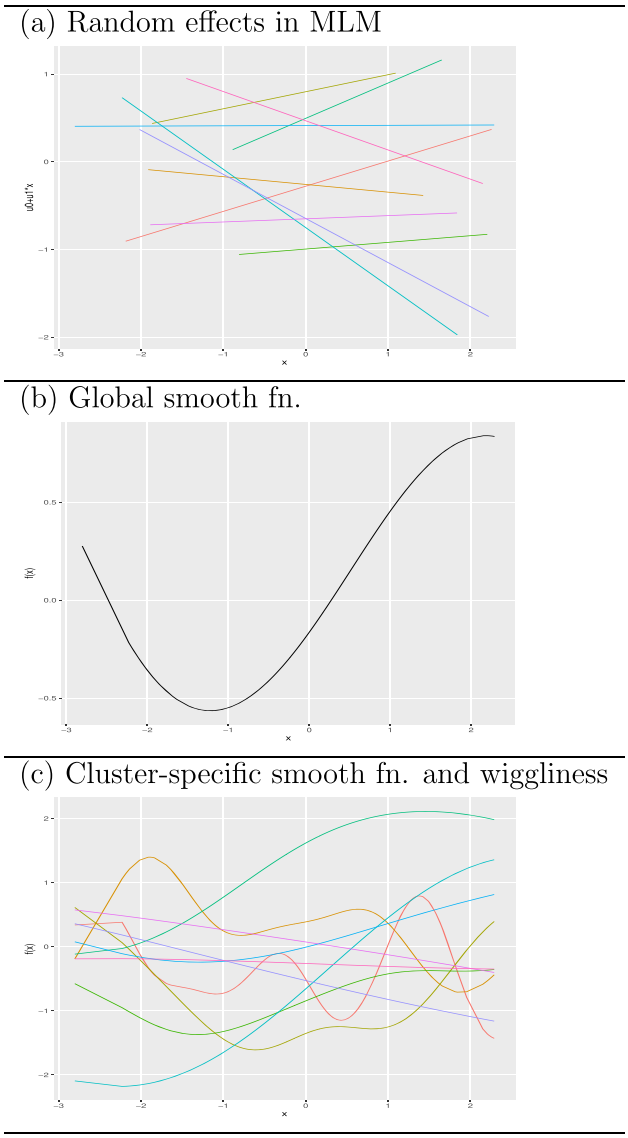


**Fig. 1** Random effects in MLM (a) vs. global smooth function (b) and by-variable smooth functions in GAMM (c)

for all kinds of by-variable smooth functions in the current study. TPRS estimates a smooth function $f_h(x_h)$ for each level of a factor by finding the predicted function $\tilde{f}_h(x_h)$ that minimizes

$$\|\mathbf{y} - f_h(x_h)\|^2 + \lambda_h J(f), \tag{4}$$

where $\mathbf{y}$ is a vector of data, $\lambda_h$ is a *smoothing parameter* controlling the trade-off between goodness of fit and smoothness of $f_h(x_h)$, and $J(f)$ is a penalty function measuring the "wiggliness" of $f_h(x_h)$. Having $\lambda_h \approx \infty$ results in a straight line estimate for $f_h(x_h)$ (i.e., totally smooth), whereas having $\lambda_h = 0$ leads to an un-penalized piecewise linear regression estimate.

The number of basis functions ($K$) for smooth functions should be selected to obtain a good fit. Oversmoothing will occur when $K$ is too small, and computation time slows if $K$ is too large. Therefore, the value of the $k$-index should be assessed in order to determine whether the selected $K$ is appropriate. The $k$-index is a measure of the remaining pattern in the residuals (see Cho et al. [2022] for details). A $k$-index below 1 indicates that there is a pattern remaining in the residuals that has been missed due to the specified $K$ being too small, and a larger $K$ should be considered. The $k$-index can be obtained through the `gam.check` function in mgcv. In addition, the corrected Akaike information criterion (corrected AIC; Wood et al., 2016) is used to select a model among candidate models differing in $K$.

The 'wiggliness' of a by-variable smooth function $f_h(x_h)$ $factor_j$ is controlled by a quadratic smoothing penalty (e.g., Wood, 2017). The quadratic smoothing penalty for the model can be written as:

$$\lambda_{h.j} \boldsymbol{\delta}_{h.j}^T \mathbf{S}_{h.j} \boldsymbol{\delta}_{h.j}, \tag{5}$$

where $\lambda_{h.j}$ is a smoothing parameter, $\boldsymbol{\delta}_{h.j}$ is a vector of basis coefficients, and $\mathbf{S}_{h.j}$ is a penalty matrix embedded as a diagonal block in a matrix, for a by-variable smooth function $f_h(x_h)factor_j$. The elements of $\mathbf{S}_{h.j}$ are known and are determined by the chosen basis functions (for TPRS in this study). The penalty matrix $\mathbf{S}_{h.j}$ can be extracted using the smoothCon function in the mgcv package. The parameter $\lambda_{h.j}$ controls the trade-off between goodness of fit and model smoothness.

Imposing the penalty for a smooth function (e.g., $f_{h.j}(x_h)$ $factor_j$) is equivalent to having a prior on basis coefficients $\boldsymbol{\delta}_{h.j}$ using a multivariate normal ($MVN$) distribution with mean vector $\mathbf{0}$ and the variance matrix $(\lambda_{h.j}\mathbf{S}_{h.j})^{-1}$ (Kimeldorf & Wahba, 1970; Silverman, 1985; Wahba, 1983):

$$\boldsymbol{\delta}_{h.j} \sim MVN(\mathbf{0}, (\lambda_{h.j}\mathbf{S}_{h.j})^{-1}). \tag{6}$$

The smoothing penalty is a measure of how much the basis coefficients $\boldsymbol{\delta}_{h.j}$ deviate from 0. This implies shrinkage in $\widehat{\boldsymbol{\delta}}_{h.j}$ towards 0, as occurs in random effects in mixed-effects models. A random effect is equivalent to a smooth function with penalty matrix $I_h$ called *ridge penalty* (i.e., $\mathbf{S}_{h.j} = I_h$, where $I_h$ is an identity matrix) (Wood, 2017). Having $\mathbf{S}_{h.j} = I_h$ means that there is a *pure* shrinkage penalty on basis coefficients $\boldsymbol{\delta}_j$ (as random effects) which penalizes all deviations from 0 regardless of any patterns in those deviations. As a result, a random intercept ($u_{0j}$) in Equation 2 can be estimated as basis coefficients $\boldsymbol{\delta}_h$ in GAMM:

$$\boldsymbol{\delta}_h = u_{0j} \sim N(0, (\lambda_h \mathbf{I}_h)^{-1}), \tag{7}$$

Here, one can see that the variance of the random effect is the inverse of the smoothing parameter ($\lambda_h$). The `gam.vcomp` function in mgcv converts smoothing parameter estimates to the variance estimates of the random effects.

In Appendix S.1, we explain specifications of smoothers for the by-variable smooth functions ($f_1(x_{.j})z_j$, $f_2(x_{ij} - x_{.j})z_j$, and $f_3(x_{ij} - x_{.j})Cluster_j$) and the random intercept ($u_{0j}$) in Equation 2 in the mgcv package. The `gam` function was selected as a main GAMM fitting routine in the mgcv package.

### Parameter estimation

The specified GAMM in Equation 2 can be rewritten as follows for parameter estimation:

$$y_{ij} = \mathbf{X}'\boldsymbol{\beta} + r_{ij}, \tag{8}$$

where $\mathbf{X}$ is a design matrix having all components of the model and all the basis functions for the by-variable smooth functions ($f_1(x_{.j})z_j$, $f_2(x_{ij} - x_{.j})z_j$, and $f_3(x_{ij} - x_{.j})Cluster_j$), $\boldsymbol{\beta}$ is a set of parameters including the coefficients of fixed effects ($[\gamma_{00}, \gamma_{02}]$), a random effect ($u_{0j}$; estimated as basis coefficients with a random effect smoother), and the basis coefficients ($\boldsymbol{\delta}$) (i.e., $\boldsymbol{\beta} = [\gamma_{00}, \gamma_{02}, u_{0j}, \boldsymbol{\delta}]'$). In the mgcv package, given smoothing parameters ($\boldsymbol{\lambda}$) estimated with REML and the variance ($\tau_{00}$) of the random effect $u_{0j}$, the default option for estimating the parameters ($\boldsymbol{\beta}$) is PIRLS: `optimizer=c("outer","newton")`. The standard errors of $\widehat{\boldsymbol{\beta}}$ are obtained with the diagonal terms in the square root of the estimated covariance matrix of $\mathbf{X}'\boldsymbol{\beta}$ (Wood, 2017, p. 341).

### Testing

The following null hypothesis can be tested to determine whether or not a smooth function $f_h(x_h)$ of a covariate $x_h$ is distinguishable from zero: $H_0 : f_h(x_h)factor_j = 0$ for all $x_h$ in the range of interest. A test statistic for $f_h(x_h)factor_j$ is:

$$T_r = \widehat{\mathbf{f}}_h^T \mathbf{V}_{f_h}^- \widehat{\mathbf{f}}_h, \tag{9}$$

where $r$ is the rounded effective degrees of freedom ($edf$; the number of parameters to represent a smooth function) of $f_h(x_h)$, the $\widehat{\mathbf{f}}_h$ is the vector of $f_h(x_h)$ evaluated at the $x_h$ values, and $\mathbf{V}_{f_h}^-$ is a rank $r$ pseudo-inverse of $\mathbf{V}_{f_h}$ ($\mathbf{V}_{f_h} = \mathbf{X}_h \mathbf{V}_{\delta h} \mathbf{X}_h^T$, where $\mathbf{X}_h$ are basis functions and $\mathbf{V}_{\delta h}$ is the covariance matrix of basis coefficient estimates) (Wood, 2017, pp. 305-306). Under $H_0$, the test statistic $T_r$ follows a chi-square distribution ($T_r \sim \chi_r^2$) (Wood, 2013).

