Psychometrika (2026), **91**, 1151–1190
doi:10.1017/psy.2026.10108

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Mixed-Effects XGBoost with Group-Aware Permutation Importance and Cross-Validation for Multilevel Cross-Classified Continuous Outcomes

Sun-Joo Cho[1] and Sophia Mueller[2]

[1]Vanderbilt University, USA; [2]The University of North Carolina at Chapel Hill, USA

Corresponding author: Sun-Joo Cho; Email: sj.cho@vanderbilt.edu

(Received 7 December 2025; revised 21 March 2026; accepted 26 March 2026)

Abstract

This article proposes a mixed-effects machine-learning framework for modeling complex, nonlinear relations between predictors and continuous outcomes in multilevel cross-classified data. The proposed method, termed LMM–XGBoost, embeds extreme gradient boosting (XGBoost) within a linear mixed model (LMM) to combine flexible modeling of nonlinear and interaction effects with random effects that model dependence. In addition, an iterative estimation procedure for LMM–XGBoost is developed, a group-aware permutation importance measure that respects multilevel dependence is proposed, and a combined-group cross-validation (CV) strategy for hyperparameter tuning, out-of-fold (OOF) prediction, and importance estimation is developed for cross-classified designs. The simulation study shows that the proposed estimation method for LMM–XGBoost yields good parameter recovery under non-zero random-effect variances. In addition, relative to standard LMM and XGBoost, LMM–XGBoost achieves lower OOF prediction error and more accurate recovery of variable importance. The study further shows that combined-group CV and group-aware permutation importance yield less biased error estimates and substantially higher agreement with the true importance rankings than conventional permutation measures. An empirical application using the Add Health study illustrates how the proposed methods can identify important factors across multiple domains associated with adolescent depressive symptoms.

Keywords: cross-validation; importance measure; linear mixed model; variable selection; XGBoost

1. Introduction

1.1. Study motivation

In social and behavioral sciences, researchers use data collection strategies to represent the complexity of human behavior in real-world settings. A frequent design samples from multiple clusters, each reflecting a distinct subgroup of the target population (e.g., by demographic, geographic, or institutional characteristics). Many such datasets exhibit *multilevel cross-classified* structures in which individuals are simultaneously nested within more than one non-nested higher-level unit (e.g., Barker et al., 2020; Goldstein, 1994; Raudenbush & Bryk, 2002). For example, students can be cross-classified by schools and neighborhoods, so that each student is associated with a specific school–neighborhood combination. In these settings, outcomes typically display statistical *dependence* arising from shared contextual or structural influences within clusters (e.g., school policies, teacher engagement, neighborhood socioeconomic conditions, and community norms). Accurately estimating effects of both

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

individual-level and contextual-level variables therefore requires models that explicitly represent cross-classified clustering. Cross-classified multilevel models partition variance across multiple grouping factors, yielding more valid estimation and inferences and enhancing the generalizability of results obtained from observational, cross-sectional data (e.g., Raudenbush & Bryk, 2002).

Arguably, the most widely used approach for modeling dependence is the *linear mixed model* (LMM; Laird & Ware, 1982), which accounts for clustering-induced dependence through random effects. In the social and behavioral sciences, these models are also termed multilevel or hierarchical linear models. In many applications, a primary aim is to identify important predictors of an outcome across multiple social contexts (e.g., neighborhoods and schools; Harris & Halpern, 2022). LMMs accommodate both individual-level and cluster-level predictors to explain or predict variability in the outcome. Typical practice compares a set of candidate mixed-effects specifications that differ in the fixed-effect sets at each level, after which researchers interpret regression coefficients and test effects of substantive interest—procedures that remain standard in social and behavioral research.

Applying LMMs to investigate predictor–outcome relations presents challenges. First, functional form is often unknown: relations may be linear or nonlinear (e.g., logarithmic and exponential), and misspecification undermines model adequacy and interpretability. Second, predictors can interact in complex, non-additive ways, including higher-order interactions, which are difficult to specify a priori. Third, when the predictor set is large, identifying a parsimonious subset that meaningfully explains outcome variability is essential to avoid redundancy and overfitting while maintaining explanatory performance.

