Modeling Multilevel Nonlinear Treatment-by-Covariate Interactions in Cluster Randomized Controlled Trials using a Generalized Additive Mixed Model

Sun-Joo Cho, Vanderbilt University
Kristopher J. Preacher, Vanderbilt University
Haley E. Yaremych, Vanderbilt University
Matthew Naveiras, Vanderbilt University
Douglas Fuchs, Vanderbilt University
Lynn S. Fuchs, Vanderbilt University

December 29, 2021

The authors' accepted version of the article currently in press
in *British Journal of Mathematical and Statistical Psychology*

Authors' note: The illustrative data collection was supported in part by Grant No. H324V980001 (Center on Accelerating Student Learning) from the Office of Special Education Programs, U.S. Department of Education. Nothing in this article necessarily reflects the positions or policies of the agency, and no endorsement by it should be inferred.
Order of Authors with Contributor Roles: **Cho** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft; **Preacher** Conceptualization, Methodology, Validation, Visualization, Writing – review & editing; **Yaremych** Validation, Writing – review & editing; **Naveiras** Validation, Writing – review & editing; **Fuchs** Funding acquisition, Resources; and **Fuchs** Funding acquisition, Resources.
Data Availability: Data sharing is not available. However, all code is available as supplementary material.

Abstract

A cluster randomized controlled trial (C-RCT) is common in educational intervention studies. Multilevel modeling (MLM) is a dominant analytic method to evaluate treatment effects in a C-RCT. In most MLM applications intended to detect an interaction effect, a single interaction effect (called a *conflated* effect) is considered instead of level-specific interaction effects in a multilevel design (called *unconflated multilevel interaction* effects), and the linear interaction effect is modeled. In this paper, we present a generalized additive mixed model (GAMM) that allows an unconflated multilevel interaction to be estimated without assuming a prespecified form of the interaction. `R` code is provided to estimate the model parameters using maximum likelihood estimation and to visualize the nonlinear treatment-by-covariate interaction. The usefulness of the model is illustrated using instructional intervention data from a C-RCT. Results of simulation studies showed that GAMM outperformed an alternative approach to recover an unconflated logistic multilevel interaction. In addition, the parameter recovery of GAMM was relatively satisfactory in multilevel designs found in educational intervention studies, except when the number of clusters, cluster sizes, and intraclass correlations were small. When modeling a linear multilevel treatment-by-covariate interaction in the presence of a nonlinear effect, biased estimates (such as overestimated standard errors and overestimated random effect variances) and incorrect predictions of the unconflated multilevel interaction were found.

Keywords: cluster randomized controlled trial, generalized additive mixed model, nonlinear interaction, multilevel interaction, smooth function, treatment effects

1. Introduction

Study Motivation

Intervention studies in education (concerning, e.g., curriculum, policy, or instructional programs) have increased over the past two decades. The evaluation setting is often a randomized controlled trial (RCT). One popular RCT design in education research is a cluster RCT (C-RCT) with control and treatment groups. In the C-RCT design, clusters (e.g., schools) are randomized to the control or treatment group, and the inferential goal is to test hypotheses related to treatment effects.

The effectiveness of the intervention measures whether a program, policy, or approach improves outcomes. In education, outcomes are commonly continuous variables (e.g., continuous scale scores of students' achievement). Multilevel modeling (MLM) is a dominant analytic method because it allows researchers to model complex patterns of variability within and across different levels of analysis (e.g., students in schools) to evaluate whether outcomes are improved due to the effects of covariates. Education researchers often report an average treatment effect (ATE) or an ATE that is conditional on a preintervention covariate as a moderator (MOD; e.g., pretest scores, demographic information) (hereafter denoted TRT $\times$ MOD, indicating that treatment effects [TRT] differ based on the values of MOD). MODs have an important role in understanding variability in treatment effects, and they are commonly used as covariates in MLM to explain variability in treatment effects. In multilevel settings commonly found in education, MODs can be assessed either at the individual level or at the cluster level. In addition, MODs can be categorical (e.g., grade levels, school types) or continuous (e.g., student pretest scores, teaching experience in years). In this study, we focus on continuous MODs assessed at the individual level and a categorical TRT at the cluster level in a C-RCT.

Current Issues

In multilevel designs, the TRT $\times$ MOD effect represents *multilevel interaction* or *moderation* (e.g., Preacher et al., 2006; Raudenbush & Bryk, 2002). In testing hypotheses of multilevel interaction, several conceptual and statistical problems have been discussed (Preacher et al., 2016).

Preacher et al. (2016) noted that problems occur because most applications testing multilevel interaction do not separate level-1 and level-2 effects into their orthogonal components (in a two-level design as an example) and instead combine them into a single coefficient (called *conflation*). As a conceptional problem, conflation results in insensitivity to theoretically meaningful ways that multilevel interaction can occur. As a statistical problem, conflation leads to estimates of a weighted average of within- and between-cluster effects in the presence of the level-specific interaction effects. As a solution, Preacher and Sterba (2019) suggested modeling level-specific interaction (TRT × MOD) for fixed effects by centering a covariate (e.g., pretest) at its cluster mean (called the *unconflated* solution).

When a continuous MOD (e.g., pretest scores) and a treatment variable (e.g., control vs. treatment groups) are considered in detecting TRT × MOD interaction, the linear effect of the TRT × MOD interaction is often modeled in MLM to detect interaction effects (Preacher & Sterba, 2019). When the treatment variable is dummy coded (control group=0; treatment group=1), the MOD effect is the linear effect of MOD in the control group and the (MOD + TRT × MOD) effect is the linear effect of MOD in the treatment group. As such, the expected difference between the control and the treatment condition is the marginal TRT effect + TRT × MOD effect for each value of MOD. If the relationship between an outcome and MOD is incorrectly assumed to be linear, estimates of treatment effects are expected to be inaccurate (Harrell, 2015). To model nonlinear TRT × MOD interaction, Preacher and Sterba (2019) presented a logistic function in MLM. The parametric logistic function may work well when there are floor and ceiling effects. As a more flexible approach, *smooth functions* can be used for MODs that are known to predict an outcome nonlinearly. To the best of our knowledge, smooth functions have not been applied to MLM in the context of detecting unconflated TRT × MOD interaction effects.

Study Purpose

The purpose of this study is to illustrate modeling of nonlinear multilevel TRT × MOD interaction with unconflated effects in intervention studies from C-RCT designs. With multilevel TRT × MOD interaction with unconflated effects, the TRT effect is estimated as a function of

MOD at each level. A modeling framework to detect the nonlinear multilevel interaction effect is a generalized additive mixed model (GAMM; Lin & Zhang, 1999; Wood, 2017) with an identity link function. A GAMM can be considered to be MLM (having fixed and random effects) with an identity link in which the linear predictor partly depends on some unknown smooth functions. The *nonlinear* multilevel TRT × MOD interaction using GAMM is illustrated using an instructional intervention data set in a C-RCT and it is compared with the *linear* multilevel TRT × MOD interaction from MLM. For parameter estimation, we utilize the `gamm` function in the `mgcv` package (Wood, 2019) in `R` (R Core Team, 2020) for maximum likelihood estimation. In the `gamm` function, smooth functions in GAMM are reformulated as random effects and parameters of GAMM are estimated as parameters of generalized linear mixed-effects models (GLMM) (Wood, 2019). In this study, the key derivations on the reformulation of smooth functions as random effects by Wood (2004; 2006; 2017, p. 239) in the statistics literature are illustrated for researchers in the social and behavioral sciences. Furthermore, the `R` code is provided to visualize nonlinear multilevel TRT × MOD based on results from the `gamm` function. In addition, the accuracy and precision of parameter estimates is evaluated and the consequences of modeling *linear* multilevel TRT × MOD are presented in the presence of nonlinear multilevel TRT × MOD via simulation.

The remainder of this paper is organized as follows. In Section 2, we present the GAMM specification, provide the estimation method using `R`, and describe model checking and testing. In Section 3, the model is illustrated using an empirical data set. In Section 4, the design of simulation studies and their results are presented. In Section 5, we end with a summary and a discussion.

2. Methods

In this section, GAMM is specified with a comparison to MLM, and its parameter estimation method in the `gamm` function is described. In addition, testing for nonlinear TRT × MOD interaction is explained.

GAMM

A GAMM with an identity link and univariate smooth functions is written as

$$\mu = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}\mathbf{u} + \sum_{h=1}^{H} f_h(x_h), \mathbf{u} \sim MN(\mathbf{0}, \Sigma), y \sim N(\mu, \sigma^2). \quad (1)$$

where h is an index for the smooth function ($h = 1, \ldots, H$); y is an outcome variable; $\mathbf{X}$ is a design matrix for fixed effects; $\mathbf{Z}$ is a design matrix for random effects; $\boldsymbol{\gamma}$ is the vector of fixed parameters; $\mathbf{u}$ is the vector of random parameters; f_h is the univariate smooth function for covariate x_h; Σ is a covariance matrix of the random parameters in a multivariate normal (MN) distribution; σ^2 is an error variance; and $N(\mu, \sigma^2)$ indicates a normal (N) distribution with mean μ and an error variance σ^2. Here, one can see that the GAMM is a GLMM in which the linear predictor partly depends linearly on some unknown smooth functions (f_h).

A general form of GAMM with an identity link (Equation 1) can be presented using an MLM specification with a smooth function for a nonlinear interaction of categorical TRT $\times$ continuous MOD. To illustrate MLM specifications as GAMM in testing a nonlinear multilevel interaction, a level-1 MOD (x_{ij}; moderator) and a level-2 TRT (z_j; focal covariate) in a C-RCT are considered for a level-1 outcome (y_{ij}) in a two-level nested design in which an individual i is nested within a cluster j. For an unconflated solution for a nonlinear multilevel interaction of categorical TRT $\times$ continuous MOD, a level-1 MOD x_{ij} can be decomposed into uncorrelated level-1 and level-2 components by subtracting the cluster average $x_{.j}$ from x_{ij} (i.e., $x_{ij} - x_{.j}$) and using $x_{.j}$ as a level-2 MOD: $x_{ij} = (x_{ij} - x_{.j}) + x_{.j}$. In this design, researchers can test whether the level-1 part of MOD ($x_{ij} - x_{.j}$) or level-2 part of MOD ($x_{.j}$) moderates a level-2 TRT (z_j) effect on an outcome variable (y_{ij}). Below, with the unconflated level-1 MOD ($x_{ij} - x_{.j}$) and the level-2 MOD ($x_{.j}$), we first present a MLM specification for a *linear* multilevel interaction of categorical TRT $\times$ continuous MOD for comparison purposes and then present a MLM specification for a *nonlinear* multilevel interaction with smooth functions as a special case of GAMM.

Following the MLM specification, notations, and symbols in Raudenbush and Bryk (2002), MLM with a random intercept β_{0j}, a random slope β_{1j}, and a *linear* interaction of dummy-coded TRT $\times$ continuous MOD is written as follows for a two-level nested design:

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

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

Reduced Form:

$$y_{ij} = \gamma_{00} + \gamma_{10}(x_{ij} - x_{.j}) + \gamma_{01}x_{.j} + \gamma_{02}z_j + \gamma_{03}x_{.j}z_j + \gamma_{11}(x_{ij} - x_{.j})z_j + u_{0j} + (x_{ij} - x_{.j})u_{1j} + r_{ij}, \quad (2)$$

where γ_{00} is a fixed intercept; γ_{10} is a fixed effect of a level-1 component $(x_{ij} - x_{.j})$ of MOD x_{ij} where $z_j = 0$; γ_{01} is a fixed effect of a level-2 component $x_{.j}$ of MOD x_{ij} where $z_j = 0$; γ_{02} is a fixed effect of a dummy coded level-2 TRT z_j (a conditional treatment effect where $x_{.j} = 0$); γ_{03} is a fixed linear interaction of a level-2 component of MOD $(x_{.j})$ and a dummy coded level-2 TRT z_j; γ_{11} is a fixed linear interaction of a level-1 component of MOD $(x_{ij} - x_{.j})$ and a dummy coded level-2 TRT z_j; u_{0j} is a random intercept; u_{1j} is a random slope of a level-1 component $(x_{ij} - x_{.j})$ of MOD x_{ij}; and r_{ij} is random error. The random effects, $[u_{0j}, u_{1j}]'$, are assumed to follow a multivariate normal (MN) distribution, $[u_{0j}, u_{1j}]' \sim MN(\mathbf{0}, \Sigma)$, with a random intercept variance τ_{00}, a random slope variance τ_{11}, and a covariance τ_{01} in Σ. The random error, r_{ij}, is assumed to follow a normal (N) distribution, $r_{ij} \sim N(0, \sigma^2)$.