## Empirical study

In this section, illustrations of a GAMM specification are presented using an empirical data set from Baranov et al. (2020a) to detect a categorical level-2 TRT × a continuous level-1 COV in a C-RCT design. The empirical data set was downloaded from Baranov et al. (2020b). The purpose of the study in Baranov et al. (2020a) was to evaluate the effect of a psychotherapy intervention (the Thinking Healthy Program) on treating maternal depression in rural Pakistan. The intervention was designed to reduce the incidence of depression among prenatally depressed mothers, and follow-up surveys were conducted at 6 and 12 months postpartum to evaluate effectiveness of the intervention. The analytic goal is to detect the treatment effect, which is done by comparing differences between control and treatment groups in changes in depression scores from pre- to post- treatment.

### Participants and measures

The subset of the data set in Baranov et al. (2020a) we analyzed included 818 women (individuals) nested within 40 communities (clusters). The trial was randomized across 40 communities: 20 clusters were randomly assigned to the intervention arm, with the remaining 20 clusters assigned to the control arm. There were 400 women in the control arm and 418 in the intervention arm. Cluster size ranged from 26 to 35.

Among tests to measure maternal mental health in Baranov et al. (2020a), Hamilton Depression Rating (HDR) scores were chosen as a continuous measure of depression severity. Lower HDR scores indicate lower degrees of depression. HDR scores were obtained at baseline, 6-month follow-up, and 12-month follow-up. Baseline and 6-month follow-up HDR were selected in this study for illustrative purposes. There are no missing data in baseline and 6-month follow-up. The mean HDR score was 14.575 ($SD = 4.072$) at the baseline, and 6.534 ($SD = 7.035$) at the 6-month follow up.

### Analyses and results

The R code for the empirical example is shown in Appendix S.2.

#### Step 1: Fitting unconditional GAMM and exploratory graphical analysis

To show dependency in 6-month follow-up HDR scores (outcomes) due to communities (clusters), the unconditional GAMM[2] was fitted to the data to calculate an intraclass cor-

[2] Because a parametric random intercept is considered in Equation 2, the unconditional GAMM is the same as the unconditional MLM.

relation coefficient *(ICC)*. *ICC* was .164 (= 8.118/[8.118 + 41.323] where $\widehat{\tau}_{00} = 8.118$, $\widehat{\sigma}^2 = 41.323$), which indicates that there is non-ignorable dependency in the 6-month follow-up HDR scores due to community clustering.

The starting analysis to consider the cluster-specific functional covariate effects using the GAMM specification (Equation 2) is to create scatter plots of the outcome ($y_{ij}$) vs. a covariate of interest ($x_{ij}$) by cluster at level 1 and by control vs. treatment groups ($z_j$) at level 2. For an unconflated solution, $x_{ij}$ was cluster-mean centered: $x_{ij} - x_{\cdot j}$ as a level-1 COV and $x_{\cdot j}$ as a level-2 COV. A scatter plot can be used to observe variability in the linear covariate effects (in the linear model) vs. functional covariate effects (in the by-variable smooth function) across clusters for the level-1 COV, and control vs. treatment groups for the level-1 and level-2 COV, by adding the predicted linear model and the predicted smooth function by clusters and by groups. As shown in Fig. 2, the shapes of linear or nonlinear relationships between cluster-mean cen-

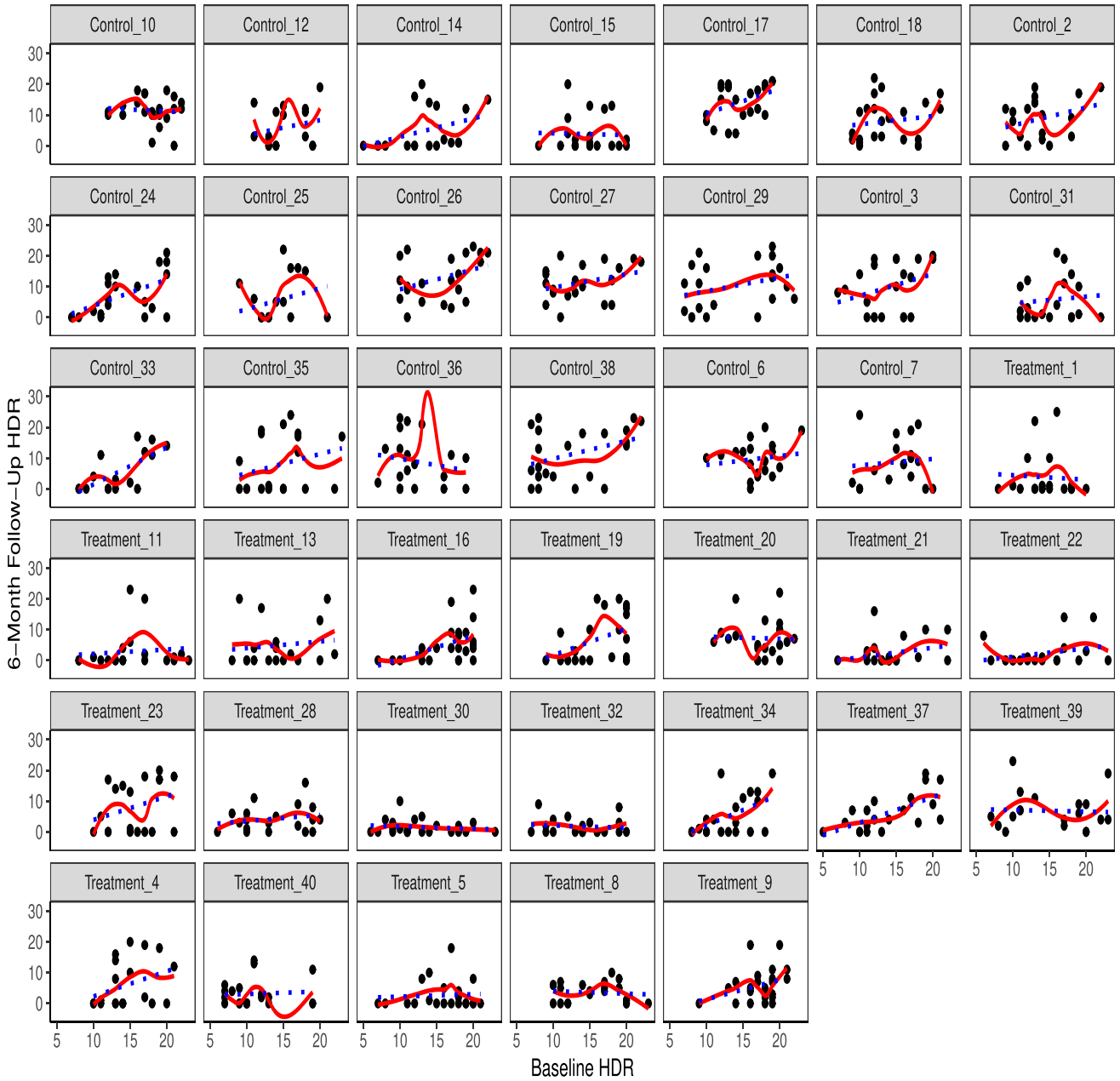


**Fig. 2** Empirical Study: Scatter plots of $y_{ij}$ vs. $x_{ij} - x_{\cdot j}$ by community (denoted by $z_j$_community id). *Note.* The red curves indicate the fitted smooth lines and the dotted blue lines indicate the fitted linear lines

tered baseline HDR (level-1 COV $x_{ij} - x_{.j}$) and 6-month follow-up HDR ($y_{ij}$) differ across communities (clusters). Ignoring the variability in shapes across communities may result in biased estimates of focal parameters, TRT × COV in C-RCTs, $\gamma_{02} + \{f_1(x_{.j})(z_j = 1) - f_1(x_{.j})(z_j = 0)\}$. In addition, it is apparent that there are functional relationships, presented with smooth lines, deviating from the linear dotted lines in most communities. For example, in the panel of community id=25 (from a control group) in Fig. 2, there is an increasing relationship predicted with a linear regression model. However, based on the predicted smooth function there is a decreasing relationship for baseline HDR scores of 10 or less and an increasing relationship for the middle range of HDR scores. Figure 3 presents a scatter plot of 6-month follow-up HDR ($y_{ij}$) vs. cluster-mean centered baseline HDR (level-1 COV $x_{ij} - x_{.j}$) by groups ($z_j$) (top) and a scatter plot of baseline HDR cluster means (level-2 COV $x_{.j}$) by groups ($z_j$) (bottom). These figures show that the effects differ over the ranges of $x_{ij} - x_{.j}$ and $x_{.j}$ (i.e., functional effects), although there are small differences in the two predicted lines by the linear regression model and smooth functions for each group. GAMM was fit in the next step based on empirical evidence of variability in the shape and wiggliness of smooth functions across clusters shown in Fig. 2.