Machine-learning (ML) methods can capture complex, nonlinear associations between predictors and outcomes (e.g., Kuhn & Johnson, 2018). Among these, tree-ensemble approaches—random forests (RFs; Breiman, 2001) and gradient boosting (GB; Friedman, 2001)—aggregate many decision trees to improve predictive accuracy over a single tree (Breiman et al., 1984). By exploring diverse feature[1] subsets and splits, tree ensembles automatically accommodate nonlinearities and interactions. Moreover, RF and GB conduct implicit feature selection during training by preferentially splitting on informative variables, thereby elevating the influence of the most relevant predictors. Applications of RF and GB are increasingly common in the social and behavioral sciences (e.g., Jacobucci et al., 2023). A well-known limitation, however, is that standard ML implementations typically assume independent observations.

Taken together, current practice tends to rely on a limited set of prespecified effects within LMMs, with constrained support for systematic discovery of nonlinearities and interactions, limited scalability when many predictors are considered, and incomplete integration of data-adaptive diagnostics tailored to clustered designs. These limitations motivate complementary exploratory steps to enhance robustness and interpretability in exploratory analyses of multilevel data.

It is useful to clarify that the motivation for considering nonlinear and interaction effects in the present study is substantive. For example, in adolescent development, interactions (e.g., two level-1 predictors) and cross-level interactions are theoretically plausible and have been empirically documented, such as the moderating role of school connectedness in the association between bullying and aggressive or suicidal behaviors among sexual minority youth (e.g., Duong & Bradshaw, 2014).

1.2. Limitations of existing methods

Several modeling families address aspects of challenges of investigating predictor–outcome, yet each introduces practical tradeoffs. Parametric extensions (e.g., polynomial terms) require *a priori* choices about which predictors to model nonlinearly. Generalized additive mixed models (GAMMs; Hastie & Tibshirani, 1986) and regularized LMMs provide flexible or shrinkage-based alternatives but still require decisions about smoothness, basis specification, and interaction structure, and their performance hinges on tuning choices. Nonparametric, data-adaptive ensembles (e.g., RF and GB), which are commonly treated as ML methods, can recover higher-order interactions and complex effect shapes without

[1] In this study, the terms "predictors" and "features" are used interchangeably.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

prespecifying functional forms; however, they likewise require principled complexity control and careful validation to mitigate overfitting. Within this nonparametric, data-adaptive class, interpretable ML methods can help to identify influential predictors, characterize predictor–outcome relationships (linear, monotonic, or complex), and probe interactions (e.g., Apley & Zhu, 2020; Friedman & Popescu, 2008; Molnar, 2019) however, their use is not yet routine in exploratory workflows for multilevel data and often lacks cluster-aware implementation.

Because regular ML methods assume that all observations are independent, hybrid methods have been developed to incorporate ML techniques (e.g., tree-based or ensemble learning methods) into LMMs to account for dependencies due to clustering in multilevel data, as summarized in Table 1.[2] These hybrid methods leverage the strengths of ML while accounting for cluster-level dependencies in multilevel data. Simulation studies in prior work, as shown in Table 1, demonstrated that the hybrid models provided superior prediction accuracy compared to both standalone ML methods and LMMs. This prior work emphasized the importance of incorporating random effects to fully leverage the advantages of ML methods. However, as reviewed in Table 1, prior work has focused primarily on nested multilevel structures and has not addressed multilevel cross-classified structures. In addition, existing estimation methods are not directly applicable in this setting because prior work integrating ML with LMMs has primarily focused on tree-based or RF approaches under nested data structures (see Table 1). Furthermore, prior work has not demonstrated the interpretability of ML methods, such as importance measures, when they are used within hybrid approaches that combine ML techniques with LMMs.