As a special case of GAMM, MLM with a random intercept, a random slope of a level-1 covariate, and a smooth function for a *nonlinear* multilevel interaction of dummy-coded TRT $\times$ continuous MOD is written as follows:

$$y_{ij} = \gamma_{00} + \gamma_{02}z_j + f_1(x_{.j})(z_j = 0) + f_1(x_{.j})(z_j = 1) + f_2(x_{ij} - x_{.j})(z_j = 0) + f_2(x_{ij} - x_{.j})(z_j = 1)$$
$$+ u_{0j} + (x_{ij} - x_{.j})u_{1j} + r_{ij}, \quad (3)$$

where γ_{02} is a 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 a smooth function of a level-2 component $x_{.j}$ of MOD x_{ij} where $z_j = 0$ (i.e., nonlinear level-2 interaction for a control group); $f_1(x_{.j})(z_j = 1)$ is a smooth function of a level-2 component $x_{.j}$ of MOD x_{ij} where $z_j = 1$ (i.e., nonlinear level-2 interaction for a treatment group); $f_2(x_{ij} - x_{.j})(z_j = 0)$ is a smooth function of a level-1 component $x_{ij} - x_{.j}$ of MOD x_{ij} where $z_j = 0$ (i.e., nonlinear level-1 interaction for a control group); and $f_2(x_{ij} - x_{.j})(z_j = 1)$ is a smooth function of a level-1 component $x_{ij} - x_{.j}$ of MOD x_{ij} where $z_j = 1$ (i.e., nonlinear level-1 interaction for a treatment group).

In MLM (Equation 2) and GAMM (Equation 3), the same fixed intercept terms for TRT (z_j), $\gamma_{00}+\gamma_{02}z_j$, are specified. However, different slope terms of MOD (x_{ij}) for TRT (z_j) are specified in MLM and GAMM. In MLM, $\gamma_{03}x_{.j}z_j$ and $\gamma_{11}(x_{ij}-x_{.j})z_j$ terms are for modeling *linear* multilevel interaction: the $\gamma_{03}x_{.j}z_j$ is for the linear TRT $\times$ level-2 MOD interaction and $\gamma_{11}(x_{ij}-x_{.j})z_j$ is for the linear TRT $\times$ level-1 MOD interaction. In GAMM, the $f_1(x_{.j})(z_j=0)$, $f_1(x_{.j})(z_j=1)$, $f_2(x_{ij}-x_{.j})(z_j=0)$, and $f_2(x_{ij}-x_{.j})(z_j=1)$ are for modeling *nonlinear* multilevel interactions: the $f_1(x_{.j})(z_j=0)$, and $f_1(x_{.j})(z_j=1)$ are for the nonlinear TRT $\times$ level-2 MOD interaction, and $f_2(x_{ij}-x_{.j})(z_j=0)$ and $f_2(x_{ij}-x_{.j})(z_j=1)$ are for the nonlinear TRT $\times$ level-1 MOD interaction.

Smooth Functions for Categorical TRT $\times$ Continuous MOD

The univariate smooth function $f_h(x_h)$ of a covariate x is specified as a weighted sum of a set of basis functions over the covariate x_h:

$$f_h(x_h)=\sum_{k=1}^{K}\delta_{hk}b_{hk}(x_h), \tag{4}$$

where k is an index for a basis function ($k=1,\ldots,K$), x_h is a covariate for a smooth function h, δ_{hk} is a basis coefficient, and $b_{hk}(x)$ is the kth basis function for smooth function h. The basis functions ($\mathbf{b}_h=[b_{h1},\ldots,b_{hK}]'$) are a set of known curves to represent $f_h(x_h)$ and they are functioned as covariates to estimate basis coefficients ($\boldsymbol{\delta}_h=[\delta_{h1},\ldots,\delta_{hK}]'$). In the mgcv package, a smooth function is estimated with an identification constraint such that f_h sums to 0 over the observed covariate values (i.e., $\sum_v f_h(x_{hv})=0$ for each h, where v as a subscript for observations); otherwise, $f_h(x)$ can be confounded with the intercept. When the TRT is specified as a factor, the `mgcv` `R`-package automatically computes a separate smooth function for the MOD effect, for every level in TRT (Wieling, 2018).

In GAMM applications using the mgcv package, a cubic regression spline (CRS; Wood, 2017) and a thin plate regression spline (TPRS; Wood, 2017, 5.5.1) are commonly used splines for the univariate smooth function ($f_h(x)$). The CRS is a smooth curve made up of sections of cubic polynomials. The sections are joined together at locations referred to as *knots*. At each knot, the joined sections of the cubic polynomials have equivalent values, first and second derivatives (Wood, 2017, 5.3.1). In the mgcv package, the default is for the knots to be equally spaced over

the entire range of the observed covariate, and the number of knots is the same as the number of basis functions (K). The CRS and the TRPS yield comparable results for the univariate smooth function (e.g., Finch & Finch, 2018), however, the CRS yields better computational efficiency; therefore, we used the CRS in the current study.

As shown in Equation 3, a nonlinear categorial TRT $\times$ continuous MOD interaction is specified by including different smooth functions of a continuous MOD multiplied by a dummy-coded TRT, $f_1(x_{.j})z_j$ and $f_2(x_{ij} - x_{.j})z_j$. In Appendix S1, the CRS for a smooth function of $z_j = 0$ or $z_j = 1$ is illustrated.

For the selected basis functions for smooth functions, the number of basis functions (K) should be selected to obtain a good fit. The dimensionality of the basis expansion is determined by K. When K is too small, oversmoothing will occur, and when K is too large, computation time is slowed. $K = 10$ is the default in mgcv, and is often sufficient in generalized additive modeling (e.g., Bringmann et al., 2017). Thus, we set $K = 10$ in the current study. To determine whether a selected K is large enough, the value of the k-index can be assessed. The k-index is a measure of the remaining pattern in the residuals. Let $\tilde{\mathbf{r}}$ denote the vector of residuals r_{ij}, ordered according to the value of covariate x_h and define *differencing residuals* that are near neighbours according to the covariate of the smooth as $\Delta_{ij} = \tilde{r}_{(ij)+1} - \tilde{r}_{ij}$. The k-index is calculated as the ratio of (a) an estimate of the mean of the squared differencing residuals ($\sigma^2_\Delta = E[\Delta^2_{ij}]$) to (b) an estimate of residual variance from a model-fit (σ^2) (Wood, 2017, p. 243, p. 330). A k-index below 1 indicates that there is a missed pattern left in the residuals with a specified K, and a larger K should be considered in the case of a k-index below 1. The k-index can be obtained through the `gam.check` function in mgcv.

In addition to the k-index, the corrected Akaike information criterion (corrected AIC; Wood, Pya, & Säfken, 2016) is considered to select a model with an adequate amount of smoothing from the data among candidate models differing in K, as a commonly used model selection criterion in generalized additive modeling (Ruppert, Wand, & Carroll, 2003, p. 120). The corrected AIC for GAMM uses the effective degrees of freedom (*edf*) as the number of parameters needed to

represent smooth functions in the penalty term of Akaike information criterion (AIC; Akaike, 1974). The corrected AIC is specified as follows:

$$\text{CorrectedAIC} = -2ll + (2 \times edf), \tag{5}$$

where ll is the log-likelihood. The log-likelihood and edf in the corrected AIC for GAMM can be extracted using the function `logLik.gam` for a fitted model in the `mgcv` package.

Parameter Estimation

The `glmm` function in the `mgcv` package was used for maximum likelihood estimation. Below, we describe the details of the implementation of the `glmm` function for the specified GAMM (Equation 3).

The 'wiggliness' of smooth function $f_h(x)$ is controlled less by K (the number of basis functions) than by a quadratic smoothing penalty (e.g., Wood, 2017). The quadratic smoothing penalty for the model can be written as:

$$\lambda_h \boldsymbol{\delta}_h^T \mathbf{S}_h \boldsymbol{\delta}_h, \tag{6}$$

where λ_h is a smoothing parameter, $\boldsymbol{\delta}_h$ is a vector of basis coefficients, and $\mathbf{S}_h$ is a penalty matrix embedded as a diagonal block in a matrix. For smooth functions, the elements of $\mathbf{S}_h$ are known and are determined by the chosen basis functions. The parameter λ_h controls the trade-off between goodness of fit and model smoothness.

For the identity link, the `glmm` function uses a GLMM formulation to fit GAMM. Wood (2004; 2006; 2017, p. 239) presented how a smooth function in GAMM can be reformulated into fixed and random effects in GLMM. Key derivations in Wood (2004; 2006; 2017, p. 239) are explained and illustrated in Appendix S2.

Testing for nonlinear TRT $\times$ MOD Interactions

To determine whether or not the smooth function $f_h(x)$ is distinguishable from zero, the following null hypothesis can be tested: $H_0 : f_h(x) = 0$ for all x in the range of interest. A test statistic for $f_h(x)$ is

$$T_r = \widehat{\mathbf{f}}_h^T \mathbf{V}_{f_h}^- \widehat{\mathbf{f}}_h, \tag{7}$$

where r is the rounded effective degrees of freedom (*edf*) of $f_h(x)$ (integer; e.g., $r = 1$ in the case of $edf = 1.45$)[1], the $\widehat{\mathbf{f}}_h$ is the vector of $f_h(x)$ evaluated at the observed predictor values, and $\mathbf{V}_{f_h}^-$ is a rank r pseudo-inverse of $\mathbf{V}_{f_h}$ ($\mathbf{V}_{f_h} = \mathbf{X}\mathbf{V}_{\boldsymbol{\delta}}\mathbf{X}^T$ where $\mathbf{X}$ are basis functions and $\mathbf{V}_{\boldsymbol{\delta}}$ 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).

Smooth functions have confidence intervals around them, which are obtained by taking the quantiles from the posterior distribution of the $f_h(x)$ (Marra & Wood, 2012). To calculate the distribution of the $f_h(x)$, a large number (e.g., 1,000) of basis coefficient parameters ($\boldsymbol{\delta}_h$) are simulated from the posterior distributions of basis coefficients using a multivariate normal (MN) distribution:

$$\boldsymbol{\delta}_h \sim MN(\widehat{\boldsymbol{\delta}}_h, \mathbf{V}_{\boldsymbol{\delta}}), \tag{8}$$

where $\widehat{\boldsymbol{\delta}}_h$ contains basis coefficient estimates. Then, a large number of the $f_h(x)$ can be calculated using sampled basis coefficients and basis functions using Equation 4. The 0.025 and 0.975 quantiles of the posterior distribution can be used for the lower and upper bounds of a 95% confidence interval of the smooth functions.