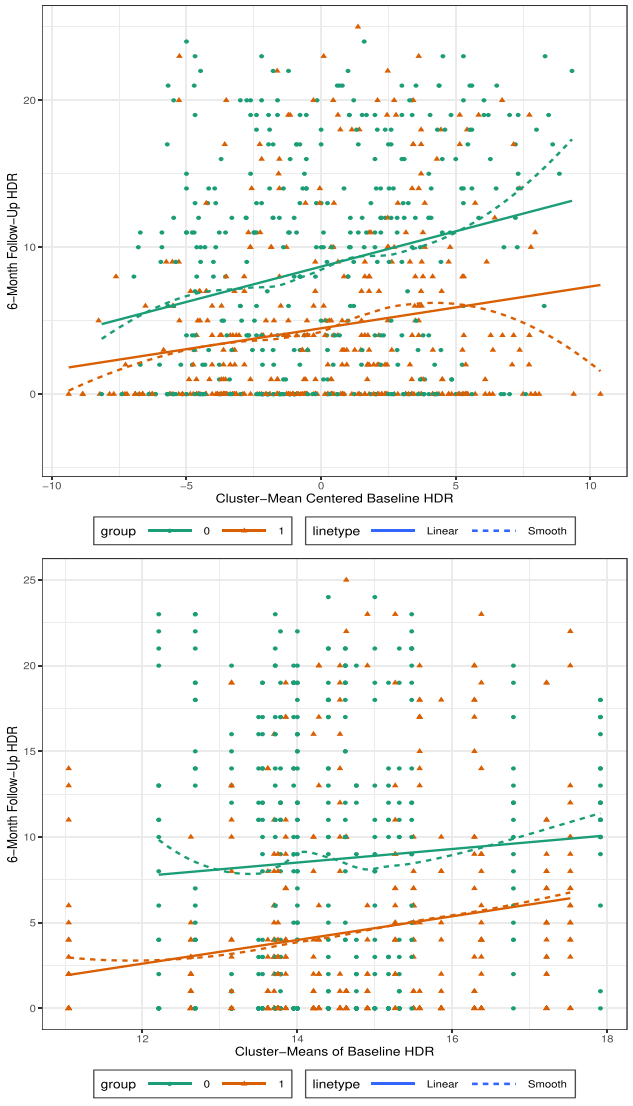


**Fig. 3** Empirical Study: Scatter plots of $y_{ij}$ vs. $x_{ij} - x_{.j}$ by $z_j$ (top) and $y_{ij}$ vs $x_{.j}$ by $z_j$ (bottom)

### Step 2: Adding covariates (TRT and COV) to the unconditional GAMM

A dummy-coded level-2 TRT ($z_j = 0$ for a control group; $z_j = 1$ for a treatment group) and a cluster-centered level-1 COV (baseline HDR $x_{ij}$; $x_{ij} - x_{.j}$ and $x_{.j}$) were selected as covariates in GAMM. Based on evidence in Figs. 2 and 3, a by-variable smooth function of $x_{ij} - x_{.j}$ ($f_2(x_{ij} - x_{.j})Cluster_j$) was added to the unconditional GAMM. Smooth functions of $x_{.j}$ and $x_{ij} - x_{.j}$ by $z_j$ ($f_1(x_{.j})z_j$ and $f_1(x_{ij} - x_{.j})z_j$) were considered because the differences between the two fitted lines appeared to be different depending on the levels of baseline HDR scores (that is, functional covariate effects).

The GAMM (Equation 2) was fitted with different values of $K$ ranging from 3 to 12. Five basis functions ($K = 5$) were chosen for all by-variable smooth functions in the model for having a $k$-index close to 1 and the smallest correctedAIC. These results indicate that $K = 5$ is adequate to obtain a good fit. Figure 4 shows the prediction by GAMM for each community to present model-data fit by GAMM (thick line) and by MLM (dotted line). In the figure, it is observed that data were better predicted with GAMM than MLM, especially for community id=16, 18, 25, 26, 36, and 38. As controlling effects for a focal parameter (TRT × COV), a smooth function of level-1 COV for the control group ($f_2(x_{ij} - x_{.j})(z_j = 0)$) is statistically significant (see Table 2). In addition, smooth functions for community id=2, 3, 6, 7, 10, 12, 14, 15, 17, 18, 24-27, 29, 31, 33, 35, 36, and 38 were distinguishable from zero based on a chi-square test (see Table 2).

### Step 3: Interpreting results

Table 2 shows the results of GAMM. A significant fixed TRT effect was found ($\widehat{\gamma}_{02} = -4.173$, SE=0.717), indicating that women in the treatment group of the Thinking Healthy Program have 4.173 lower HDR scores on average than women in the control group. The corresponding effect size, Hedges' $g$, is 0.702. A smooth function for a treatment group ($f_1(x_{.j})(z_j = 1)$) was statistically significant ($T_1 = 4.615$, $p$-value=.032). The effective degrees of free-

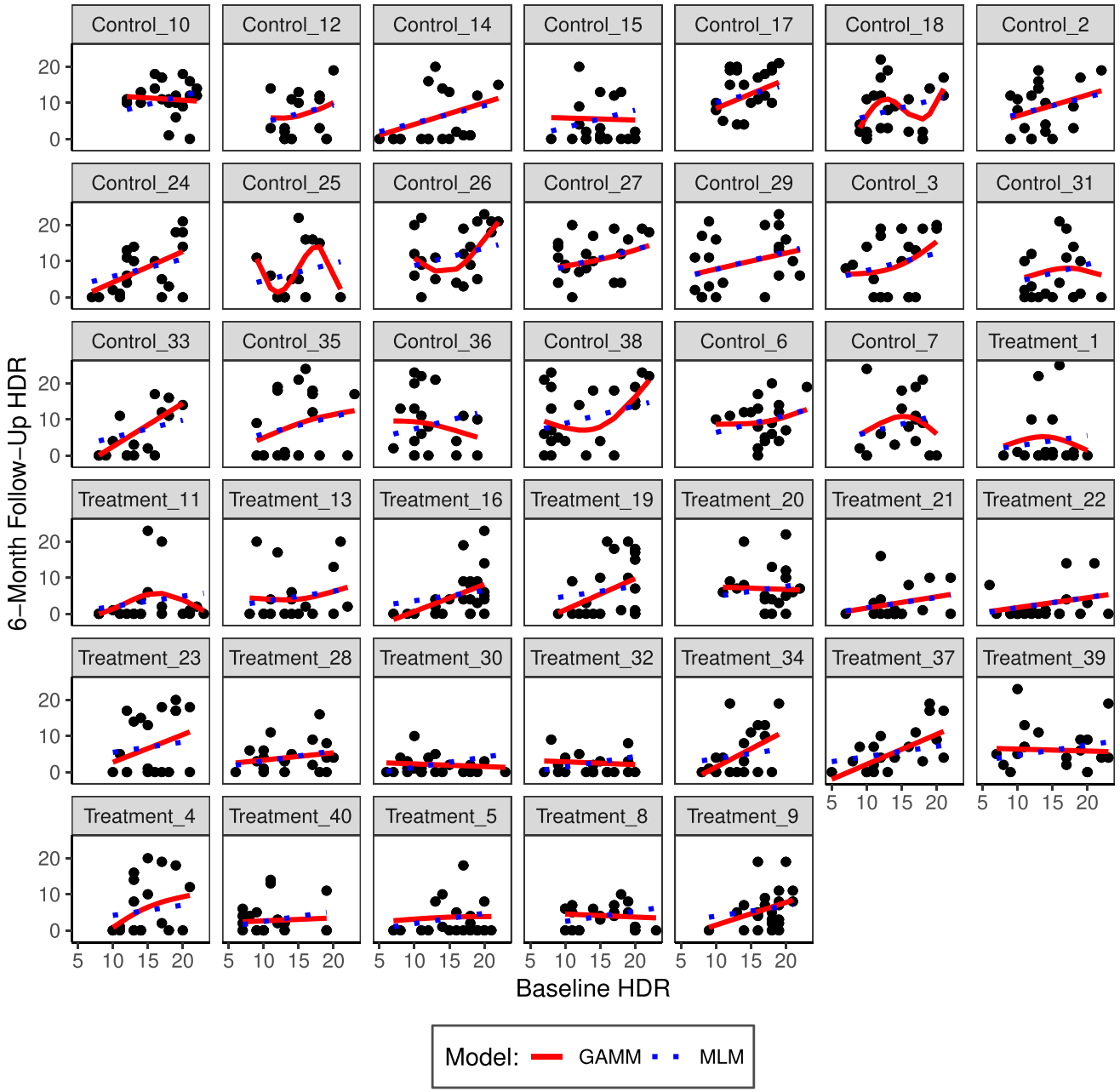


**Fig. 4** Empirical Study: Prediction by GAMM (thick line) and MLM (dotted line) for each community (denoted by $z_j$_community id)

dom ($edf$) being 1 for the smooth function indicates that there is a linear relationship. It is of interest to interpret the level-2 TRT effect ($z_j$) on cluster means ($y_{.j}$) at any chosen value of the level-2 part of COV ($x_{.j}$). Figure 5 (top) shows the effect of $z_j$ on $y_{.j}$ by quantiles of $x_{.j}$ (0.1, 0.25, 0.5, 0.75, 0.9). Figure 5 (bottom) shows the level-2 TRT effects ($\gamma_{02} + \{f_1(x_{.j})(z_j = 1) - f_1(x_{.j})(z_j = 0)\}$) across all levels of $x_{.j}$. The region of significance for the effect of cluster-level HDR ($x_{.j}$) was [11.048, 16.803], shown in the vertical lines of Fig. 5 (bottom). As shown in Fig. 5 (bottom), the level-2 TRT effects decrease as cluster means of baseline HDR increase. This result suggests that the Thinking Healthy Program is more effective at the low-medium levels of baseline HDR than at the higher levels of baseline HDR.

## Comparisons between GAMM and MLM

For MLM, restricted maximum likelihood (REML) estimation was implemented to obtain unbiased estimates of