Extreme gradient boosting (XGBoost; Chen & Guestrin, 2016) is a scalable implementation of gradient-boosted decision trees that sequentially fits shallow trees to the negative gradients (pseudo-residuals) of a specified loss function. Compared to earlier boosting frameworks (e.g., Friedman, 2001), XGBoost introduces systematic ℓ_1/ℓ_2 regularization on leaf weights, shrinkage and column/row subsampling to reduce variance, and a second-order Taylor approximation for line search. These design features typically yield robustness to multicollinearity and complex interactions, and favorable bias–variance tradeoffs relative to single trees, RF, or unregularized boosting, while remaining compatible with smooth loss functions for continuous outcomes. To our knowledge, no prior hybrid method combines XGBoost with LMM for continuous responses—particularly not in multilevel cross-classified settings with crossed random effects—nor do prior studies develop estimation methods that preserve the random-effects structure while leveraging XGBoost's regularized boosting and interaction modeling. Consequently, the interpretability layer specific to XGBoost (e.g., permutation importance aligned with clustered dependence) has also not been demonstrated within an LMM-ML hybrid framework.

ML methods, including XGBoost, are prone to memorize idiosyncratic patterns or noise in the training data. Cross-validation (CV) provides a principled way to quantify generalization error by evaluating model performance on held-out folds that were not used during training. Applying CV in multilevel cross-classified settings poses several challenges. Standard K-fold CV that randomly partitions individual observations breaks the multilevel structure, allowing units from the same cluster (e.g., school and neighborhood) to appear in both training and validation sets, which leads to information leakage and overly optimistic estimates of generalization error. In cross-classified designs, additional complications arise because each observation belongs to the cross of two or more higher-level units, making it nontrivial to define folds that simultaneously hold out entire combinations of clusters in a balanced way. Existing practice often relies on naïve observation-level CV or cluster-blocking on only one clustering factor (e.g., Roberts et al., 2017), both of which fail to fully respect the dependence structure and can misrepresent out-of-sample performance for new cross-classified structures (e.g., schools and neighborhoods). Moreover, widely used software defaults and tuning workflows rarely implement principled group- or combined-group CV for cross-classified multilevel data.

[2]In this review, we did not include hybrid methods for categorical responses that use generalized LMMs, such as those in Speiser et al. (2019), because their model specification and estimation methods are not directly applicable to continuous responses.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Table 1. Overview of literature reviews on LMM-ML for continuous outcomes

Paper	ML method	Interpretable ML?	Focused data structure	Application area
Segal (1992)	Tree	No	Nested longitudinal	Human immune virus (HIV) vaccine trial
De'Ath (2002)	Tree	No	Nested cross-sectional	Hunting spider
Galimberti and Montanari (2002)	Tree	No	Nested longitudinal	No application
Hajjem et al. (2011)	Tree	No	Nested cross-sectional	Box office revenues of movies
Sela and Simonoff (2012)	Tree	No	Nested longitudinal	Amazon web services; Accident fatality
Eo and Cho (2014)	Tree	No	Nested longitudinal	Wages
Hajjem et al. (2014)	RF	No	Nested cross-sectional	Box office revenues of movies
Fu and Simonoff (2015)	Tree	No	Nested longitudinal	Wages; Traffic fatality
Wang et al. (2016)	RF	No	Nested cross-sectional	Imputing multitissue expression
Fokkema et al. (2018)	Tree	No	Nested cross-sectional	Treatment–subgroup interaction detection
Capitaine et al. (2021)	RF	No*	Nested longitudinal	HIV vaccine trial
Krennmair and Schmid (2022)	RF	No	Nested cross-sectional	Small area estimation
Kilian et al. (2023)	Tree	No	Nested cross-sectional	Educational outcome
Salditt et al. (2023)	GB	No	Nested longitudinal	Wages

Note: RF indicates random forests; GB indicates gradient boosting; *Capitaine et al. (2021) used a variable selection procedure based on random forests, which is not typically categorized as an interpretable ML method in the sense of providing a formal variable importance measure.

1.3. Study purposes and novel contributions

The purpose of this article is to develop a mixed-effects ML framework for modeling complex, nonlinear relationships between predictors and outcomes in multilevel, cross-classified data with continuous outcomes. The study makes four methodological contributions: (a) integrating XGBoost into the LMM (hereafter, LMM–XGBoost); (b) developing an estimation procedure for LMM–XGBoost; (c) developing a group-aware permutation importance measure; and (d) proposing combined-group CV for hyperparameter tuning, out-of-fold (OOF) prediction, and permutation importance. The accompanying `R` (R Core Team, 2025) functions, which fully automate these four components, is publicly available on the Open Science Framework (OSF), https://osf.io/jkctg/.