In addition to significance testing for the smooth function, we can visualize over which ranges of MOD the smooth functions differ significantly (called *the region of significance*). As an example, varying treatment effects depending on the levels of x can be estimated for each level of a TRT variable (one for a control group and another for a treatment group) using the GAMM specification (Equation 3), as shown in Figure 1 (left panel). Based on the result of the smooth functions, *differences* in the smooth functions in Figure 1 (left panel) (a smooth function for a treatment group - a smooth function for a control group) can be presented as in Figure 1 (right panel). Windows of significant differences are found in ranges between 2.606 and 16.505, and between 26.495 and 40.828 (noted with vertical bars in Figure 1 [right]) in x.

3. Empirical Study

[1]In the `gamm4` output, the degrees of freedom used in computing test statistic and the p-values are presented as Ref.df.

In this section, a GAMM specification is illustrated using an empirical data set to detect a nonlinear multilevel interaction of categorical level-2 TRT and continuous level-1 MOD in a C-RCT design. The data set comes from D. Fuchs et al. (2021). The purpose of the study was to evaluate the efficacy of two revised versions of First-Grade Peer-Assisted Learning Strategies (PALS) called the PALS-Only and PALS+Fluency programs. Using a subset of the data from D. Fuchs et al. (2021), an analysis goal in the present study is to test whether the contrast of both PALS conditions vs. control moderates the effect of students' pre-treatment scores of phonological awareness (prePA) on students' post-treatment scores of phonological awareness (postPA).

Data Description

Participants

Teachers from 33 first-grade classrooms in 8 elementary schools and their 491 students participated. In a C-RCT, the 33 teachers were assigned randomly within schools to a control group (11 teachers and their 171 students) and two treatment groups (the PALS-Only [11 teachers and their 168 students] and PALS+Fluency groups [11 teachers and their 152 students]). Cluster size (the number of students per classroom) ranged from 1 to 18 (median=15, semi-interquartile range=1). One classroom had one student in a PALS+Fluency group. D. Fuchs et al. (2021) reported that there were no statistically significant differences among the three study groups on student demographics, teacher demographics, or pretreatment reading performance.

Measures

Students were tested before and immediately following the 22-week treatment period. The same measures, used by Yopp (1998) for segmenting sounds in words and by L. Fuchs et al. (2001) for blending sounds in words, were used to assess phonological awareness at two time points. In this study, the same continuous scores of phonological awareness were used in GAMM and MLM analyses as were used in D. Fuchs et al. (2021). There are no missing data in prePA or postPA. As descriptive information, the normality assumption of the postPA scores is tested. A Shapiro-Wilk test indicated that the postPA scores were significantly non-normal ($W = 0.950$, $p < .0001$). However, the deviance from normality is not large in the Q-Q plot and it is localized

in the tails of the distribution. As employed in D. Fuchs et al. (2021), the zero-centering of the prePA and postPA scores was used. The mean and standard deviation of prePA scores are 0 and 0.911, respectively, and the mean and standard deviation of postPA scores are 0 and 0.842, respectively.

Analyses and Results

Below, analysis steps are described to test whether the contrast of both PALS conditions vs. control moderates the effect of prePA on students' postPA. The `R` code for empirical data analyses is presented in Appendix S3.

Step 1: Unconditional GAMM

The postPA results are from a nested data structure, 491 students (level 1) nested within 33 classes (level 2), nested within 8 schools (level 3). Dependencies in the postPA scores due to clusters (classes and schools) can be accounted for in the three-level model. However, there was a convergence problem with estimating the variance of the random intercept for schools in the unconditional three-level random intercept model.[2] This problem may be due to having only 8 schools, which is a lower number of upper-level units than the number recommended in multilevel modeling (Snijders & Bosker, 2011). In this circumstance, it is recommended to replace a random intercept with $L-1$ (where L is the number of clusters) dummy codes for cluster membership (e.g., McNeish & Stapleton, 2016). The unconditional GAMM with the $L-1$ dummy codes for 8 schools is specified as follows (the model can be called MLM because smooth functions are not introduced yet):

$$y_{ij} = \gamma_{00} + \sum_{l=2}^{L} \alpha_l D_l + u_{0j} + r_{ij}, \tag{9}$$

where l is an index for a school ($l = 2, \ldots, L$; $L = 8$ in this example), D_l is a dummy code for school membership with the first school as a reference school, and α_l is the fixed effect of D_l. The intraclass correlation coefficient (ICC), calculated based on the results of Equation 9 was 0.107 ($= 0.066/[0.066 + 0.552]$ where 0.066 is $\widehat{\tau}_{00}$ and 0.552 is $\widehat{\sigma}^2$), which suggests that there is

[2] Warning message from the `lmer` function in the `lme4` package is "Model failed to converge with max—grad— = 0.00206106 (tol = 0.002, component 1)."

non-ignorable dependency in postPA scores due to class clustering.

Step 2: Adding Covariates (TRT and MOD), and Comparing and Checking Models

In Step 2, a dummy-coded level-2 TRT ($z_j = 0$ for a control group; $z_j = 1$ for PALS-Only and PALS+Fluency groups) and a cluster (class)-centered level-1 MOD (prePA x_{ij}; $x_{ij} - x_{.j}$ and $x_{.j}$) were considered as covariates. Prior to modeling, the relationship between postPA scores (y_{ij}) and prePA scores ($x_{ij} - x_{.j}$ or $x_{.j}$) was explored by TRT groups (z_j) using scatter plots. Figure 2 (top) presents a scatter plot of y_{ij} vs. $x_{ij} - x_{.j}$ by z_j and Figure 2 (bottom) presents a scatter plot of y_{ij} vs. $x_{.j}$ by z_j. This figure shows nonlinear relationships presented with smooth lines deviant from the linear dotted lines at each level. In addition, the figure shows that the differences in postPA between the two groups (presented with 95% confidence bands) differ depending on the levels of prePA. Given the patterns identified in Figure 2, smooth functions of $f_1(x_{.j})(z_j = 0)$, $f_1(x_{.j})(z_j = 1)$, $f_2(x_{ij} - x_{.j})(z_j = 0)$, and $f_2(x_{ij} - x_{.j})(z_j = 1)$ (along with z_j) were added to Equation 9.

The the random-intercept-and-slope GAMM and the random-intercept GAMM were fit with 10 basis functions ($K = 10$ as the `mgcv` default). The k-index was close to 1 and the corrected AIC differs in the first decimal place for the models differing $K = 10$ (correctedAIC = 897.706), 12 (correctedAIC = 897.683), 14 (correctedAIC = 897.606), and 16 (correctedAIC = 897.453). These results indicate that $K = 10$ is adequate to obtain a good fit in both models.

Results of the random-intercept-and-slope GAMM were compared with those of the random-intercept GAMM. In the random-intercept-and-slope GAMM, $\widehat{\tau}_{11}$ was $3.699829e-09$. In addition, estimates and standard errors of fixed effects differed in the second or third decimal places and patterns in the smooth functions were similar between the two models. Furthermore, AIC and the Bayesian information criterion (BIC; Schwarz, 1978) suggested that the random-intercept GAMM fits better than the random-intercept-and-slope GAMM (see AIC and BIC values in Table 1 (top)). Thus, the random-intercept GAMM was chosen for result interpretations. Residual analysis of the model indicates that there is evidence of good model-data fit (see Appendix S4).

Step 3: Interpreting Results

Table 1 presents results of the random-intercept GAMM, compared with the MLM results we will discuss in the following subsection. Below, the GAMM results are interpreted.

A significant fixed TRT (z_j) effect was found ($\widehat{\gamma}_{02} = 0.439$, SE=0.071). This result means that students in PALS-Only and PALS+Fluency programs together outperformed control students. The corresponding effect size (Hedges's g) is 0.848, following the guideline suggested by What Works Clearinghouse (WWC, 2017). When TRT is a focal covariate at the class level (level 2) and MOD is a moderator, the level-2 TRT effect (z_j) on class means ($y_{.j}$) at any chosen value of the level-2 part of MOD ($x_{.j}$) is of interest to interpret. Figure 3 (a) presents the effect of z_j on $y_{.j}$ by quantiles (0.1, 0.25, 0.5, 0.75, 0.9) of $x_{.j}$ and Figure 3 (b) shows the level-2 TRT effects ($\widehat{\gamma}_{02} + \{\tilde{f}_1(x_{.j})(z_j = 1) - \tilde{f}_1(x_{.j})(z_j = 0)\}$) vs. $x_{.j}$ from GAMM. The region of significance of class-level prePA ($x_{.j}$) was $[-0.494, 0.568]$, presented in the vertical lines of Figure 3 (b). PostPA scores were higher for the treatment groups than for the control group in the range of $[-0.494, 0.568]$. However, the difference at the extremes was not significant (see Figure 3 (b)).

Comparisons between GAMM and MLM

Because MLM is a dominant analytic method to evaluate TRT effects in a C-RCT, results of GAMM and MLM are compared for instructive purposes in Table 1. Similarly to the GAMM, there was a convergence problem with estimating a random intercept variance for classes in the unconditional three-level MLM. $L-1$ (where L is the number of schools) dummy codes for school memberships were considered to account for dependency due to school clustering. In addition, there was a convergence problem with estimating τ_{11} in the random-intercept-and-slope model (Equation 2).[3] Thus, the random-intercept MLM with dummy codes for school memberships was considered for comparison with the random-intercept GAMM.

Regarding model selection between the random-intercept GAMM and the random-intercept MLM as shown in Table 1 (top), the random-intercept GAMM fits better than the random-intercept MLM based on AIC, whereas the random-intercept MLM fits better than the random-intercept GAMM based on BIC. However, differences in BIC between the two models were small

[3]The error message from the `lme` function was "nlminb problem, convergence error code = 1 message = iteration limit reached without convergence (10)."

(975.32 for GAMM - 974.93 for MLM = 0.39), indicating that there is no strong evidence that the random-intercept MLM fits better than the random-intercept GAMM based on BIC.

There were similar patterns in the effects of school memberships and variances of random effects between GAMM and MLM, as shown in Table 1 (bottom). Standard errors of fixed effects were larger in MLM than in GAMM. For the comparison with an interaction between a level-2 TRT (z_j) and the level-2 part of MOD ($x_{.j}$) from GAMM, Figure 3 (c) presents the effect of z_j on $y_{.j}$ by quantiles (0.1, 0.25, 0.5, 0.75, 0.9) of $x_{.j}$ and Figure 3 (d) shows the level-2 TRT effects ($\widehat{\gamma}_{02} + \widehat{\gamma}_{03}x_{.j}$) vs. $x_{.j}$ from MLM. In Figure 3 (d), 95% confidence bands were calculated as $(\widehat{\gamma}_{02} + \widehat{\gamma}_{03}x_{.j}) \pm z_{crit.}\sqrt{Var(\widehat{\gamma}_{02}) + Var(\widehat{\gamma}_{03})x_{.j}^2 + 2x_{.j}Cov(\widehat{\gamma}_{02}, \widehat{\gamma}_{03})}$. Unlike GAMM (Figure 3 (a)), postPA scores were higher for the treatment groups than for the control group over all quantiles of $x_{.j}$, as shown in Figure 3 (c). The region of significance of class-level prePA ($x_{.j}$) was $[-1.690, 0.800]$, presented in the vertical lines of Figure 3 (d).

4. Simulation Study

A simulation study was designed to investigate the following: (a) relative performance of GAMM to detect a nonlinear multilevel TRT × MOD interaction compared with an alternative approach (MLM with a logistic function [MLM-Logistic]) in the presence of the level-specific logistic (parametric) form of the interaction and (b) the accuracy of GAMM (Equation 3) parameter estimates and their precision (standard errors). For both foci, the results of modeling a *nonlinear* TRT × MOD in GAMM are compared with those of modeling a *linear* TRT × MOD in MLM.