**Table 2** Empirical study: Results of GAMM and MLM

| Fixed effects | GAMM | | MLM | |
|---|---|---|---|---|
| | EST | SE | EST | SE |
| Intercept[$\gamma_{00}$] | **8.593** | 0.513 | 2.942 | 5.893 |
| $x_{ij} - x_{.j}$[$\gamma_{10}$] | - | | **0.481** | 0.084 |
| $x_{.j}$[$\gamma_{01}$] | - | | 0.392 | 0.407 |
| $z_j$[$\gamma_{02}$] | **-4.173** | 0.716 | −8.534 | 7.752 |
| $x_{.j}z_j$[$\gamma_{03}$] | - | | 0.295 | 0.531 |
| $(x_{ij} - x_{.j})z_j$[$\gamma_{11}$] | - | | −0.197 | 0.115 |
| Random effects | | | | |
| | EST | | EST | |
| $\sqrt{\tau_{00}}$ | 1.778 | | 1.908 | |
| $\sqrt{\tau_{11}}$ | - | | 0.0004 | |
| $\tau_{01}$ | - | | 0.000 | |
| $\sigma$ | 6.124 | | 6.256 | |
| Smooth functions | | | | |
| | $Ref.edf$ | $T_r(p\text{-value})$ | | |
| $f_1(x_{.j})(z_j = 0)$ | 1.000 | 1.460(0.227) | - | |
| $f_1(x_{.j})(z_j = 1)$ | 1.000 | 4.615(0.032) | - | |
| $f_2(x_{ij} - x_{.j})(z_j = 0)$ | 1.000 | 7.780(0.005) | - | |
| $f_2(x_{ij} - x_{.j})(z_j = 1)$ | 1.000 | 1.789(0.181) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 1)$ | 1.937 | 1.474(0.291) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 2)$ | 1.001 | 8.826(0.003) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 3)$ | 1.997 | 5.263(0.006) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 4)$ | 1.624 | 0.169(0.751) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 5)$ | 1.170 | 0.903(0.394) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 6)$ | 1.566 | 6.179(0.012) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 7)$ | 2.189 | 4.542(0.009) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 8)$ | 1.001 | 1.543(0.214) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 9)$ | 0.000 | 0.039(0.998) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 10)$ | 1.000 | 7.318(0.007) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 11)$ | 2.421 | 1.187(0.237) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 12)$ | 1.561 | 6.455(0.010) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 13)$ | 1.665 | 0.740(0.582) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 14)$ | 1.000 | 8.958(0.003) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 15)$ | 1.000 | 7.463(0.006) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 16)$ | 1.000 | 0.044(0.834) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 17)$ | 1.000 | 9.268(0.002) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 18)$ | 3.325 | 4.134(0.003) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 19)$ | 1.000 | 0.157(0.692) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 20)$ | 1.000 | 1.165(0.281) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 21)$ | 1.000 | 0.272(0.602) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 22)$ | 1.000 | 0.412(0.521) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 23)$ | 1.000 | 0.046(0.831) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 24)$ | 1.000 | 9.516(0.002) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 25)$ | 2.483 | 4.225(0.010) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 26)$ | 2.573 | 5.350(0.002) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 27)$ | 1.333 | 5.826(0.007) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 28)$ | 1.000 | 0.561(0.454) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 29)$ | 1.000 | 8.652(0.003) | - | |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 30)$ | 1.000 | 1.630(0.202) | - | |

**Table 2** continued

| | | | |
|---|---|---|---|
| $f_3(x_{ij} - x_{.j})(Cluster_j = 31)$ | 1.795 | 5.495(0.014) | - |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 32)$ | 1.000 | 1.438(0.231) | - |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 33)$ | 1.001 | 10.222(0.001) | - |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 34)$ | 1.000 | 0.377(0.540) | - |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 35)$ | 1.526 | 6.870(0.008) | - |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 36)$ | 1.486 | 3.967(0.021) | - |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 37)$ | 1.000 | 0.132(0.716) | - |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 38)$ | 2.488 | 5.930(0.002) | - |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 39)$ | 1.000 | 1.606(0.205) | - |
| $f_3(x_{ij} - x_{.j})(Cluster_j = 40)$ | 1.000 | 0.854(0.356) | - |

*Note* - indicates a parameter or a smooth function which was not considered; Significance for fixed effects in bold based on $t$-test at $\alpha = .05$; $T_r$ is a test statistic for a smooth function and $Ref.edf$ is a reference $edf$ used in computing test statistic and the $p$-values

variance and covariance parameters using the `gamm` function in the `mgcv` package, which calls the `lme` function in the `nlme` package (Pinheiro et al., 2021). Table 2 presents results of MLM in comparison with GAMM. As described earlier, the level-2 TRT effects detected by MLM are $\gamma_{02} + \gamma_{03}x_{.j}$. Unlike GAMM, the level-2 TRT effects were not significant when detected with MLM. For the comparisons between GAMM and MLM with respect to model-data fit, the root mean squared error as a model-data fit index (RMSEI)[3] was calculated as a measure of differences between observed data and model predicted values. The RMSEI can be understood as the standard deviation of the part of the data that remains unexplained by a model; therefore, a lower value of RMSEI signals better model-data fit. The RMSEI was 5.815 for GAMM and 6.490 for MLM, which indicates that there is improvement of model-data fit with by-variable smooth functions as presented in Fig. 5. However, Akaike information criterion (AIC; Akaike, 1974) for GAMM is 5386.947 and AIC for MLM is 5383.156, which indicates that the model fit can be similar when taking model complexity (the number of parameters) into account.

## Simulation study

A simulation study was designed to answer the following three research questions: (a) are parameters and standard errors of the GAMM specification recovered well in a C-RCT design commonly found in intervention research?, (b) what are the consequences of modeling linear relationships between covariates and outcomes to detect TRT × COV effects in the presence of cluster-specific functional relationships?, and (c) can parameter recovery in GAMM be equally

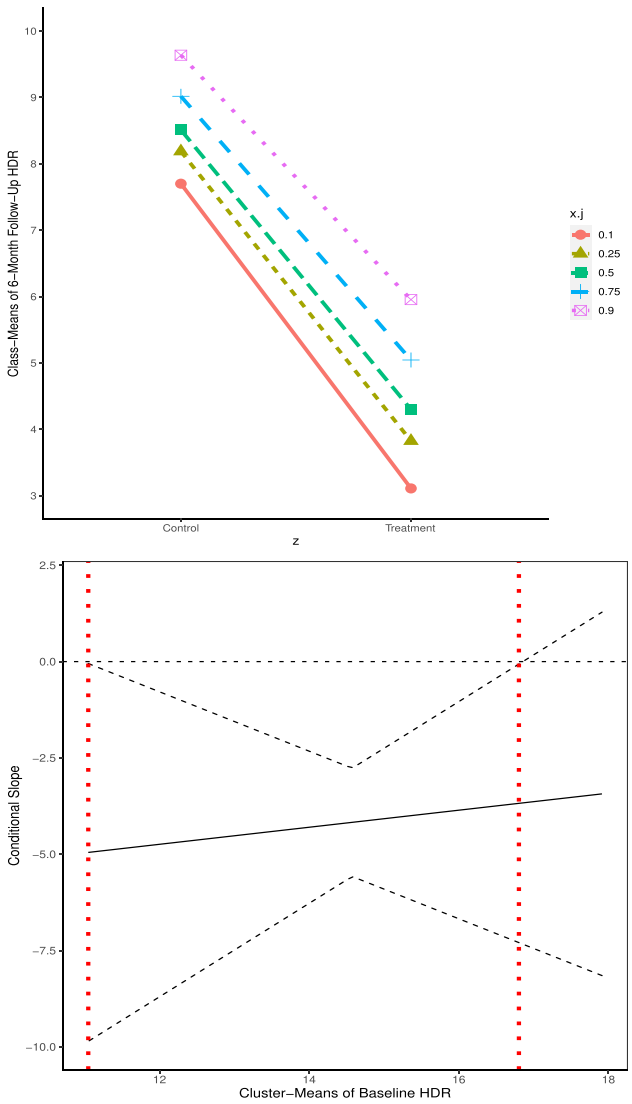


**Fig. 5** Empirical Study: Probing the interaction between a level-2 TRT and the level-2 part of COV for GAMM. Vertical lines in Fig. 5 (bottom) indicate windows of significant differences

[3] We use the abbreviation RMSEI to distinguish it from the root mean squared error of an estimator (RMSE) in the simulation study.

as good as in MLM when there are varying linear relationships between covariates and outcomes across clusters?

## Simulation design

Three varying simulation conditions which affect accuracy of parameter estimates in multilevel modeling were selected (e.g., Geldhof et al., 2014): (a) the number of clusters ($J$), selected as 20, 40, and 80; (b) balanced cluster sizes ($n_j$), selected as 15 and 30; and (c) $ICC$ of outcomes, selected to be at $ICC = .05, .10$, or $.30$. The levels of these simulation conditions were chosen based on literature reviews on study designs in RCTs, which are reported in Cho et al. (2022).

For Research Questions (a) and (b), the data-generating model is the specified GAMM (Equation 2). Fixed parameters were generated as $\gamma_{00} = 2.942$ and $\gamma_{02} = -8.534$ as found in the empirical study (see Table 2). Increasing nonlinear functions were generated using Equation 3 with $K = 10-1$ (1 is for an identification constraint; $K = 10$ should be set in estimation) for 'true' smooth functions. The $K = 10-1$ was chosen in the simulation study as a default setting in the `mgcv` package. Intervention studies are often designed with the intention of improving outcomes at the lower end of a covariate (e.g., improvement of learning for students with low-achieving levels). To mimic this pattern in a C-RCT design, for smooth functions at level 2 ($f_1(x_{.j})z_j$), data were generated such that the lower end of the covariate corresponded with larger nonlinear treatment effects, whereas smaller treatment effects were generated at the other ranges of a covariate. This data structure simulates an intervention that is more effective for individuals at lower levels of the covariate. For smooth functions at level 1 ($f_2(x_{ij} - x_{.j})z_j$), small differences in the two smooth functions from control and treatment groups were generated. For a cluster-specific smooth function ($f_3(x_{ij} - x_{.j})Cluster_j$)), nine basis coefficients (for $K = 10 - 1$ with an identification constraint) were generated from a uniform distribution to generate cluster-specific functional effects. For illustrative purposes, generated smooth functions were presented for one simulation condition ($J = 80, n_j = 30, ICC = 0.30$) in Appendix S.3. For Research Question (c), the data-generating model is the specified MLM (Equation 1). Estimates of fixed effects in MLM for the empirical study ($\widehat{\boldsymbol{\gamma}}$ reported in Table 2) were considered to be true parameters.

For both GAMM and MLM as data-generating models, $ICC$ was varied by manipulating the 'true' variances of random intercept and random errors. Specifically, given the error variance $\sigma^2 = 0.6$, the three levels of $\tau_{00}$ were 0.032, 0.067, and 0.257 which corresponded with $ICC$ =.05, .10, and .30, respectively. When MLM was the data-generating model, the slope variance $\tau_{11}$ was set as 0.1 and the intercept-slope covariance $\tau_{01}$ was set to be 0. The $x_{.j}$ and $x_{ij} - x_{.j}$ were drawn from standard normal distributions. For all replications within each simulation condition, the same COV ($x_{ij} - x_{.j}$ and $x_{.j}$) and generated by-variable smooth functions were used, whereas random effects were generated at each replication.