LMM–XGBoost is designed to capture nonlinear and interaction effects among predictors (a core strength of ML) while accommodating random effects due to clustering (a core feature of LMMs). As noted above, LMM–XGBoost outputs can be used in exploratory stages to enhance robustness and interpretability in subsequent analyses. LMM–XGBoost is illustrated with an empirical analysis investigating factors across multiple domains associated with adolescent depressive symptoms, with the aim of informing targeted mental-health interventions. In addition, simulation studies evaluate (a) the benefits of modeling nonlinear and interaction fixed effects in LMM–XGBoost relative to a standard LMM, (b) the advantages of incorporating crossed random effects in LMM–XGBoost relative to XGBoost alone, and (c) parameter recovery for LMM–XGBoost together with the accuracy of the proposed group-aware permutation importance.

The remainder of this article is organized as follows. In Section 2, LMM–XGBoost is specified, following presentations of LMM and XGBoost, respectively. In addition, the estimation method and conditional group-aware permutation importance measure are presented. In Section 3, the performance of the LMM–XGBoost relative to competing methods (LMM and XGBoost), the parameter recovery of LMM–XGBoost, and the accuracy of the conditional group-aware permutation importance measure are evaluated through simulation. In Section 4, LMM–XGBoost is applied using the Add Health data (Harris et al., 2019), with comparisons to LMM and XGBoost. In Section 5, we conclude with a summary and a discussion.

2. Method

This section begins by specifying LMM for a cross-classified multilevel structure and XGBoost, respectively, followed by the hybrid model, LMM–XGBoost. Subsequently, the estimation method is described for LMM–XGBoost. In addition, the proposed group-aware permutation importance for LMM–XGBoost and combined-group CV methods are presented.

2.1. Model specification

2.1.1. LMM

LMM for a cross-classified multilevel structure (Goldstein, 1994) is written as

$$y_{ijk} = \beta_0 + \mathbf{x}_{ijk}\boldsymbol{\beta} + u_{1j} + u_{2k} + \varepsilon_{ijk}, \tag{1}$$

where y_{ijk} is the outcome for the i-th observation in cross-classified clusters j and k, β_0 is the fixed intercept, $\mathbf{x}_{ijk}$ is the matrix of predictors, $\boldsymbol{\beta}$ is the vector of fixed effects, u_{1j} is the random effect associated with cluster j, u_{2k} is the random effect associated with cluster k, and ε_{ijk} is the random residual. The u_{1j} and u_{2k} are assumed to be additive and independent random effects; $u_{1j} \sim \mathcal{N}(0, \sigma^2_{u1})$ and $u_{2k} \sim \mathcal{N}(0, \sigma^2_{u2})$ (Goldstein, 1994). The ε_{ijk} is assumed to follow $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

2.1.2. XGBoost

XGBoost (Chen & Guestrin, 2016) is a highly optimized, distributed implementation of GB. It is used to approximate an unknown target function $f(\mathbf{x})$ that may involve complex nonlinearities and high-order interactions among predictors. Within the boosting framework formulated by Friedman (2001), XGBoost constructs an additive ensemble of decision trees in a forward, stage-wise fashion. At boosting iteration m, a new tree $f_m(\mathbf{x})$ is appended to the current ensemble to further reduce residual error, so that after M iterations, the prediction is

$$\hat{y}^{(M)} = \sum_{m=1}^{M} f_m(\mathbf{x}). \tag{2}$$

The boosting iterations proceed until a convergence criterion is met or a prespecified stopping rule terminates the process.