Simulation Design

A two-level nested design (e.g., students nested within schools) is chosen for the simulation study. As is common in education intervention studies, a balanced design for control and treatment groups was used. The following simulation conditions were varied because they are expected to affect parameter recovery and precision in multilevel designs (Lüdtke et al., 2011; Preacher et al., 2011): the number of clusters, cluster sizes (i.e., the number of individuals within a cluster), and ICC for outcomes. The levels of these three simulation conditions were chosen based on literature reviews on study designs for educational intervention studies in 30 papers published in *Journal of*

Educational Psychology, *American Educational Research Journal*, and *Exceptional Children*, and in other related MLM work. The number of clusters in 30 papers we reviewed ranged from 28 to 225 (median=70, semi-interquartile range=25). To mimic these numbers of clusters, the number of clusters was selected as 30 (small), 70 (medium), and 200 (large). Cluster sizes in 30 papers ranged from 13 to 35 (median=16, semi-interquartile range=7). Balanced cluster sizes were selected as 15 and 30 in the simulation study. Eight of 30 papers reported ICCs ranging from 0.08 to 0.25. ICC of outcomes was selected to be at $ICC = .05, .10$, or .30. ICC values are rarely greater than .30 in educational and organizational studies (e.g., Fox, 2010).

For the first focus (performance of GAMM in the presence of the logistic form of the TRT $\times$ MOD), the data-generating model is MLM-Logistic for the level-specific TRT $\times$ MOD. The MLM with a logistic function (MLM-Logistic) as a data-generating model is written as:

$$y_{ij} = \frac{1}{1 + exp[-\mu]} + u_{0j} + (x_{ij} - x_{.j})u_{1j} + r_{ij}, \tag{10}$$

where $\mu = \gamma_{00} + \gamma_{10}(x_{ij} - x_{.j}) + \gamma_{01}x_{.j} + \gamma_{02}z_j + \gamma_{03}x_{.j}z_j + \gamma_{11}(x_{ij} - x_{.j})z_j$, and the error term r_{ij} is assumed to be distributed independently as $N(0, \sigma^2)$. Fixed parameters were selected to generate no differences in logistic functions between control and treatment groups at level 1 and to have large differences in logistic functions between control and treatment groups at level 2 as expected in a C-RCT: $\gamma_{00} = 0$, $\gamma_{10} = 1.5$, $\gamma_{01} = 7.2$, $\gamma_{02} = 1.7$, $\gamma_{03} = 5.1$, and $\gamma_{11} = 0.1$. The three levels of ICC were manipulated using the 'true' variances of random intercept and random errors. That is, given the error variance $\sigma^2 = 0.6$, the three levels of τ_{00} were calculated as 0.032, 0.067, and 0.257 for ICC =.05, .10, and .30, respectively. The slope variance τ_{11} was set as 0.1 and the covariance τ_{01} was set to be 0. In generating the logistic functions for control and treatment groups at Level 2, it was assumed that students with low and high values of MOD would not benefit from TRT and students in the middle values of MOD may gain from TRT, but to some degrees conditional on TRT (Preacher & Sterba, 2019). The generated logistic function in MLM with $K = 10$ is shown for one condition (number of clusters=200, cluster size=30, $ICC = 0.30$) in Appendix S5 as an example.

For the second focus (parameter recovery of GAMM), the data-generating model is a random-

intercept-and-slope GAMM (Equation 3). Fixed parameters, $\gamma_{00} = 0.4$ and $\gamma_{02} = 0.4$, were selected to mimic results of the empirical study. The same σ^2, τ_{00}, τ_{11}, and τ_{01} used for the first focus were also used in the second focus. For 'true' smooth functions, increasing nonlinear functions were generated using Equation 4 with $K = 10$. According to literature reviews on educational intervention studies in the 30 papers mentioned earlier, educational interventions are implemented mainly to improve learning for students with low-achieving levels. Thus, larger nonlinear treatment effects were generated at the lower end and smaller differences were generated in the other ranges of a covariate, assuming that the intervention is the most effective for them. Parameters of basis coefficients ($\boldsymbol{\delta}_h$) and generated smooth functions are shown for one condition (number of clusters=200, cluster size=30, $ICC = 0.30$) in Appendix S5 for illustrative purpose.

In the two data-generating models, the x_{ij} was generated from a standard normal distribution and then $x_{.j}$ and $x_{ij} - x_{.j}$ were calculated. For each simulation condition, the same MOD ($x_{ij} - x_{.j}$ and $x_{.j}$) and generated functions for TRT $\times$ MOD were used across replications, and random effects were generated at each replication.

The simulation conditions regarding multilevel designs were fully crossed, yielding 18 (= 3 number of clusters $\times$ 2 cluster sizes $\times$ 3 ICCs) conditions. Five hundred replications were simulated for each condition. For MLM-Logistic and GAMM as data-generating models, GAMM (Equation 3) was fit to the generated data sets. In addition, for MLM-Logistic and GAMM as data-generating models, MLM (Equation 2) was fit to the same generated data sets to demonstrate the consequences of modeling *linear* categorical TRT $\times$ continuous MOD interactions in the presence of nonlinear categorical TRT $\times$ continuous MOD interactions. And for MLM-Logistic as a data-generating model, MLM-Logistic (see Appendix S5 for estimation in `R`) was fit the same generated data sets to compare its results with GAMM's results. In addition, five candidate models regarding Ks ($K = 6, 8, 10, 12, 14$) were fitted to the generated data sets for each replication in a condition to check whether the K used in generated smooth functions was adequate based on the corrected AIC. Thus, the total number of fitted models is $225,000$ (18 multilevel designs $\times$ 500 replications $\times$ 3 models [MLM-Logistic, GAMM, and MLM] $\times$ 5 models differing Ks for

MLM-Logistic as a data-generating model; 18 multilevel designs × 500 replications × 2 models [GAMM and MLM] × 5 models differing Ks for GAMM as a data-generating model).

Analysis

For the first focus, the 'true' level-specific TRT × MOD generated using MLM-Logistic is compared with predicted level-specific TRT × MOD generated using MLM-Logistic, MLM, and GAMM, respectively. As an evaluation measure for the TRT × MOD, the root mean squared difference (RMD) between predicted values (calculated based on estimates of fixed effects for TRT × MOD) and true values (calculated based on parameters of fixed effects for TRT × MOD) was obtained. The RMD is interpreted as the standard deviation of the differences between predicted and true values. Equations to calculate the level-specific TRT × MOD in the MLM-logistic, GAMM, and MLM are presented in Table 2 and equations to calculate the RMD are presented in Table 3 (top). As a summary of the RMD, the mean of RMDs over 500 replications were reported. For variance and covariance estimates of random intercept and slope ($\widehat{\tau}_{00}, \widehat{\tau}_{11}, \widehat{\tau}_{01}$), and variance estimate of random residuals ($\widehat{\sigma}^2$) in the MLM-logistic, GAMM, and MLM, the bias (calculated $\sum_{rep=1}^{500}(\widehat{\tau}_{00.rep} - \tau_{00})/500$ where rep indicates a replication number as an example) and the root mean square error (RMSE; calculated $\sqrt{\sum_{rep=1}^{500}(\widehat{\tau}_{00.rep} - \tau_{00})^2/500}$ as an example) was used to evaluate overall accuracy (i.e., bias and variability) patterns with respect to levels of simulation conditions.

To evaluate the accuracy of estimates in the parametric part of GAMM ($\widehat{\boldsymbol{\gamma}}$, $\widehat{\Sigma}$, and $\widehat{\sigma}^2$) in the second focus, the bias[4] and RMSE were calculated and compared across simulation conditions. To evaluate the accuracy of standard errors, 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 close to 1 suggests that the estimated standard errors are approximately correct. Because smooth functions are generated based on basis functions and basis coefficients, accuracy of basis coefficient estimates ($\widehat{\boldsymbol{\delta}}$ which are calculated based on $\widehat{\boldsymbol{\gamma}}_s$ and $\widehat{\boldsymbol{\lambda}}$) was evaluated using the bias and RMSE for the smooth functions. As shown in Equation 8,

[4]Relative percentage bias was not considered because it leads to scaling problems in the case of parameters close to 0 as in our simulation study.

the precision of the smooth functions depends on the standard errors of basis coefficient estimates. Thus, for the smooth functions, the standard errors of the basis coefficient estimates were evaluated using the ratio of M(SE) to SD. In addition, the RMD between predicted values of the level-specific TRT × MOD under GAMM or MLM and 'true' smooth functions was obtained (see RMD calculations in Table 3 (bottom)). To summarize the results of the RMD, its mean over 500 replications was obtained. For GAMM, the k-index for $K = 10$ used in generating smooth functions was close to 1 for all smooth functions in all conditions and a model with $K = 10$ was selected among candidate models differing in K ($K = 6, 8, 10, 12, 14$) based on the corrected AIC. These results indicate that $K = 10$ is adequate. In addition, to show the effects of modeling *linear* TRT × MOD interactions on MLM parameter estimates in the presence of nonlinear TRT × MOD interactions, the bias, RMSE, and the ratio of M(SE) to SD were calculated for MLM estimates which are not part of the linear interaction effects and are comparable with GAMM estimates: the intercept ($\widehat{\gamma}_{00}$), the effect of TRT ($\widehat{\gamma}_{02}$), the covariance matrix of random effects ($\widehat{\Sigma} = [\widehat{\tau}_{00}, \widehat{\tau}_{01}, \widehat{\tau}_{10}, \widehat{\tau}_{11}]'$), and the residual variance ($\widehat{\sigma}^2$) in Equation 2.

Results

Below, results are summarized by data-generating models. No convergence problems occurred in any simulation condition for MLM-Logistic, GAMM, or MLM.

Results for MLM with a Logistic Function as a Data-Generating Model

Table 4 shows the average RMD across 500 replications for level-specific TRT × MOD interactions, and bias and RMSE of $\widehat{\tau}_{00}$, $\widehat{\tau}_{11}$, $\widehat{\tau}_{01}$, and $\widehat{\sigma}^2$ in MLM-Logistic, GAMM, and MLM. In Table 4, the averaged results are reported by the levels of simulation conditions in order to understand the main effects of each condition. Results of all 18 simulation conditions are presented in the figures of Appendix S6.

Performance of GAMM compared with MLM-Logistic for Predictions of Level-Specific TRT × MOD Interactions. When a data-generating model is MLM-Logistic, the RMD in GAMM are smaller than the RMD in MLM-Logistic as an alternative approach to GAMM in all simulation conditions (RMD ranged from 0.042 to 0.126 at level 1 and ranged from 0.106 and

0.281 at level 2 in GAMM; RMD ranged from 0.082 to 0.301 at level 1 and ranged from 0.250 and 0.410 at level 2 in MLM-Logistic). This result indicates that a logistic form of the level-specific TRT × MOD interactions can be recovered better using smooth functions in GAMM than using non-linear fitting in MLM. Regarding patterns in RMD by the levels of simulation conditions, the following is observed. First, in MLM-Logistic, RMD decreased with increasing number of clusters (J) and decreasing cluster size (n_j) at level 1, whereas it decreased with decreasing the number of clusters (J) and increasing cluster size (n_j) at level 2. In GAMM, RMD decreased with increasing number of clusters (J) and cluster size (n_j) at both levels. Second, RMD decreased with increasing ICCs at level 1 and at level 2 (from $ICC = 0.1$ to $ICC = 0.3$) in MLM-Logistic, and it decreased with decreasing ICCs at level 2 in GAMM; in GAMM, ICC had no effect on RMD at level 1.