The simulation conditions regarding multilevel designs were fully crossed, yielding 18 (= 3 number of clusters × 2 cluster sizes × 3 $ICC$s) conditions. Five hundred replications were simulated for each condition. To answer Research Question (a), GAMM was fitted to the generated data sets under GAMM as a data-generating model. To answer Research Question (b), MLM was fitted to the generated data sets under GAMM as a data-generating model. For Research Question (c), MLM and GAMM were fitted to the generated data sets under MLM as a data-generating model. Thus, the total number of fitted models is 36,000 (18 multilevel designs × 500 replications × 2 models [GAMM and MLM] for GAMM as a data-generating model; and 18 multilevel designs × 500 replications × 2 models [GAMM and MLM] for MLM as a data-generating model).

## Evaluation measures and analysis

For the accuracy of estimates in the parametric part of GAMM ($\widehat{\gamma}_{00}$, $\widehat{\gamma}_{02}$, $\widehat{\tau}_{00}$ and $\widehat{\sigma}^2$) (Research Questions (a) and (c)), the bias[4] (calculated as $\sum_{rep=1}^{500}(\widehat{\gamma}_{02} - \gamma_{02})/500$ where $rep$ is a replication number as an example) and the root mean square error (RMSE; obtained as $\sqrt{\sum_{rep=1}^{500}(\widehat{\gamma}_{02} - \gamma_{02})^2/500}$ as an example) were calculated. For the accuracy of standard errors for the fixed effects ($\gamma_{00}$ and $\gamma_{02}$), the mean standard error of the estimates (M(SE)) across 500 replications was compared with the standard deviation of the estimates (SD) across 500 replications. A ratio of M(SE) to SD is reported. For the accuracy of fitted smooth functions in GAMM ($\tilde{f}_1(x_{.j})z_j$, $\tilde{f}_2(x_{ij} - x_{.j})z_j$, and $\tilde{f}_3(x_{ij} - x_{.j})Cluster_j$), the root mean square difference (RMD) between predicted smooth functions and true smooth functions was obtained at level 1 and level 2. It was assumed that smooth functions were specified with $K = 10$ because basis functions were generated for $K = 10$. The same $K = 10$ was used in all replications of each simulation condition. The RMD calculations at each level are presented in Table 3 (top).

To present the consequences of modeling random effects in the presence of cluster-specific functional effects (Research Question (b)) and to show the accuracy of estimates in the presence of cluster-specific linear effects (Research Question (c)), the bias and RMSE were obtained based on MLM estimates ($\widehat{\gamma}_{00}$, $\widehat{\gamma}_{02}$, $\widehat{\tau}_{00}$, and $\widehat{\sigma}^2$, which are comparable with GAMM estimates) and the ratio of M(SE) to SD was calcu-

[4] In this study, we did not consider relative bias due to its scaling artifact. When true parameters are close to 0 (e.g., $\tau_{00}$ =0.032, 0.067, and 0.257 and basis coefficients for smooth functions in the simulation study), relative bias will be inflated by the true parameters close to 0 in the denominator even for small differences from the true parameters.

**Table 3** Simulation study: Root mean squared differences (RMD) between predicted values and true values under GAMM as a data-generating model (top) and MLM as a data-generating model (bottom)

| Fitting model | Level | Group | RMD |
|---|---|---|---|
| GAMM | Level 1 | $z_j$ | $\sqrt{\sum_{j=1}^{J}\sum_{j=1}^{n_j} d_{ij.G1}^2/(Jn_j)}$ |
| | | | where $d_{ij.G1} = \{\tilde{f}_2(x_{ij} - x_{.j})(z_j = 0) - f_2(x_{ij} - x_{.j})(z_j = 0)\} + \{\tilde{f}_2(x_{ij} - x_{.j})(z_j = 1) - f_2(x_{ij} - x_{.j})(z_j = 1)\}$ |
| | Level 2 | $z_j$ | $\sqrt{\sum_{j=1}^{J} d_{j.G2}/J}$ |
| | | | where $d_{j.G2} = \{\tilde{f}_1(x_{.j})(z_j = 0) - f_1(x_{.j})(z_j = 0)\} + \{\tilde{f}_1(x_{.j})(z_j = 1) - f_1(x_{.j})(z_j = 1)\}$ |
| | | Cluster | $\sqrt{\sum_{j=1}^{J} d_{j.G3}/J}$ |
| | | | where $d_{j.G3} = \tilde{f}_3(x_{ij} - x_{.j})(Cluster = j) - f_3(x_{ij} - x_{.j})(Cluster = j)$ |
| MLM | Level 1 | $z_j$ | $\sqrt{\sum_{j=1}^{J}\sum_{j=1}^{n_j} d_{ij.M1}^2/(Jn_j)}$ |
| | | | where $d_{ij.M1} = \{\widehat{\gamma}_{10}(x_{ij} - x_{.j}) - f_2(x_{ij} - x_{.j})(z_j = 0)\} + \{\widehat{\gamma}_{11}(x_{ij} - x_{.j})z_j - f_2(x_{ij} - x_{.j})(z_j = 1)\}$ |
| | Level 2 | $z_j$ | $\sqrt{\sum_{j=1}^{J} d_{j.M2}/J}$ |
| | | | where $d_{j.M2} = \{\widehat{\gamma}_{01}x_{.j} - f_1(x_{.j})(z_j = 0)\} + \{\widehat{\gamma}_{03}x_{.j}z_j - f_1(x_{.j})(z_j = 1)\}$ |
| | | Cluster | $\sqrt{\sum_{j=1}^{J} d_{j.M3}/J}$ |
| | | | where $d_{j.M3} = (x_{ij} - x_{.j})\tilde{u}_{1j} - f_3(x_{ij} - x_{.j})(Cluster = j)$ |
| GAMM | Level 1 | $z_j$ | $\sqrt{\sum_{j=1}^{J}\sum_{j=1}^{n_j} d_{ij.G1}^2/(Jn_j)}$ |
| | | | where $d_{ij.G1} = \{\tilde{f}_2(x_{ij} - x_{.j})(z_j = 0) - \gamma_{10}(x_{ij} - x_{.j})\} + \{\tilde{f}_2(x_{ij} - x_{.j})(z_j = 1) - \gamma_{11}(x_{ij} - x_{.j})z_j\}$ |
| | Level 2 | $z_j$ | $\sqrt{\sum_{j=1}^{J} d_{j.G2}/J}$ |
| | | | where $d_{j.G2} = \{\tilde{f}_1(x_{.j})(z_j = 0) - \gamma_{01}x_{.j}\} + \{\tilde{f}_1(x_{.j})(z_j = 1) - \gamma_{03}x_{.j}z_j\}$ |
| | | Cluster | $\sqrt{\sum_{j=1}^{J} d_{j.G3}/J}$ |
| | | | where $d_{j.G3} = \tilde{f}_3(x_{ij} - x_{.j})(Cluster = j) - (x_{ij} - x_{.j})u_{1j}$ |
| MLM | Level 1 | $z_j$ | $\sqrt{\sum_{j=1}^{J}\sum_{j=1}^{n_j} d_{ij.M1}^2/(Jn_j)}$ |
| | | | where $d_{ij.M1} = \{\widehat{\gamma}_{10}(x_{ij} - x_{.j}) - \gamma_{10}(x_{ij} - x_{.j})\} + \{\widehat{\gamma}_{11}(x_{ij} - x_{.j})z_j - \gamma_{11}(x_{ij} - x_{.j})z_j\}$ |
| | Level 2 | $z_j$ | $\sqrt{\sum_{j=1}^{J} d_{j.M2}/J}$ |
| | | | where $d_{j.M2} = (\widehat{\gamma}_{01}x_{.j} - \gamma_{01}x_{.j}) + (\widehat{\gamma}_{03}x_{.j}z_j - \gamma_{03}x_{.j}z_j)$ |
| | | Cluster | $\sqrt{\sum_{j=1}^{J} d_{j.M3}/J}$ |
| | | | where $d_{j.M3} = (x_{ij} - x_{.j})\tilde{u}_{1j} - (x_{ij} - x_{.j})u_{1j}$ |

lated with the fixed effects of MLM ($\widehat{\gamma}_{00}$, $\widehat{\gamma}_{02}$). In addition, the RMD of the predicted values of the linear TRT × COV effects and the random slope under MLM was calculated (see Table 3 [bottom] for the RMD calculations at level 1 and level 2).

The GAMM estimates were obtained using the `gam` function in the `mgcv` package, as described in the method section. As in the empirical study, REML estimation was implemented to obtain unbiased estimates of variance and covariance parameters in MLM using the `gamm` function.

## Simulation results

There were no convergence problems in any simulation conditions for MLM and GAMM. Results indicated that $K = 10$ was adequate. For all smooth functions in all simulation conditions, the $k$-index was close to 1. The model with $K = 10$ was selected among models with differing $K$ values ($K = 6, 8, 10, 12, 14$) based on the correctedAIC. When GAMM is a data-generating model, RMSEIs of GAMM were smaller than those of MLM in all replications for all conditions, which indicates that RMSEI can be used to evaluate model-data fit in the presence of functional covariate effects.

Below, simulation results are summarized by research question. Table 4 (for Research Questions (a) and (b)) and Table 5 (Research Question (c)) show bias and RMSE of $\widehat{\gamma}_{00}$, $\widehat{\gamma}_{02}$, $\widehat{\tau}_{00}$ and $\widehat{\sigma}^2$, which are comparable parameter estimates between MLM and GLMM; ratio for standard error evaluation of $\widehat{\gamma}_{00}$ and $\widehat{\gamma}_{02}$; and RMD for predictions. In the tables, the average results are presented by the levels of simulation conditions to interpret results by main effects of the conditions. Results of all simulation results (disaggregated results) are presented in the figures of Appendix S.4.

### Research question (a)

As presented under the GAMM columns in Table 4 (top), the bias of $\widehat{\gamma}_{00}$, $\widehat{\gamma}_{02}$, $\widehat{\tau}_{00}$, and $\widehat{\sigma}^2$ was close to 0 (ranging from −0.072 to 0.068) across the 18 simulation conditions. Except for RMSE of $\widehat{\sigma}^2$, clear patterns regarding simulation conditions arose, such that bias and RMSE of these estimates decreased as the number of clusters ($J$) and cluster size ($n_j$) increased. The ratio of M(SE) to SD for $\widehat{\gamma}_{00}$ and $\widehat{\gamma}_{02}$ ranged from 0.993 to 1.029 across 18 simulation conditions, which indicates that the estimated SE is approximately correct.