XGBoost fits a single tree at each boosting iteration by optimizing a second-order Taylor expansion of the loss function with a structural regularization term. Specifically, at iteration m, the objective for the new tree $f_m(\mathbf{x}_n)$, for observation n ($n = 1, \ldots, N$; n indexes observational units, each uniquely determined by the triplet (i,j,k)), is

$$\mathcal{L}^{(m)} = \sum_{n=1}^{N} \left[g_n f_m(\mathbf{x}_n) + \frac{1}{2} h_n f_m^2(\mathbf{x}_n) \right] + \Omega(f_m), \tag{3}$$

where g_n and h_n denote the first- and second-order derivatives of the loss with respect to the current prediction, respectively, and $\Omega(f_m)$ controls the complexity of the newly added tree. The regularization term, which mitigates overfitting, is specified as

$$\Omega(f_m) = \kappa V + \frac{1}{2} \lambda \sum_{v=1}^{V} ||\omega_v||^2, \tag{4}$$

where v indexes terminal nodes (leaves), V is the number of leaves, ω_v is the prediction assigned to leaf v, κ penalizes tree size, and λ is an ℓ_2 (ridge) penalty on the leaf values. Optionally, an additional ℓ_1 (lasso) penalty, $\alpha \sum_{v=1}^{V} |\omega_v|$, may be included, with α shrinking some leaf outputs exactly to zero and thereby inducing sparsity.

To derive optimal leaf predictions and tree structure, the objective is simplified by aggregating gradient statistics at the leaf level. For a tree with V leaves, the per-tree objective (up to an additive constant) reduces to

$$\tilde{\mathcal{L}} = \sum_{v=1}^{V} \left[G_v \omega_v + \frac{1}{2} (H_v + \lambda) \omega_v^2 \right] + \kappa V, \tag{5}$$

where G_v and H_v are the sums of the first- and second-order derivatives over all observations assigned to leaf v. This formulation leads to closed-form solutions for the optimal leaf weights and yields an efficient split-gain criterion for evaluating candidate partitions of the predictor space.

Key hyperparameters include the learning rate (`eta`), maximum depth of trees (`max depth`), minimum child weight (`min child weight`), regularization coefficients (`lambda` [ℓ_2 penalty], `alpha` [ℓ_1 penalty]), minimum loss reduction required to create a split (`gamma`), subsampling fractions (`subsample`, `colsample bytree`), the number of boosting rounds (`num boost round`), and the number of early stopping rounds. For general introductions to tree-based boosting and the role of these hyperparameters, see Friedman (2001), Hastie et al. (2009, Chapters 9 and 10), and Chen and Guestrin (2016).

2.1.3. LMM–XGBoost

In LMM–XGBoost, the predicted values from XGBoost using an input $\mathbf{x}$ are incorporated into LMM, while allowing for random effects over cross-classified clusters. The LMM–XGBoost is presented as

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

follows:

$$y_{ijk} = \beta_0 + \mathrm{XG}(\mathbf{x}_{ijk}) + u_{1j} + u_{2k} + \varepsilon_{ijk}, \tag{6}$$

where y_{ijk} is the outcome for the i-th observation in cross-classified clusters j and k, β_0 is the fixed intercept, $\mathrm{XG}(\mathbf{x}_{ijk})$ denotes the predicted value from XGBoost based on predictors $\mathbf{x}_{ijk}$, u_{1j} is the random effect associated with cluster j, u_{2k} is the random effect associated with cluster k, and ε_{ijk} is the random residual. The u_{1j} and u_{2k} are assumed to be additive and independent random effects; $u_{1j} \sim \mathcal{N}(0, \sigma^2_{u1})$ and $u_{2k} \sim \mathcal{N}(0, \sigma^2_{u2})$ (Goldstein, 1994). The ε_{ijk} is assumed to follow $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

In Equation (6), the crossed random effects u_{1j} and u_{2k} are assumed to be uncorrelated with $\mathbf{x}_{ijk}$ and to account fully for intracluster correlation, as in LMMs. Unlike conventional LMMs, however, parametric random slopes for predictor effects are not included. In LMMs, heterogeneity in predictor effects across clusters is typically modeled through random slopes, which represent unexplained cluster-specific deviations in the linear effects of predictors. In contrast, the LMM–XGBoost framework assumes that part of the apparent slope heterogeneity arises from systematic (non)linearities, interactions, or higher-order relationships among predictors. Such systematic heterogeneity can be flexibly captured by the XGBoost component, which learns complex predictor–outcome relationships without imposing a prespecified parametric form. Under this perspective, variability in predictor effects across clusters is partly explained through the ML component, while remaining unexplained cluster-level variation is captured through the random intercepts.