Effects of Modeling Linear TRT × MOD on Predictions for Level-Specific TRT × MOD Interactions. Under MLM-Logistic as a data-generating model, the RMD in MLM with linear TRT × MOD are larger than the RMD in MLM-Logistic or in GAMM in all simulation conditions (RMD ranged from 0.499 to 0.597 at level 1 and ranged from 0.512 and 0.663 at level 2 in MLM). In MLM, there were small differences in RMDs across simulation conditions (differed in the second decimal place of RMDs). These results suggest that misspecifying the functional forms for TRT × MOD interactions in MLM leads to biased predictions of the interactions in all multilevel designs we considered.

Random Effects Comparisons. Overall, bias and RMSE of $[\widehat{\tau}_{00}, \widehat{\tau}_{11}, \widehat{\sigma}^2]'$ were smaller in GAMM than in MLM-Logistic. For $\widehat{\tau}_{01}$, smaller bias was observed in GAMM than MLM-Logistic while larger RMSE was found in GAMM than MLM-Logistic. In addition, bias and RMSE of $[\widehat{\tau}_{00}, \widehat{\tau}_{11}, \widehat{\sigma}^2]'$ in MLM were similar to those in GAMM because misspecification in MLM is not for these variances but for the level-specific TRT × MOD interactions. With respect to simulation conditions, bias and RMSE of $[\widehat{\tau}_{00}, \widehat{\tau}_{11}, \widehat{\sigma}^2]'$ tended to decrease as the number of clusters (J) and cluster size (n_j) increased in all three models, except for a few cases: (a) the bias and RMSE of $\widehat{\tau}_{00}$ in MLM-Logistic regarding n_j, (b) bias of $\widehat{\tau}_{11}$ and $\widehat{\sigma}^2$ regarding n_j in MLM-Logistic, and (c) bias of $\widehat{\sigma}^2$ with respect to J and n_j in MLM. As noticeable patterns regarding ICCs, bias and

RMSE of $\widehat{\tau}_{00}$ tended to decrease with decreasing ICCs in MLM-Logistic and GAMM. And bias and RMSE of $\widehat{\tau}_{11}$ decreased with increasing ICCs in MLM-Logistic and those of $\widehat{\tau}_{01}$ decreased with increasing ICCs mainly in GAMM.

Results for GAMM as a Data-Generating Model

Averaged bias, RMSE, and the ratio of M(SE) to SD by levels of simulation conditions are reported in Table 5 for fixed and random effects of GAMM and MLM, and in Table 6 for averaged basis coefficients across 9 basis coefficients of each smooth function in GAMM. Results of all 18 simulation conditions are presented in the figures of Appendix S7.

Accuracy of Parameter Estimates and Precision of GAMM. As shown under GAMM columns in Table 5 and the figures of Appendix S7, the bias of intercept estimate ($\widehat{\gamma}_{00}$), TRT estimate ($\widehat{\gamma}_{02}$), variance and covariance estimates of random intercept and slope ($\widehat{\tau}_{00}, \widehat{\tau}_{11}, \widehat{\tau}_{01}$), and variance estimate of random residuals ($\widehat{\sigma}^2$) was close to 0 (ranging from $-.086$ to .109 for all fixed and random estimates in all 18 conditions). Overall, bias and RMSE of these estimates decreased with increasing number of clusters (J) and cluster size (n_j). For $\widehat{\gamma}_{00}$, $\widehat{\gamma}_{02}$, and $\widehat{\tau}_{00}$, bias and RMSE of the estimates decreased with smaller ICCs. However, this pattern was not observed for the other parameter estimates. Except for two of the conditions with the smallest number of clusters and smallest cluster size ($J = 30$, $n_j = 15$, $ICC = 0.05$ and $J = 30$, $n_j = 15$, $ICC = 0.1$), the ratios of M(SE) to SD for both $\widehat{\gamma}_{00}$ and $\widehat{\gamma}_{02}$ were close to 1 (ranging from 0.950 to 1.067 and from 0.964 to 1.024 across 16 conditions, respectively). The ratio approached 1 as the number of clusters and cluster size increased and as ICC decreased.

As presented in Table 6, bias of the average basis coefficient estimate (across 9 basis coefficient estimates) for each smooth function was relatively small (ranging from -0.012 to 0.278 across 18 conditions and 4 smooth functions). For all level-1 ($f_2(x_{ij} - x_{.j})(z_j = 0)$; $f_2(x_{ij} - x_{.j})(z_j = 1)$) and level-2 ($f_1(x_{.j})(z_j = 0)$; $f_1(x_{.j})(z_j = 1)$) smooth functions, bias and RMSE decreased with increasing the number of clusters and cluster size. However, different patterns were found regarding ICCs at level 1 and level 2. Bias and RMSE decreased with increasing ICCs for the smooth functions at level 1, whereas they decreased with decreasing ICCs for the smooth functions at level

2. Because a larger ICC corresponds with greater between-cluster variability, the pattern at level 2 indicates that the accuracy of basis coefficients for level-2 smooth functions can decrease when there is greater between-cluster variability. The ratio of M(SE) to SD for level-1 smooth functions ranged from 0.900 to 1.003 and the ratio of M(SE) to SD for level-2 smooth functions ranged from 0.920 to 1.033. In addition, as shown in Table 6, the mean RMD across 500 replications are close to 0 (ranged from 0.037 to 0.326) across all simulation conditions, indicating that the predicted smooth functions are close to the true smooth functions. The RMD decreased with increasing the number of clusters (J) and cluster size (n_j) for all smooth functions. They decreased with decreasing ICC for the level-2 smooth functions, whereas they were not affected by ICC for the level-1 smooth functions.

Effects of Modeling Linear TRT $\times$ MOD on Estimates of MLM and on Prediction for Level-Specific TRT $\times$ MOD Interactions. As presented in Table 5, a larger bias, RMSE, and ratio of M(SE) to SD were observed in estimates of MLM than in those of GAMM. First, for all parameter estimates reported in Table 5, bias decreased with decreasing number of clusters and cluster sizes, except for $\widehat{\tau}_{01}$ and $\widehat{\sigma}^2$ regarding the number of clusters. Second, for all parameter estimates, RMSE decreased with increasing the number of clusters and cluster sizes, with a few exceptions for $\widehat{\tau}_{00}$ with respect to the number of clusters and cluster sizes, and for $\widehat{\sigma}^2$ with respect to cluster sizes. Third, bias tended to be larger with decreasing levels of ICCs for all parameter estimates except $\widehat{\gamma}_{00}$. Fourth, $\widehat{\tau}_{00}, \widehat{\tau}_{11}$, and $\widehat{\tau}_{01}$ from MLM were overestimated. Fifth, the ratio of M(SE) to SD in MLM ranged from 1.210 to 2.241 for $\widehat{\gamma}_{00}$ and from 1.110 to 2.233 for $\widehat{\gamma}_{02}$ across 18 simulation conditions, indicating that standard errors of $\widehat{\gamma}_{00}$ and $\widehat{\gamma}_{02}$ were overestimated. The degree of overestimation of standard errors increased mainly with decreasing number of clusters and ICCs. To conclude, these results suggest that misspecifying the functional forms for TRT $\times$ MOD interactions leads to biased estimates of MLM parameters. In addition, as shown in Table 6, larger RMDs were found in MLM than in GAMM across all simulation conditions, indicating that modeling linear TRT $\times$ MOD interactions leads to biased predictions of the interactions in the presence of the nonlinear interactions.

5. Summary and Discussion

In this paper, we presented a GAMM specification to model a nonlinear multilevel TRT × MOD interaction with unconflated effects in intervention studies from C-RCT designs. The nonlinear multilevel TRT × MOD interaction was modelled using smooth functions in GAMM. Maximum likelihood estimation was implemented using the `gamm` function in the `mgcv` package. Because the smooth functions are reformulated as random effects in the `gamm` function, it may be challenging for researchers in the social and behavioral sciences to interpret results from the software output. Thus, core derivations from the statistical literature were explained.

The GAMM specification and its estimation were illustrated using instructional intervention data from a C-RCT. We provided the `R` code to visualize the nonlinear multilevel TRT × MOD interaction with unconflated effects. Because MLM is a dominant analytic method to detect a multilevel TRT × MOD interaction in education, the GAMM results were contrasted with MLM results. In GAMM, TRT effects were different depending on the values of MOD (pretest scores). However, this pattern was obscured when the linear multilevel TRT × MOD interaction was modelled in MLM. In addition, simulation studies were implemented to evaluate the performance of GAMM in recovering the level-specific logistic (parametric) form of the TRT × MOD interaction, compared with MLM-Logistic as an alternative approach and MLM as a misspecification approach. We found that GAMM outperformed MLM-Logistic to recover the level-specific logistic form of TRT × MOD interaction and MLM led to incorrect prediction of the interaction. Simulation studies were also conducted to evaluate parameter recovery in GAMM and to show consequences of modeling a nonlinear multilevel TRT × MOD as a linear multilevel TRT × MOD. The parameter recovery in GAMM was relatively satisfactory in most multilevel designs typical of educational intervention studies except designs with small number of clusters, small cluster size, and small ICC. When ignoring the nonlinear multilevel TRT × MOD interaction, biased estimates such as overestimated standard errors and overestimated variance estimates of random effects were found. These bias patterns were also observed in the empirical study.

The following methodological limitations remain because this paper is the first attempt to

apply GAMM to model a nonlinear multilevel TRT × MOD interaction with unconflated effects for educational intervention studies. First, we presented the GAMM specification for two groups (control vs. treatment groups) in a two-level nested design. As in D. Fuchs et al. (2021), there can be three groups (control, treatment 1, and treatment 2 groups). In this example, the two contrasts (e.g., control vs. treatment 1 + treatment 2; treatment 1 vs. treatment 2) can be created for the nonlinear multilevel TRT × MOD interaction with unconflated effects. In addition, there are more complex multilevel designs than the two-level nested design, such as three levels with cross-classified units (e.g., student [level 1] nested in a cross-classification of rater and classroom [both level 2] nested in school [level 3]). Further studies are needed to apply GAMM to detect a nonlinear multilevel TRT × MOD interaction with unconflated effects for more than two groups and in more complex multilevel designs. Second, the simulation study results are limited to the selected simulation conditions and the selected parameters and nonlinear functions in this study. More extensive simulations that vary these limited conditions should be conducted to make solid generalizations. Third, when newly specified GAMMs are presented to researchers in substantive areas, it is important to plan sample sizes to ensure high power for detecting hypothesized magnitudes of ATEs and variability in treatment effects. In a C-RCT design, 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 × MOD (Raudenbush & Liu, 2000). Equations for power calculation have been provided for TRT × MOD. For example, Raudenbush and Liu (2000) derived a F-statistic with the noncentrality parameter for the *conflated* fixed effects and variances of random effects of site-level TRT × MOD in a multisite randomized trial (MRT) in which individuals are randomly assigned within sites. Dong, Kelcey, and Spybrook (2020) provided power calculation formulas for level-1 TRT × binary and continuous MOD in MRTs. Bloom (2005) presented power calculation formula for TRT × binary level-1 or level-2 MOD in two-level C-RCTs. Spybrook, Kelcey, and Dong (2016) provided power calculation formulas for level-2 TRT × Level-1 binary MOD and in C-RCT. Dong, Kelcey, and Spybrook (2018) presented power calculation formulas for level-3 TRT × level-1 binary and continuous MOD

in C-RCT. However, existing formulas for power calculation have not been designed for detecting *unconflated* fixed and variances of random effects for TRT × MOD in the C-RCT design. Further studies are needed to provide equations of power calculations to detect such effects.