For the three kinds of by-variable smooth functions, the mean RMD across 500 replications ranged from 0.002 to 0.365 across the 18 simulation conditions (presented in the figures of Appendix S4.4), which suggests that the predicted smooth functions are close to the true smooth functions. As shown in Table 4 (bottom), the RMD for $f_1(x_{.j})z_j$ and $f_2(x_{ij} - x_{.j})z_j$ decreased with increasing number of clusters ($J$) and cluster size ($n_j$). Regarding levels of $ICC$, the RMD decreased with decreasing $ICC$ for $f_2(x_{ij} - x_{.j})z_j$ while there was no clear pattern for $f_1(x_{.j})z_j$. For $f_3(x_{ij} - x_{.j})Cluster_j$, the RMD increased as the number of clusters ($J$) and cluster size ($n_j$) increased. This pattern may be due to the fact that there are more observations for model-data fit predicted by $f_3(x_{ij} - x_{.j})Cluster_j$ for each cluster with increasing the number of clusters ($J$) and cluster size ($n_j$). In addition, the RMD decreased as $ICC$ increased, indicating that observations can be predicted better by $f_3(x_{ij} - x_{.j})Cluster_j$ when there is greater between-cluster variability. To conclude, parameters of GAMM were recovered well and functional covariate effects were predicted well by the by-variable smooth functions in the considered multilevel designs.

### Research question (b)

Except for the bias of $\widehat{\gamma}_{00}$ in conditions with a smaller number of clusters ($J$) and smaller cluster size ($n_j$) ($J = 30$ or $J = 70$ with $n_j = 15$), a larger bias, RMSE, and ratio of M(SE) to SD were found in MLM estimates than in GAMM estimates. The following patterns were found in MLM as a misspecified model, as shown in Table 4 (top). First, bias tended to be smaller as the number of clusters ($J$) and cluster size ($n_j$) decreased, except for $\widehat{\tau}_{00}$ regarding the number of clusters ($J$). Second, for $\widehat{\gamma}_{00}$, $\widehat{\gamma}_{02}$, and $\widehat{\tau}_{00}$ in MLM, RMSE decreased as the number of clusters ($J$) and cluster size ($n_j$) increased. Third, the ratio of M(SE) to SD in MLM ranged from 1.027 to 2.120 for $\widehat{\gamma}_{00}$ and from 1.170 to 2.038 for $\widehat{\gamma}_{02}$ (presented in the figures of Appendix S4.3). These results suggest that standard errors of these fixed effects are overestimated. The degree of overestimation increased as the number of clusters ($J$), cluster size ($n_j$), and $ICC$ decreased.

For predictions, MLM had a larger RMD than GAMM in all 18 simulation conditions, as shown in Table 4 (bottom). For $f_1(x_{.j})z_j$ and $f_3(x_{ij} - x_{.j})Cluster_j$, the RMD in MLM increased as the number of clusters ($J$) and cluster size ($n_j$) increased. However, the opposite pattern was found in MLM for $f_2(x_{ij} - x_{.j})z_j$. Furthermore, the RMD for $f_1(x_{.j})z_j$ decreased with increasing $ICC$, whereas the RMD for $f_2(x_{ij} - x_{.j})z_j$ increased with increasing $ICC$. To summarize, these results indicate that misspecifying the functional form of level-specific TRT × COV and cluster-specific covariate effects leads to biased estimates of MLM parameters (including a focal parameter $\gamma_{02}$) and standard errors of fixed effects in MLM.

### Research question (c)

The following patterns, presented in Table 5, were observed when comparing the results for MLM and GAMM when MLM was the data-generating model. First, overall, larger

**Table 4** Simulation Study: Results for Fixed and Random Effects (top) and RMD (bottom) of GAMM ('True' Model) and MLM (Misspecified Model) under GAMM as a Data-Generating Model

| Parameters | Conditions | GAMM | | | MLM | | |
|---|---|---|---|---|---|---|---|
| | | Bias | RMSE | Ratio | Bias | RMSE | Ratio |
| Fixed effects | | | | | | | |
| $\gamma_{00}$ | $J = 20$ | −0.035 | 0.122 | 0.999 | 0.008 | 0.157 | 1.553 |
| | $J = 40$ | −0.025 | 0.089 | 1.004 | 0.012 | 0.143 | 1.462 |
| | $J = 80$ | −0.008 | 0.058 | 1.006 | 0.032 | 0.066 | 1.430 |
| | $n_j = 15$ | −0.028 | 0.091 | 1.003 | 0.007 | 0.151 | 1.644 |
| | $n_j = 30$ | −0.017 | 0.088 | 1.003 | 0.027 | 0.094 | 1.319 |
| | $ICC = 0.05$ | −0.011 | 0.050 | 1.000 | 0.028 | 0.117 | 1.676 |
| | $ICC = 0.1$ | −0.029 | 0.086 | 1.000 | 0.017 | 0.117 | 1.545 |
| | $ICC = 0.3$ | −0.027 | 0.132 | 1.009 | 0.006 | 0.133 | 1.224 |
| $\gamma_{02}$ | $J = 20$ | −0.011 | 0.161 | 0.997 | 0.057 | 0.205 | 1.551 |
| | $J = 40$ | −0.010 | 0.119 | 0.998 | 0.036 | 0.149 | 1.379 |
| | $J = 800$ | −0.009 | 0.095 | 1.008 | −0.055 | 0.107 | 1.297 |
| | $n_j = 15$ | −0.013 | 0.127 | 1.002 | 0.072 | 0.177 | 1.459 |
| | $n_j = 30$ | −0.007 | 0.124 | 1.001 | −0.047 | 0.130 | 1.359 |
| | $ICC = 0.05$ | 0.010 | 0.104 | 1.004 | −0.015 | 0.106 | 1.515 |
| | $ICC = 0.1$ | −0.024 | 0.132 | 0.995 | 0.022 | 0.157 | 1.463 |
| | $ICC = 0.3$ | −0.016 | 0.139 | 1.004 | 0.031 | 0.198 | 1.249 |
| Random effects | | | | | | | |
| $\tau_{00}$ | $J = 20$ | 0.017 | 0.068 | | 0.107 | 0.184 | |
| | $J = 40$ | 0.002 | 0.033 | | 0.106 | 0.123 | |
| | $J = 80$ | −0.001 | 0.028 | | 0.102 | 0.113 | |
| | $n_j = 15$ | 0.009 | 0.047 | | 0.104 | 0.146 | |
| | $n_j = 30$ | 0.003 | 0.039 | | 0.107 | 0.134 | |
| | $ICC = 0.05$ | 0.013 | 0.029 | | 0.142 | 0.178 | |
| | $ICC = 0.1$ | 0.008 | 0.027 | | 0.088 | 0.122 | |
| | $ICC = 0.3$ | −0.003 | 0.073 | | 0.085 | 0.120 | |
| $\sigma^2$ | $J = 20$ | 0.043 | 0.067 | | 0.190 | 0.218 | |
| | $J = 40$ | 0.035 | 0.050 | | 0.213 | 0.248 | |
| | $J = 80$ | 0.027 | 0.049 | | 0.235 | 0.261 | |
| | $n_j = 15$ | 0.040 | 0.054 | | 0.204 | 0.233 | |
| | $n_j = 30$ | 0.030 | 0.056 | | 0.222 | 0.252 | |
| | $ICC = 0.05$ | 0.043 | 0.064 | | 0.267 | 0.280 | |
| | $ICC = 0.1$ | 0.029 | 0.051 | | 0.188 | 0.224 | |
| | $ICC = 0.3$ | 0.033 | 0.051 | | 0.183 | 0.223 | |

| Predictions | Conditions | GAMM | MLM |
|---|---|---|---|
| $f_1(x_{.j})z_j$ | $J = 20$ | 0.022 | 0.084 |
| | $J = 40$ | 0.021 | 0.116 |
| | $J = 80$ | 0.016 | 0.129 |
| | $n_j = 15$ | 0.023 | 0.096 |
| | $n_j = 30$ | 0.017 | 0.123 |
| | $ICC = 0.05$ | 0.021 | 0.122 |
| | $ICC = 0.1$ | 0.018 | 0.115 |
| | $ICC = 0.3$ | 0.020 | 0.093 |
| $f_2(x_{ij} - x_{.j})z_j$ | $J = 30$ | 0.028 | 0.109 |
| | $J = 70$ | 0.015 | 0.098 |

**Table 4** continued

| Predictions | Conditions | GAMM | MLM |
|---|---|---|---|
| | $J = 200$ | 0.009 | 0.092 |
| | $n_j = 15$ | 0.019 | 0.108 |
| | $n_j = 30$ | 0.015 | 0.092 |
| | $ICC = 0.05$ | 0.011 | 0.073 |
| | $ICC = 0.1$ | 0.011 | 0.094 |
| | $ICC = 0.3$ | 0.030 | 0.132 |
| $f_3(x_{ij} - x_{.j})Cluster_j$ | $J = 20$ | 0.220 | 0.441 |
| | $J = 40$ | 0.256 | 0.483 |
| | $J = 80$ | 0.266 | 0.500 |
| | $n_j = 15$ | 0.243 | 0.461 |
| | $n_j = 30$ | 0.252 | 0.489 |
| | $ICC = 0.05$ | 0.290 | 0.488 |
| | $ICC = 0.1$ | 0.235 | 0.464 |
| | $ICC = 0.3$ | 0.217 | 0.472 |