MLM is frequently used to detect a linear multilevel TRT × MOD interaction with conflated effects in educational intervention studies. However, conflation results in insensitivity to theoretically meaningful interactions, and estimates a weighted average of within- and between-cluster effects in the presence of level-specific interactions effects. In addition, a linear interaction is a misspecification in the existence of a more complex nonlinear interaction. The main goal of this study is to illustrate the applicability of GAMM to detect a nonlinear multilevel TRT × MOD interaction with unconflated effects. We hope that this paper can serve as an example of modelling nonlinear effects using smooth functions in GAMM for educational intervention research.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). Russell Sage Foundation.

Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (2017). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological Methods, 22(3)*, 409–425. https://doi.org/10.1037/met0000085

Dong, N., Kelcey, B., & Spybrook, J. (2018). Power analyses of moderator effects in three-level cluster randomized trials. *Journal of Experimental Education, 86,* 489–514. http://doi.org/10.1080/00220973.2017.1315714

Dong, N., Kelcey, B., & Spybrook, J. (2020). Design considerations in multisite randomized trials probing moderated treatment effects. *Journal of Educational and Behavioral Statistics.* https://doi.org/10.3102/1076998620961492

Finch, W. H., & Finch, M. H. (2018). A simulation study evaluating the generalized additive model for assessing intervention effects with small samples. *Journal of Experimental Education, 86(4)*, 652–670. https://doi.org/10.1080/00220973.2017.1339010

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications.* New York, NY: Springer. https://doi.org/10.1007/978-1-4419-0742-4

Fuchs, D., Cho, E., Toste, J. R., Fuchs, L. S., Gilbert, J. K., McMaster, K. L., Svenson, E., & Thompson, A. (2021). A quasiexperimental 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

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5(3)*, 239–256. https://doi.org/10.1207/S1532799XSSR05033

Harrell, F. E., Jr. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed). New York, NY: Springer.

Hilden-Minton, J. A. (1995). *Multilevel diagnostics for mixed and hierarchical linear models.* Doctoral dissertation, University of California Los Angeles.

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

Longford, N. T. (1993). *Random coefficient models.* Oxford University Press.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods, 16(4)*, 444–467. https://doi.org/10.1037/a0024376

Marra, G., & Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics, 39(1)*, 53–74. https://doi.org/10.1111/j.1467-9469.2011.00760.x

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28(2)*, 295–314. https://doi.org/10.1007/s10648-014-9287-x

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D, & R Core Team (2020). Package nlme: Linear and nonlinear mixed effects models. https://cran.r-project.org/web/packages/nlme/nlme.pdf

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent

curve analysis. *Journal of Educational and Behavioral Statistics, 31(4)*, 437–448. https://doi.org/10.3102/10769986031004437

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

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling, 18(2)*, 161–182. https://doi.org/10.1080/10705511.2011.557329

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods, 21(2)*, 189–205. https://doi.org/10.1037/met0000052

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5,* 199–213. http://doi.org/10.1037/1082-989x.5.2.199

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). Semiparametric regression. New York: NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511755453

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

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

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage Publications.

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,* 605–627. http://doi.org/10.3102/1076998616655442

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics, 70*, 86–116. https://doi.org/10.1016/j.wocn.2018.03.002

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association, 99(467)*, 673–686. https://doi.org/10.1198/016214504000000980

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics, 62(4)*, 1025–1036. https://doi.org/10.1111/j.1541-0420.2006.00574.x

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika, 100(1)*, 221–229. https://doi.org/10.1093/biomet/ass048

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed). Chapman & Hall/CRC

Wood, S. N. (2019). *Package 'mgcv': Mixed GAM computation vehicle with automatic smoothness estimation.* https://cran.r-project.org/web/packages/mgcv/mgcv.pdf

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

What Works Clearinghouse. (2017). *What Works Clearinghouse: Procedures handbook (Version 4.0).* U.S. Department of Education, Institute of Education Sciences. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc procedures handbook v4.pdf

Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. *Reading Research Quarterly, (23)2*, 159–177. https://doi.org/10.2307/747800

Table 1: Empirical Study: Results of Model Selection (top) and Results (bottom) of GAMM and MLM

Model	GAMM		MLM	
	AIC	BIC	AIC	BIC
Random-Intercept	895.59	975.32	911.98	974.93
Random-Intercept-and-Slope	899.59	987.71	*	*

	GAMM		MLM	
Fixed Effects				
	EST	SE	EST	SE
Intercept[γ_{00}]	**0.384**	0.133	**0.318**	0.149
$x_{ij} - x_{.j}$[γ_{10}]	-		**0.693**	0.056
$x_{.j}$[γ_{01}]	-		**0.537**	0.189
z_j[γ_{02}]	**0.439**	0.071	**0.463**	0.081
$x_{.j}z_j$[γ_{03}]	-		-0.183	0.209
$(x_{ij} - x_{.j})z_j$[γ_{11}]	-		**-0.212**	0.070
D_2[α_2]	-0.216	0.173	-0.304	0.206
D_3[α_3]	**-1.123**	0.150	**-1.056**	0.180
D_4[α_4]	**-0.353**	0.155	**-0.381**	0.183
D_5[α_5]	**-0.928**	0.139	**-0.877**	0.161
D_6[α_6]	**-0.680**	0.152	**-0.689**	0.181
D_7[α_7]	**-0.713**	0.220	**-0.506**	0.232
D_8[α_8]	**-0.679**	0.146	**-0.594**	0.168
Random Effects				
	EST		EST	
τ_{00}	0.010		0.023	
σ^2	0.316		0.337	
Smooth Functions				
	Ref.edf	T_r(p-value)		
$f_2(x_{ij} - x_{.j})(z_j = 0)$	3.630	47.747($< 2e-16$)	-	
$f_2(x_{ij} - x_{.j})(z_j = 1)$	3.153	51.731($< 2e-16$)	-	
$f_1(x_{.j})(z_j = 0)$	2.270	7.728(.000165)	-	
$f_1(x_{.j})(z_j = 1)$	1.000	3.981(.046566)	-	

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; * indicates that AIC and BIC were not reported because there was a convergence problem with estimating the random-intercept-and-slope model.

Table 2: Simulation Study: Comparisons of Level-Specific TRT × MOD among a 'True' model (MLM-Logistic), GAMM, and MLM.

	Function Form	Level 1	Level 2
MLM-Logistic	Logistic	$\frac{1}{1+exp[-\{(\gamma_{10}+\gamma_{11}z_j)(x_{ij}-x_{.j})\}]}$	$\frac{1}{1+exp[-\{(\gamma_{00}+\gamma_{02}z_j)+(\gamma_{01}+\gamma_{03}z_j)x_{.j}\}]}$
GAMM	Nonlinear	$\hat{f}_2(x_{ij}-x_{.j})(z_j=0)+\hat{f}_2(x_{ij}-x_{.j})(z_j=1)$ $=\sum_{k=1}^{9}\widehat{\delta}_{2k(z_j=0)}b_{2k}(x_{ij}-x_{.j})+\sum_{k=1}^{9}\widehat{\delta}_{2k(z_j=1)}b_{2k}(x_{ij}-x_{.j})$	$(\widehat{\gamma}_{00}+\widehat{\gamma}_{02}z_j)+\hat{f}_1(x_{.j})(z_j=0)+\hat{f}_1(x_{.j})(z_j=1)$ $=(\widehat{\gamma}_{00}+\widehat{\gamma}_{02}z_j)+\sum_{k=1}^{9}\widehat{\delta}_{1k(z_j=0)}b_{1k}(x_{.j})+\sum_{k=1}^{9}\widehat{\delta}_{1k(z_j=1)}b_{1k}(x_{.j})$
MLM	Linear	$(\widehat{\gamma}_{10}+\widehat{\gamma}_{11}z_j)(x_{ij}-x_{.j})$	$(\widehat{\gamma}_{00}+\widehat{\gamma}_{02}z_j)+(\widehat{\gamma}_{01}+\widehat{\gamma}_{03}z_j)x_{.j}$

Note. $\widehat{\delta}_{1k(z_j=0)}$ is an estimate of a basis coefficient for $\tilde{f}_1(x_{.j})(z_j=0)$; $\widehat{\delta}_{1k(z_j=1)}$ is an estimate of a basis coefficient for $\tilde{f}_1(x_{.j})(z_j=1)$; $\widehat{\delta}_{2k(z_j=0)}$ is an estimate of a basis coefficient for $\tilde{f}_2(x_{ij}-x_{.j})(z_j=0)$; $\widehat{\delta}_{2k(z_j=1)}$ is an estimate of a basis coefficient for $\tilde{f}_2(x_{ij}-x_{.j})(z_j=1)$.

Table 3: Simulation Study: Root Mean Squared Differences (RMD) between Predicted Values and True Values for MLM-Logistic (top) and GAMM (bottom) as Data-Generating Models.

Fitting Model	Level	RMD
MLM-Logistic	Level 1	$\sqrt{\sum_{j=1}^{J}\sum_{i=1}^{n_j} d_{ij.ML1}^2/Jn_j}$ where $d_{ij.ML1} = \frac{1}{1+exp[-\{(\widehat{\gamma}_{10}+\widehat{\gamma}_{11}z_j)(x_{ij}-x_{.j})\}]} - \frac{1}{1+exp[-\{(\gamma_{10}+\gamma_{11}z_j)(x_{ij}-x_{.j})\}]}$
	Level 2	$\sqrt{\sum_{j=1}^{J} d_{j.ML2}^2/J}$ where $d_{j.ML2} = \frac{1}{1+exp[-\{(\widehat{\gamma}_{00}+\widehat{\gamma}_{02}z_j)+(\widehat{\gamma}_{01}+\widehat{\gamma}_{03}z_j)x_{.j}\}]} - \frac{1}{1+exp[-\{(\gamma_{00}+\gamma_{02}z_j)+(\gamma_{01}+\gamma_{03}z_j)x_{.j}\}]}$
GAMM	Level 1	$\sqrt{\sum_{j=1}^{J}\sum_{i=1}^{n_j} d_{ij.G1}^2/Jn_j}$ where $d_{ij.G1} = \{\tilde{f}_2(x_{ij}-x_{.j})(z_j=0) + \tilde{f}_2(x_{ij}-x_{.j})(z_j=1)\} - \frac{1}{1+exp[-\{(\gamma_{10}+\gamma_{11}z_j)(x_{ij}-x_{.j})\}]}$
	Level 2	$\sqrt{\sum_{j=1}^{J} d_{j.G2}^2/J}$ where $d_{j.G2} = (\widehat{\gamma}_{00}+\widehat{\gamma}_{02}z_j) + \{\tilde{f}_1(x_{.j})(z_j=0) + \tilde{f}_1(x_{.j})(z_j=1)\} - \frac{1}{1+exp[-\{(\gamma_{00}+\gamma_{02}z_j)+(\gamma_{01}+\gamma_{03}z_j)x_{.j}\}]}$
MLM	Level 1	$\sqrt{\sum_{j=1}^{J}\sum_{i=1}^{n_j} d_{ij.M1}^2/Jn_j}$ where $d_{ij.M1} = (\widehat{\gamma}_{10}+\widehat{\gamma}_{11}z_j)(x_{ij}-x_{.j}) - \frac{1}{1+exp[-\{(\gamma_{10}+\gamma_{11}z_j)(x_{ij}-x_{.j})\}]}$
	Level 2	$\sqrt{\sum_{j=1}^{J} d_{j.M2}^2/J}$ where $d_{j.M2} = (\widehat{\gamma}_{00}+\widehat{\gamma}_{02}z_j) + (\widehat{\gamma}_{01}+\widehat{\gamma}_{03}z_j)x_{.j} - \frac{1}{1+exp[-\{(\gamma_{00}+\gamma_{02}z_j)+(\gamma_{01}+\gamma_{03}z_j)x_{.j}\}]}$

Fitting Model	Level	Group	RMD
GAMM	Level 1	$z_j = 0$	$\sqrt{\sum_{j=1}^{J}\sum_{j=1}^{n_j} d_{ij.G10}^2/Jn_j}$ where $d_{ij.G10} = \tilde{f}_2(x_{ij}-x_{.j})(z_j=0) - f_2(x_{ij}-x_{.j})(z_j=0)$
		$z_j = 1$	$\sqrt{\sum_{j=1}^{J}\sum_{j=1}^{n_j} d_{ij.G11}^2/Jn_j}$ where $d_{ij.G11} = \tilde{f}_2(x_{ij}-x.j)(z_j=1) - f_2(x_{ij}-x.j)(z_j=1)$
	Level 2	$z_j = 0$	$\sqrt{\sum_{j=1}^{J} d_{j.G20}/J}$ where $d_{j.G20} = \tilde{f}_1(x_{.j})(z_j=0) - f_1(x_{.j})(z_j=0)$
		$z_j = 1$	$\sqrt{\sum_{j=1}^{J} d_{j.G21}/J}$ where $d_{j.G21} = \tilde{f}_1(x_{.j})(z_j=1) - f_1(x_{.j})(z_j=1)$
MLM	Level 1	$z_j = 0$	$\sqrt{\sum_{j=1}^{J}\sum_{j=1}^{n_j} d_{ij.M10}^2/Jn_j}$ where $d_{ij.M10} = \widehat{\gamma}_{10}(x_{ij}-x_{.j}) - f_2(x_{ij}-x_{.j})(z_j=0)$
		$z_j = 1$	$\sqrt{\sum_{j=1}^{J}\sum_{j=1}^{n_j} d_{ij.M11}^2/Jn_j}$ where $d_{ij.M11} = \{\widehat{\gamma}_{10}(x_{ij}-x_{.j}) + \widehat{\gamma}_{11}(x_{ij}-x_{.j})\} - f_2(x_{ij}-x.j)(z_j=1)$
	Level 2	$z_j = 0$	$\sqrt{\sum_{j=1}^{J} d_{j.M20}/J}$ where $d_{j.M20} = \widehat{\gamma}_{01}x_{.j} - f_1(x_{.j})(z_j=0)$
		$z_j = 1$	$\sqrt{\sum_{j=1}^{J} d_{j.M21}/J}$ where $d_{j.M21} = (\widehat{\gamma}_{01}x_{.j} + \widehat{\gamma}_{03}x_{.j}z_j) - f_1(x_{.j})(z_j=1)$

Table 4: Simulation Study: Results for Predictions of Level-Specific TRT × MOD Interactions (top) and for Random Effects (bottom) of MLM-Logistic, GAMM, and MLM under MLM-Logistic as a Data-Generating Model

	Conditions	MLM-Logistic	GAMM	MLM
Prediction of Interactions		RMD	RMD	RMD
Level 1	$J = 30$	0.205	0.107	0.519
	$J = 70$	0.170	0.073	0.526
	$J = 200$	0.162	0.050	0.528
	$n_j = 15$	0.173	0.089	0.522
	$n_j = 30$	0.187	0.070	0.527
	$ICC = 0.05$	0.190	0.080	0.531
	$ICC = 0.1$	0.189	0.079	0.526
	$ICC = 0.3$	0.159	0.081	0.517
Level 2	$J = 30$	0.312	0.208	0.554
	$J = 70$	0.318	0.161	0.546
	$J = 200$	0.323	0.127	0.554
	$n_j = 15$	0.347	0.180	0.554
	$n_j = 30$	0.286	0.156	0.548
	$ICC = 0.05$	0.329	0.150	0.547
	$ICC = 0.1$	0.333	0.156	0.574
	$ICC = 0.3$	0.290	0.196	0.532

	Conditions	MLM-Logistic		GAMM		MLM	
Random Effects		Bias	RMSE	Bias	RMSE	Bias	RMSE
τ_{00}	$J = 30$	0.131	0.155	-0.018	0.041	-0.007	0.038
	$J = 70$	0.116	0.126	-0.006	0.026	0.002	0.026
	$J = 200$	0.101	0.105	-0.001	0.016	0.006	0.018
	$n_j = 15$	0.109	0.122	-0.011	0.031	0.001	0.029
	$n_j = 30$	0.123	0.134	-0.006	0.025	0.000	0.025
	$ICC = 0.05$	0.099	0.104	-0.002	0.011	0.005	0.013
	$ICC = 0.1$	0.105	0.112	-0.005	0.018	0.002	0.018
	$ICC = 0.3$	0.143	0.169	-0.017	0.054	-0.006	0.051
τ_{11}	$J = 30$	0.044	0.068	-0.007	0.034	-0.007	0.034
	$J = 70$	0.045	0.056	-0.001	0.023	-0.001	0.023
	$J = 200$	0.042	0.046	0.000	0.014	0.000	0.014
	$n_j = 15$	0.043	0.058	-0.004	0.026	-0.003	0.026
	$n_j = 30$	0.045	0.056	-0.002	0.021	-0.002	0.021
	$ICC = 0.05$	0.052	0.063	-0.003	0.024	-0.002	0.024
	$ICC = 0.1$	0.047	0.059	-0.002	0.024	-0.002	0.024
	$ICC = 0.3$	0.032	0.048	-0.004	0.024	-0.004	0.024
τ_{01}	$J = 30$	0.074	0.086	0.019	0.348	0.022	0.295
	$J = 70$	0.070	0.074	0.004	0.189	0.004	0.174
	$J = 200$	0.063	0.064	0.003	0.105	0.001	0.100
	$n_j = 15$	0.068	0.075	0.009	0.256	0.014	0.216
	$n_j = 30$	0.070	0.075	0.008	0.173	0.004	0.163
	$ICC = 0.05$	0.071	0.074	0.023	0.260	0.019	0.214
	$ICC = 0.1$	0.070	0.074	-0.003	0.208	0.001	0.188
	$ICC = 0.3$	0.067	0.077	0.006	0.174	0.006	0.167
σ^2	$J = 30$	0.003	0.036	-0.004	0.036	0.001	0.036
	$J = 70$	0.003	0.023	-0.002	0.023	0.001	0.023
	$J = 200$	0.004	0.015	0.000	0.014	0.003	0.014
	$n_j = 15$	0.003	0.029	-0.003	0.029	0.000	0.029
	$n_j = 30$	0.004	0.020	-0.002	0.020	0.003	0.020
	$ICC = 0.05$	0.004	0.025	-0.002	0.024	0.002	0.024
	$ICC = 0.1$	0.004	0.024	-0.002	0.023	0.002	0.024
	$ICC = 0.3$	0.002	0.025	-0.003	0.025	0.001	0.025

Note. RMD indicates the mean the root mean squared difference between predicted values and true values across 500 replications.

Table 5: Simulation Study: Results for Fixed and Random Effects 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							
γ_{00}	$J = 30$	-0.004	0.094	1.091	-0.028	0.137	1.721
	$J = 70$	-0.005	0.070	0.980	0.054	0.112	1.633
	$J = 200$	-0.002	0.033	1.009	0.075	0.105	1.612
	$n_j = 15$	-0.005	0.079	1.045	0.022	0.144	1.546
	$n_j = 30$	-0.002	0.053	1.008	0.045	0.092	1.765
	$ICC = 0.05$	-0.002	0.048	1.016	0.011	0.114	1.951
	$ICC = 0.1$	0.004	0.058	1.056	0.026	0.105	1.687
	$ICC = 0.3$	-0.013	0.093	1.009	0.064	0.136	1.328
γ_{02}	$J = 30$	0.031	0.113	1.109	-0.011	0.145	1.429
	$J = 70$	0.014	0.077	0.981	0.024	0.066	1.151
	$J = 200$	0.006	0.043	0.979	0.030	0.058	1.250
	$n_j = 15$	0.031	0.107	1.054	0.012	0.125	1.381
	$n_j = 30$	0.003	0.049	0.992	0.017	0.055	1.173
	$ICC = 0.05$	0.014	0.063	1.026	0.056	0.091	1.397
	$ICC = 0.1$	0.016	0.076	1.051	-0.015	0.085	1.272
	$ICC = 0.3$	0.020	0.095	0.992	0.001	0.092	1.161
Random Effects							
τ_{00}	$J = 30$	0.031	0.074		0.218	0.233	
	$J = 70$	0.005	0.034		0.253	0.259	
	$J = 200$	-0.002	0.016		0.257	0.259	
	$n_j = 15$	0.022	0.052		0.209	0.218	
	$n_j = 30$	0.000	0.030		0.276	0.283	
	$ICC = 0.05$	0.004	0.018		0.260	0.262	
	$ICC = 0.1$	0.018	0.038		0.239	0.243	
	$ICC = 0.3$	0.011	0.067		0.229	0.246	
τ_{11}	$J = 30$	-0.003	0.035		0.001	0.038	
	$J = 70$	-0.002	0.023		0.005	0.026	
	$J = 200$	0.000	0.014		0.006	0.016	
	$n_j = 15$	-0.002	0.026		0.003	0.028	
	$n_j = 30$	-0.001	0.022		0.005	0.025	
	$ICC = 0.05$	-0.001	0.023		0.005	0.027	
	$ICC = 0.1$	-0.002	0.024		0.004	0.026	
	$ICC = 0.3$	-0.001	0.025		0.004	0.027	
τ_{01}	$J = 30$	-0.011	0.261		-0.094	0.278	
	$J = 70$	0.009	0.115		-0.035	0.155	
	$J = 200$	0.002	0.097		-0.058	0.106	
	$n_j = 15$	-0.001	0.164		-0.040	0.185	
	$n_j = 30$	0.000	0.151		-0.084	0.174	
	$ICC = 0.05$	0.003	0.167		-0.075	0.184	
	$ICC = 0.1$	-0.008	0.159		-0.065	0.179	
	$ICC = 0.3$	0.004	0.147		-0.047	0.175	
σ^2	$J = 30$	0.008	0.039		0.191	0.198	
	$J = 70$	0.004	0.025		0.132	0.136	
	$J = 200$	0.000	0.014		0.133	0.134	
	$n_j = 15$	0.005	0.031		0.111	0.117	
	$n_j = 30$	0.002	0.021		0.194	0.196	
	$ICC = 0.05$	0.004	0.026		0.175	0.178	
	$ICC = 0.1$	0.004	0.026		0.146	0.151	
	$ICC = 0.3$	0.004	0.026		0.135	0.140	

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.