*Note.* Ratio of M(SE) to SD was considered for fixed effects; $J$ indicates the number of clusters; $n_j$ indicates a cluster size; $ICC$ indicates an intraclass correlation coefficient; for each smooth function, RMD indicates the mean RMD across 500 replications

bias and RMSE of $\widehat{\gamma}_{00}$ and $\widehat{\gamma}_{02}$ were found in GAMM than in MLM. This could be due to the fact that there were more parameters to be estimated in GAMM than in MLM when the other fixed effects ($\gamma_{10}$, $\gamma_{01}$, $\gamma_{03}$, $\gamma_{11}$) and the random slope ($u_{1j}$) in MLM were replaced with smooth functions in GAMM (see Table 1). For example, for modeling variability in slope, the variance of the random slope was estimated in MLM, while the smoothing parameter and the 9 ($K = 10-1$) basis coefficients were estimated for each cluster. Second, the ratios of estimated standard errors to the standard deviation for $\widehat{\gamma}_{00}$ and $\widehat{\gamma}_{02}$ in MLM and GAMM were close to 1 (ranging from 0.982 to 1.109 in MLM and from 0.957 to 0.997 in GAMM). Although these results suggest that the estimated standard errors are reasonably accurate (less than 10.9% bias in MLM and less than 4.3% bias in GAMM), there was a tendency for standard errors to be overestimated in MLM and underestimated in GAMM. Third, differences in bias and RMSE for $\widehat{\tau}_{00}$ and $\widehat{\sigma}^2$ were relatively small in MLM and GAMM except the condition of $J = 20$ and $ICC = 0.3$ for bias: average differences in bias and RMSE between MLM and GAMM across the 18 simulation conditions were 0.011 and −0.003 respectively for $\widehat{\tau}_{00}$, and were 0.001 and 0 respectively for $\widehat{\sigma}^2$. Fourth, RMD for $(x_{ij}-x_{.j})z_j$ was similar between MLM and GAMM (average difference across the 18 simulation conditions=−0.006). However, different patterns in RMD for $x_{.j}z_j$ and $(x_{ij} - x_{.j})u_{1j}$ were found in MLM and GAMM. For $x_{.j}z_j$, a larger RMD was found in GAMM than in MLM (average difference across the 18 simulation conditions=−0.051). For $(x_{ij} - x_{.j})u_{1j}$, a larger RMD was found in MLM than in GAMM (average difference across 18 simulation conditions=0.239).

Regarding simulation conditions, comparable patterns in results were found in MLM and GAMM except for a few cases listed below. First, for all parameters and predictions, bias and RMSE decreased for MLM and GAMM as the number of clusters $J$ and cluster sizes $n_j$ increased, with a few exceptions.[5] Second, for all parameters and predictions overall, bias and RMSE decreased with decreasing the $ICC$s, except for bias for $\widehat{\gamma}_{00}$, $\widehat{\gamma}_{02}$, and $\widehat{\sigma}^2$ in GAMM. Third, the ratios for $\widehat{\gamma}_{00}$ and $\widehat{\gamma}_{02}$ were close to exactly 1, particularly as the number of clusters $J$ and cluster sizes $n_j$ increased for MLM and GAMM, except for $\widehat{\gamma}_{02}$ with respect to $J$.

To summarize, results from MLM and GAMM are comparable in the presence of varying linear relationships between covariates and outcomes across clusters in most conditions, except for bias and RMSE of $\widehat{\gamma}_{00}$ and $\widehat{\gamma}_{02}$ and RMD of $x_{.j}z_j$ and $(x_{ij} - x_{.j})u_{1j}$. Overall, the accuracy of the fixed effects ($\widehat{\gamma}_{00}$ and $\widehat{\gamma}_{02}$) and the prediction of $x_{.j}z_j$ were better in MLM than in GAMM. For $(x_{ij} - x_{.j})u_{1j}$, the accuracy of predictions was better when it was predicted with a smooth function in GAMM than with a random slope in MLM.

[5] Exceptional cases include bias of $\widehat{\gamma}_{00}$ and $\widehat{\gamma}_{02}$ with respect to $n_j$ and $\widehat{\sigma}^2$ with respect to $J$ in MLM; and bias of $\widehat{\gamma}_{00}$ with respect to $J$ and $n_j$, $\widehat{\gamma}_{02}$ with respect to $J$, and $\widehat{\sigma}^2$ with respect to $n_j$ in GAMM.

**Table 5** Simulation study: Results for fixed and random effects (top) and RMD (bottom) of MLM ('True' Model) and GAMM (Alternative Model) under MLM as a data-generating model

| Parameters | Conditions | MLM | | | GAMM | | |
|---|---|---|---|---|---|---|---|
| | | Bias | RMSE | Ratio | Bias | RMSE | Ratio |
| Fixed effects | | | | | | | |
| $\gamma_{00}$ | $J = 20$ | 0.003 | 0.106 | 1.109 | 0.001 | 0.139 | 0.957 |
| | $J = 40$ | −0.002 | 0.085 | 1.015 | −0.008 | 0.115 | 0.981 |
| | $J = 80$ | 0.000 | 0.061 | 0.995 | −0.015 | 0.075 | 0.994 |
| | $n_j = 15$ | 0.001 | 0.087 | 1.058 | −0.001 | 0.118 | 0.958 |
| | $n_j = 30$ | 0.000 | 0.081 | 1.022 | −0.013 | 0.101 | 0.997 |
| | $ICC = 0.05$ | −0.001 | 0.058 | 1.045 | −0.053 | 0.084 | 0.984 |
| | $ICC = 0.1$ | −0.002 | 0.065 | 1.052 | 0.000 | 0.102 | 0.963 |
| | $ICC = 0.3$ | 0.004 | 0.129 | 1.023 | 0.031 | 0.143 | 0.986 |
| $\gamma_{02}$ | $J = 20$ | −0.002 | 0.153 | 1.108 | −0.001 | 0.181 | 0.960 |
| | $J = 40$ | 0.002 | 0.122 | 0.997 | 0.004 | 0.131 | 0.987 |
| | $J = 80$ | 0.001 | 0.086 | 0.982 | −0.017 | 0.093 | 0.988 |
| | $n_j = 15$ | 0.001 | 0.128 | 1.036 | −0.010 | 0.142 | 0.969 |
| | $n_j = 30$ | 0.000 | 0.113 | 1.022 | 0.001 | 0.128 | 0.988 |
| | $ICC = 0.05$ | 0.005 | 0.085 | 1.015 | −0.032 | 0.101 | 0.979 |
| | $ICC = 0.1$ | −0.003 | 0.098 | 1.047 | 0.007 | 0.118 | 0.972 |
| | $ICC = 0.3$ | 0.000 | 0.178 | 1.025 | 0.011 | 0.187 | 0.984 |
| Random effects | | | | | | | |
| $\tau_{00}$ | $J = 20$ | −0.001 | 0.059 | | −0.023 | 0.060 | |
| | $J = 40$ | 0.000 | 0.035 | | −0.007 | 0.044 | |
| | $J = 80$ | 0.005 | 0.029 | | 0.002 | 0.029 | |
| | $n_j = 15$ | 0.005 | 0.042 | | −0.009 | 0.041 | |
| | $n_j = 30$ | −0.003 | 0.041 | | −0.010 | 0.047 | |
| | $ICC = 0.05$ | 0.008 | 0.021 | | 0.003 | 0.020 | |
| | $ICC = 0.1$ | −0.008 | 0.028 | | −0.014 | 0.037 | |
| | $ICC = 0.3$ | 0.004 | 0.074 | | −0.016 | 0.076 | |
| $\sigma^2$ | $J = 20$ | 0.000 | 0.044 | | −0.002 | 0.044 | |
| | $J = 40$ | −0.001 | 0.032 | | −0.002 | 0.031 | |
| | $J = 80$ | 0.001 | 0.022 | | 0.000 | 0.022 | |
| | $n_j = 15$ | 0.000 | 0.038 | | −0.002 | 0.039 | |
| | $n_j = 30$ | 0.000 | 0.027 | | −0.001 | 0.026 | |
| | $ICC = 0.05$ | −0.001 | 0.033 | | −0.002 | 0.033 | |
| | $ICC = 0.1$ | −0.002 | 0.031 | | −0.002 | 0.032 | |
| | $ICC = 0.3$ | 0.002 | 0.033 | | −0.001 | 0.033 | |
| Predictions | Conditions | MLM | GAMM | | | | |
| $x_{.j}z_j$ | $J = 20$ | 0.128 | 0.186 | | | | |
| | $J = 40$ | 0.089 | 0.146 | | | | |
| | $J = 80$ | 0.062 | 0.100 | | | | |
| | $n_j = 15$ | 0.101 | 0.152 | | | | |
| | $n_j = 30$ | 0.085 | 0.136 | | | | |
| | $ICC = 0.05$ | 0.067 | 0.110 | | | | |
| | $ICC = 0.1$ | 0.076 | 0.141 | | | | |
| | $ICC = 0.3$ | 0.136 | 0.180 | | | | |
| $(x_{ij} - x_{.j})z_j$ | $J = 20$ | 0.112 | 0.121 | | | | |
| | $J = 40$ | 0.080 | 0.086 | | | | |
| | $J = 80$ | 0.058 | 0.061 | | | | |
| | $n_j = 15$ | 0.085 | 0.094 | | | | |

**Table 5** continued

| Predictions | Conditions | MLM | GAMM |
|---|---|---|---|
| | $n_j = 30$ | 0.081 | 0.085 |
| | $ICC = 0.05$ | 0.082 | 0.087 |
| | $ICC = 0.1$ | 0.085 | 0.090 |
| | $ICC = 0.3$ | 0.084 | 0.091 |
| $(x_{ij} - x_{.j})u_{1j}$ | $J = 20$ | 0.419 | 0.177 |
| | $J = 40$ | 0.403 | 0.164 |
| | $J = 80$ | 0.394 | 0.157 |
| | $n_j = 15$ | 0.428 | 0.184 |
| | $n_j = 30$ | 0.383 | 0.148 |
| | $ICC = 0.05$ | 0.402 | 0.164 |
| | $ICC = 0.1$ | 0.407 | 0.167 |
| | $ICC = 0.3$ | 0.407 | 0.167 |

*Note.* Ratio of M(SE) to SD was considered for fixed effects; $J$ indicates the number of clusters; $n_j$ indicates a cluster size; $ICC$ indicates an intraclass correlation coefficient; for each term of predictions, RMD indicates the mean RMD across 500 replications

## Summary and discussion

Examining variability in treatment effects is important to better understand for whom, and under what conditions, interventions will be most effective (e.g., Spybrook et al., 2016). The effect of TRT × COV can be detected to understand such variability. This paper presented GAMM specifications and their parameter estimation methods using the freely available `R` package `mgcv` to obtain unbiased estimates of TRT × COV when cluster-specific functional covariate effects are observed graphically. In the presence of cluster-specific functional covariate effects, we showed via a simulation study (Research Question (a)) that parameters of the GAMM specifications were recovered satisfactorily in most multilevel designs from the C-RCT.