Table 6: Simulation Study: Results of Basis Coefficient Estimates in GAMM, and RMD in GAMM and MLM under GAMM as a Data-Generating Model

Smooth Functions	Conditions	GAMM				MLM
		Bias	RMSE	Ratio	RMD	RMD
$f_2(x_{ij} - x_{.j})(z_j = 0)$	$J = 30$	0.155	0.511	0.938	0.126	0.398
	$J = 70$	0.147	0.397	0.945	0.084	0.361
	$J = 200$	0.032	0.192	0.980	0.045	0.439
	$n_j = 15$	0.154	0.405	0.940	0.110	0.365
	$n_j = 30$	0.069	0.329	0.968	0.071	0.433
	$ICC = 0.05$	0.116	0.381	0.964	0.089	0.380
	$ICC = 0.1$	0.110	0.369	0.952	0.095	0.410
	$ICC = 0.3$	0.108	0.350	0.945	0.095	0.408
$f_2(x_{ij} - x_{.j})(z_j = 1)$	$J = 30$	0.177	0.444	0.936	0.130	0.326
	$J = 70$	0.133	0.348	0.991	0.084	0.297
	$J = 200$	0.037	0.207	0.985	0.045	0.359
	$n_j = 15$	0.157	0.396	0.966	0.110	0.305
	$n_j = 30$	0.074	0.270	0.976	0.077	0.350
	$ICC = 0.05$	0.123	0.344	0.980	0.095	0.308
	$ICC = 0.1$	0.115	0.334	0.967	0.095	0.334
	$ICC = 0.3$	0.109	0.320	0.965	0.095	0.340
$f_1(x_{.j})(z_j = 0)$	$J = 30$	0.105	0.355	0.953	0.217	0.676
	$J = 70$	0.101	0.278	0.967	0.126	0.555
	$J = 200$	0.075	0.236	0.983	0.071	0.633
	$n_j = 15$	0.142	0.389	0.962	0.179	0.589
	$n_j = 30$	0.045	0.191	0.974	0.118	0.654
	$ICC = 0.05$	0.067	0.223	0.996	0.095	0.689
	$ICC = 0.1$	0.081	0.293	0.960	0.158	0.578
	$ICC = 0.3$	0.133	0.353	0.947	0.184	0.597
$f_1(x_{.j})(z_j = 1)$	$J = 30$	0.083	0.387	0.977	0.217	0.536
	$J = 70$	0.051	0.322	0.992	0.138	0.416
	$J = 200$	0.003	0.169	0.996	0.077	0.430
	$n_j = 15$	0.084	0.380	0.985	0.173	0.464
	$n_j = 30$	0.007	0.205	0.992	0.134	0.457
	$ICC = 0.05$	0.036	0.231	0.998	0.118	0.465
	$ICC = 0.1$	0.043	0.272	0.990	0.138	0.487
	$ICC = 0.3$	0.058	0.375	0.978	0.195	0.430

Note. For each smooth function, the bias, RMSE, and ratio reported are averaged across 9 basis coefficient estimates; RMD indicates the mean RMD across 500 replications; J indicates the number of clusters; n_j indicates a cluster size; ICC indicates an intraclass correlation coefficient.

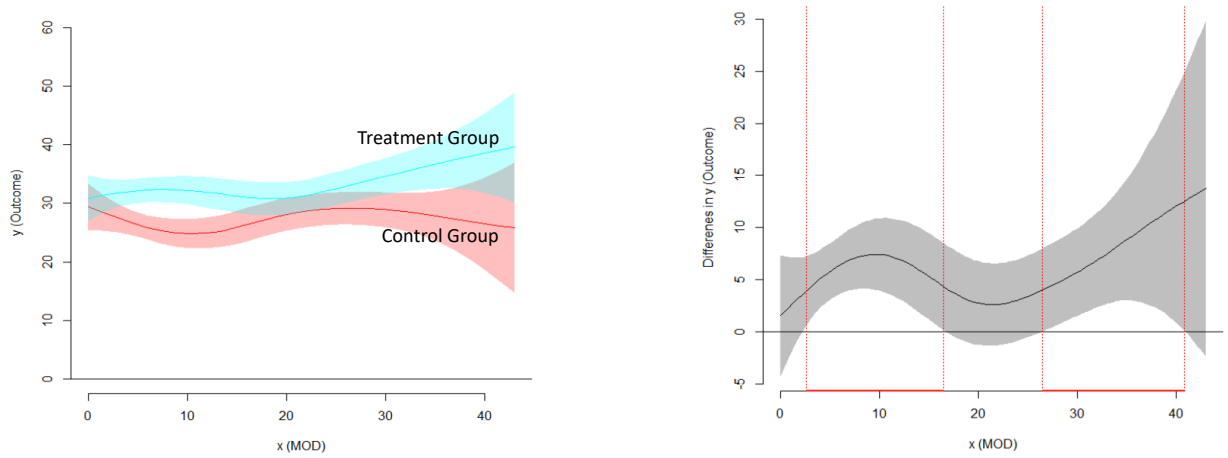


Figure 1. Varying effects of MOD x on outcome y by the TRT z modelled with smooth functions with confidence bands (left) and differences in outcome y between the two smooth functions (the smooth function for the treatment group - a smooth function for the control group) with confidence bands (right). Vertical lines in Figure 1 (right) indicate windows of significant differences.

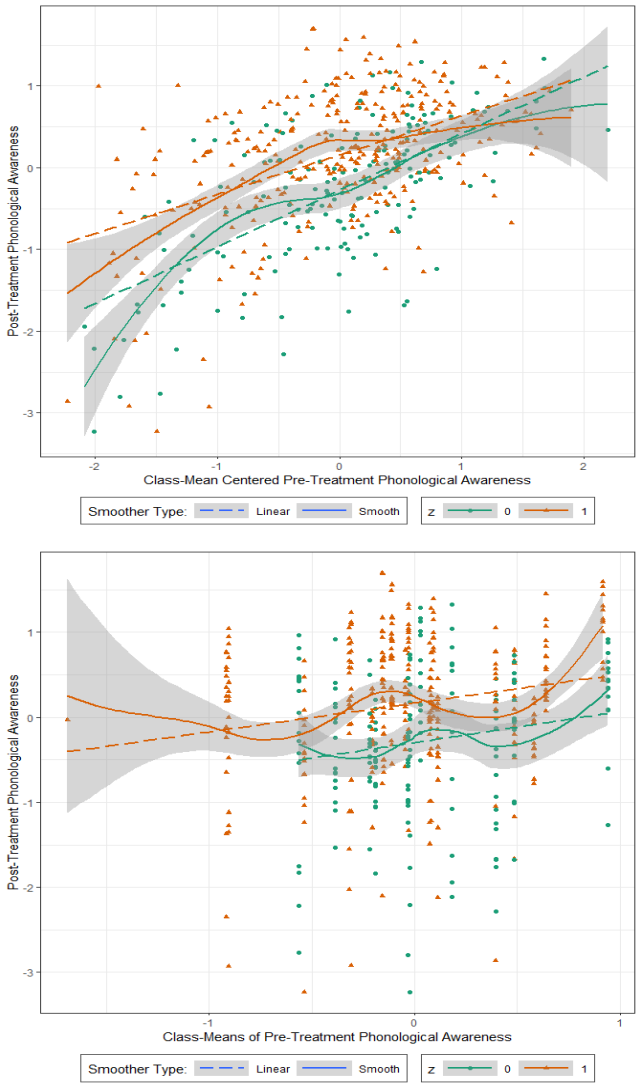


Figure 2. Empirical study: Scatter plots of y_{ij} vs. $x_{ij} - x_{\cdot j}$ by z_j (top) and y_{ij} vs. $x_{\cdot j}$ by z_j (bottom) (raw data).

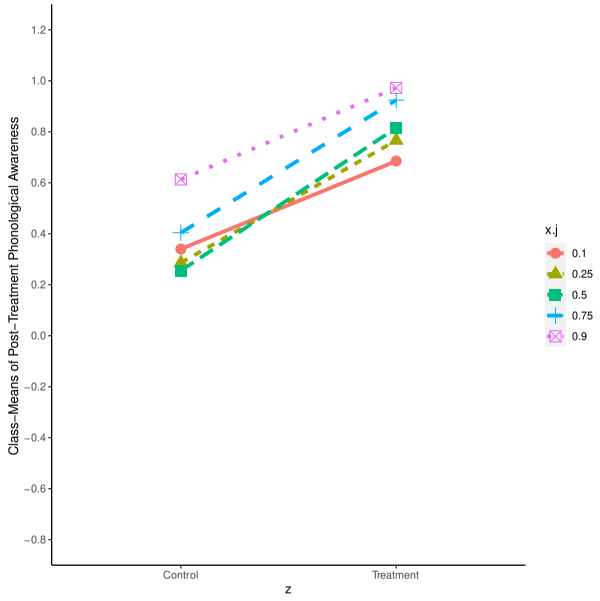
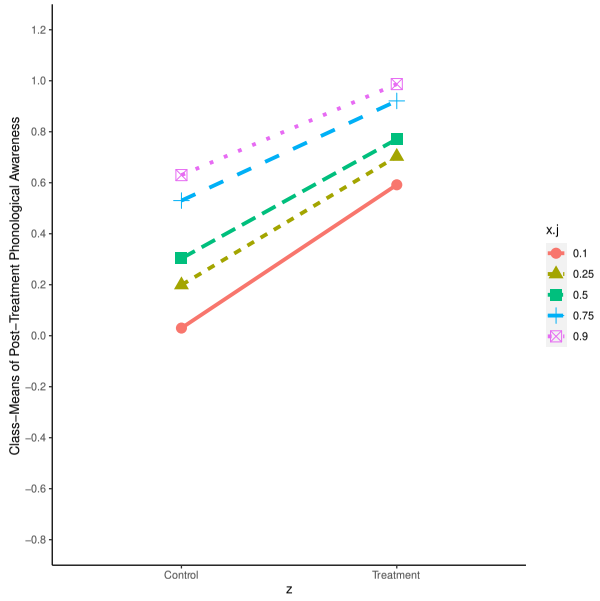
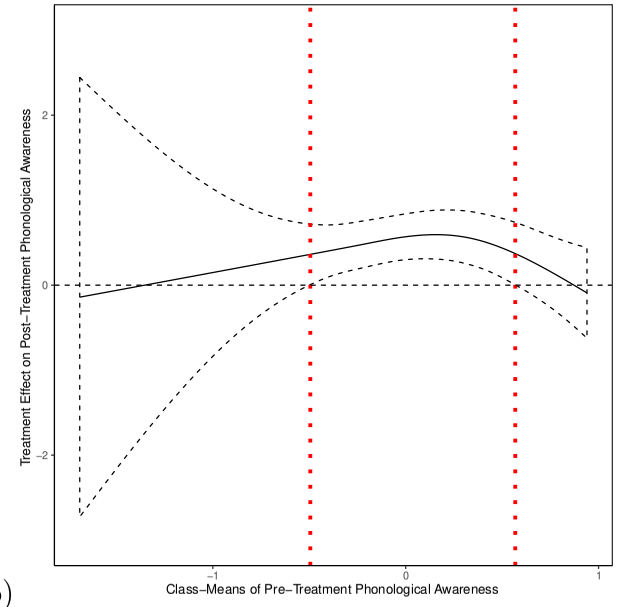
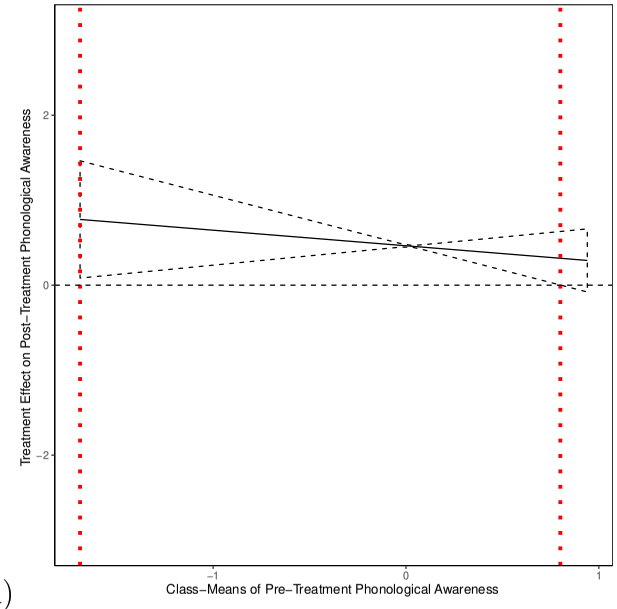
	GAMM	MLM
$y_{.j}$ vs. z_j by quantiles of $x_{.j}$	(a) 	(c) 
TRT effect vs. $x_{.j}$	(b) 	(d) 

Figure 3. Empirical study: Probing an interaction between a level-2 TRT (focal covariate) and a level-2 part of MOD for GAMM and MLM. Vertical lines in Figure 3 (b) and (d) indicate windows of significant differences.