As illustrated in the empirical study, GAMM specifications can be applied when cluster-specific functional covariate effects are observed in the cluster-specific scatter plots (e.g., Fig. 2), and then model-data fit can be compared between GAMM and MLM using RMSEI. As shown in our simulation study (for Research Question (b)), TRT × COV effects are biased in the presence of cluster-specific functional covariate effects. Thus, we recommend conducting statistical inference on TRT × COV in GAMM rather than in MLM when cluster-specific functional covariate effects are observed in the cluster-specific scatter plots and there is improvement in model-data fit by using by-variable smooth functions in GAMM. By applying GAMM specifications, a primary benefit is to have better estimates of TRT × COV effects when cluster-specific functional covariate effects exist. As a supplementary benefit, as shown in the empirical study, data can be better predicted with GAMM than MLM (e.g., community id=16, 18, 25, 26, 36, and 38 in Fig. 4). These cluster-specific predictions can be informative when a researcher is interested in understanding the effectiveness of intervention by clusters.

In addition, the simulation study (for Research Question (c)) evaluated whether GAMM can be an alternative model to MLM in the presence of varying linear relationships between covariates and outcomes across clusters. Results showed that the relationships were better predicted with a by-variable smooth function ($f_3(x_{ij} - x_{.j})Cluster_j$) in GAMM than with random effects ($(x_{ij} - x_{.j})u_{1j}$) in MLM, whereas the fixed effects ($\widehat{\gamma}_{00}$ and $\widehat{\gamma}_{02}$) were more accurately estimated in most conditions and $x_{.j}z_j$ was more accurately predicted under MLM than under GAMM. As illustrated in the empirical study, statistical significance for the by-variable smooth function can be tested in GAMM. However, one limitation of estimating the relationships with the by-variable smooth function in GAMM is that its covariance with a random intercept ($u_{0j}$) cannot be estimated as it can in MLM. When the covariance is of interest, GAMM cannot be an alternative model to MLM when the linear relationship is observed. Nevertheless, as illustrated in the empirical study, GAMM can be considered when cluster-specific functional covariate effects are observed in the cluster-specific scatter plots (e.g., Fig. 2). Thus, GAMM may not be a good alternative to MLM in practice when linear relationships are observed for all clusters in the cluster-specific scatter plots.

When a functional covariate effect is modeled using by-variable smooth functions by groups and clusters in the presence of a functional relationship between covariates and outcomes, the bias-variance trade-off can be of concern. By-variable smooth functions have more parameters (basis coefficients and penalty parameters) to be estimated than a global smooth function. As described earlier, the bias-

variance trade-off is controlled by the penalty matrix in GAMM (i.e., larger penalties correspond to smaller variance, whereas smaller penalties correspond to higher variance). In addition, having by-variable smooth functions will increase computational time as the number of clusters increases.[6] In this study, REML was used to estimate a variable-specific penalty parameter in the `mgcv` package. To reduce overall mean square errors and improve computational efficiency, shrinkage estimators (i.e., pooling information on smoothness between clusters) can be considered for the penalty parameters to estimate cluster-specific smooth functions using a hierarchical Bayes approach. Further research is needed to investigate the relative performance of REML and hierarchical Bayes approaches for the by-variable smooth functions.

As the first attempt to model cluster-specific smooth functions in detecting TRT × COV effects in C-RCT designs, the GAMM specifications were illustrated and evaluated via simulation studies for two-level nested designs. Further studies are needed to apply the GAMM with by-variable smooth functions to more complex multilevel designs than two-level nested designs, such as higher-level designs and/or cross-classified designs. In addition, simulation results presented in this study are limited to the simulation conditions, specific parameters, and generated by-variable smooth functions that we elected to include. To make generalizations, more extensive simulation studies are required with varying sets of parameters and by-variable smooth functions. As another limitation of the current study, the model-data fit measured with RMSEI was used when the GAMM was selected over the MLM in the empirical study. Relying on RMSEI can lead to selecting a model with overfitting. In the empirical study, there were small differences in AIC between GAMM and MLM. However, additional study is needed to evaluate common model selection methods accounting for the model-data fit and model complexity (e.g., model information criteria) in selecting a best-fitting model between GAMM and MLM. Furthermore, when newly specified GAMMs are presented to applied researchers, it is important to design a study to ensure sufficient power (e.g., .80) for detecting ATEs and variability in treatment effects. For example, in a C-RCT, it is important to have a large number of clusters for inferences about the ATE and to have a large number of clusters and large cluster size for inferences about TRT × COV (Raudenbush & Liu, 2000). A future study is needed to present a power formula for the model specifications in the current study.

Despite the extensive model specification work applicable for C-RCT designs in methodological journals, specifying and estimating cluster-specific functional relationships between covariates and outcomes may not be straightforward to substantive researchers to obtain unbiased statistical inference on TRT × COV when the cluster-specific functional relationships are observed as shown in the empirical study. The GAMM approach in the current study is likely to be of increasing interest to researchers when variability in treatment effects is explained using covariates.

[6] For a single replication of GAMM estimation under GAMM as a data-generating model in a simulation condition with $J = 80$, about an hour was required on a laptop computer with a 2.8 GHz Intel Core i7 CPU and 16 GB of RAM.

## Open science statement

The empirical data set is freely available from Baranov et al. (2020b) and the R code is provided in the appendix.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.3758/s13428-023-02138-w.

## References


1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705
2. Baranov, V., Bhalotra, S., Biroli, P., & Maselko, J. (2020). Maternal depression, women's empowerment, and parental investment: Evidence from a randomized controlled trial. *American Economic Review, 110*(3), 824–859. https://doi.org/10.1257/aer.20180511
3. Baranov, V., Bhalotra, S., Biroli, P., & Maselko, J. (2020b). Data and code for: Maternal depression, women's empowerment, and parental investment: Evidence from a randomized controlled trial. Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor], 2020-02-28. https://doi.org/10.3886/E111366V1
4. Bloom, H. S., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness, 10*(4), 877–902. https://doi.org/10.1080/19345747.2016.1271069
5. Cho, S.-J., Preacher, K. J., Yaremych, H. E., Naveiras, M., Fuchs, D., & Fuchs, L. S. (2022). Modeling multilevel nonlinear treatment-by-covariate interactions in cluster randomized controlled trials using a generalized additive mixed model. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12265
6. Fine, K. L., Suk, H. W., & Grimm, K. J. (2019). An examination of a functional mixed-effects modeling approach to the analysis of longitudinal data. *Multivariate Behavioral Research, 54*(4), 475–491. https://doi.org/10.1080/00273171.2018.1520626
7. Fuchs, D., Cho, E., Toste, J. R., Fuchs, L. S., Gilbert, J. K., McMaster, K. L., Svenson, E., & Thompson, A. (2021). A quasi-experimental evaluation of two versions of first-grade PALS: One with and one without repeated reading. *Exceptional Children, 87*(2), 141–162. https://doi.org/10.1177/0014402920921828
8. Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72–91. https://doi.org/10.1037/a0032138
9. Guo, W. (2002). Functional mixed effects models. *Biometrics, 58*, 121–128. https://doi.org/10.1111/j.0006-341X.2002.00121.x

10. Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (Vol. 3). Springer. https://doi.org/10.1007/978-3-319-19425-7
11. Kimeldorf, G. S., & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics, 41*(2), 495–502. https://doi.org/10.1214/aoms/1177697089
12. Lawrence, J. F., Francis, D., Paré-Blagoev, J., & Snow, C. E. (2017). The poor get richer: Heterogeneity in the efficacy of a school-level intervention for academic language. *Journal of Research on Educational Effectiveness, 10*(4), 767–793. https://doi.org/10.1080/19345747.2016.1237596
13. Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, 61*(2), 381–400. https://doi.org/10.1111/1467-9868.00183
14. Longford, N. T. (1993). *Random coefficient models*. Oxford University Press.
15. Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2021). nlme: Linear and nonlinear mixed effects models. R package version 3. 1–152. https://CRAN.R-project.org/package=nlme
16. Preacher, K. J., & Sterba, S. K. (2019). Aptitude-by-treatment interactions in research on educational interventions. *Exceptional Children, 85*(2), 248–264. https://doi.org/10.1177/0014402918802803
17. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
18. Ramsay, J. O., & Silverman, B. W. (2002). Applied functional data analysis: Methods and case studies. *Springer*. https://doi.org/10.1007/b98886
19. Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer. https://doi.org/10.1007/b98888
20. Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.
21. Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213. https://doi.org/10.1037/1082-989X.5.2.199
22. Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society, Series B, 47*(1), 1–21. https://doi.org/10.1111/j.2517-6161.1985.tb01327.x
23. Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two-and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics, 41*(6), 605–627. https://doi.org/10.3102/1076998616655442
24. Tipton, E., & Hedges, L. V. (2017). The role of the sample in estimating and explaining treatment effect heterogeneity. *Journal of Research on Educational Effectiveness, 10*(4), 903–906. https://doi.org/10.1080/19345747.2017.1364563
25. Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society: Series B (Methodological), 45*(1), 133–150. https://doi.org/10.1111/j.2517-6161.1983.tb01239.x
26. Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness, 10*(4), 843–876. https://doi.org/10.1080/19345747.2017.1300719
27. Wood, S. N. (2013). A simple test for random effects in regression models. *Biometrika, 100*, 1005–1010. https://doi.org/10.1093/biomet/ast038
28. Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9781315370279
29. Wood, S. N. (2021). Package 'mgcv.' Retrieved from https://cran.r-project.org/web/packages/mgcv/mgcv.pdf
30. Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association, 111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986