Equation (6) cannot be estimated using a standard LMM estimation procedure because the fixed-effects component is replaced by a nonparametric function learned via XGBoost. Standard LMM estimation relies on a linear predictor with a finite set of parameters, whereas a boosted tree ensemble is an algorithmically estimated function whose structure is not parameterized within the LMM likelihood.

2.2. Estimation methods

This section first outlines two key components of the iterative LMM–XGBoost algorithm and then presents the full algorithm. The complete workflow is summarized in Algorithm 1.

2.2.1. Working residuals

The LMM in LMM–XGBoost assumes a Gaussian conditional distribution with an identity link. In this setting, the general iteratively reweighted least squares (IRLS)/Newton machinery collapses to a single weighted least-squares step for the conditional mean.

The fixed-effects component in LMM–XGBoost consists only of an intercept, $\boldsymbol{\beta} = (\beta_0)$. Let n index an observation, uniquely identified by the triplet (i,j,k). For notational convenience, the multiple indices are replaced by the single observation index n. In the matrix formulation used for estimation, observations are stacked into a single response vector, and cluster membership is represented through the random-effect design matrices. The linear predictor of LMM–XGBoost is rewritten as

$$\eta_n = o_n + \beta_0 + \mathbf{z}_n^\top \mathbf{u}, \tag{7}$$

where o_n is a fixed, known offset $o_n = \mathrm{XG}(\mathbf{x}_n)$, $\mathbf{u}$ is the vector of crossed random effects ($\mathbf{u}_1$ and $\mathbf{u}_2$), and $\mathbf{z}_n^\top$ is the row of the design matrix corresponding to observation n.

The conditional log-likelihood for observation n is

$$\ell_n(\eta_n, \sigma^2) = -\tfrac{1}{2}\log(2\pi\sigma^2) - \frac{(y_n - \eta_n)^2}{2\sigma^2}. \tag{8}$$

The score (U_n) and the negative Hessian (H_n) with respect to the linear predictor η_n are

$$U_n = \frac{\partial \ell_n}{\partial \eta_n} = \frac{y_n - \eta_n}{\sigma^2}, \qquad H_n = -\frac{\partial^2 \ell_n}{\partial \eta_n^2} = \frac{1}{\sigma^2}. \tag{9}$$

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

A single Newton/IRLS step yields the working response and weight

$$e_n^{(t)} = \eta_n^{(t)} + \frac{U_n^{(t)}}{H_n^{(t)}} = \eta_n^{(t)} + \left(y_n - \eta_n^{(t)}\right) = y_n, \qquad \omega_n^{(t)} = H_n^{(t)} = \frac{1}{\sigma^2}. \tag{10}$$

Thus, for the homoscedastic Gaussian model, the working response is simply the raw outcome ($\mathbf{e} = \mathbf{y}$), and the precision weight is constant.

With the offset $\mathbf{o} = \mathrm{XG}(\mathbf{x})$, and the fixed-effect design matrix $\mathbf{L}$ being a column vector of ones (corresponding to β_0), the estimation proceeds by generalized least squares (GLS) on the shifted outcome:

$$\mathbf{y}^* = \mathbf{y} - \mathbf{o}. \tag{11}$$

The marginal covariance remains

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^\top + \sigma^2\mathbf{I}, \tag{12}$$

where $\mathbf{G} = \mathrm{blockdiag}\left(\sigma_{u1}^2\mathbf{I}_J\,,\ \sigma_{u2}^2\mathbf{I}_K\right)$. The Newton/GLS step for the fixed effects $\boldsymbol{\beta}$ solves the normal equations:

$$\left(\mathbf{L}^\top\mathbf{V}^{-1}\mathbf{L}\right)\Delta\boldsymbol{\beta} = \mathbf{L}^\top\mathbf{V}^{-1}\left(\mathbf{y}^* - \mathbf{Z}\widetilde{\mathbf{u}}\right), \tag{13}$$

where $\widetilde{\mathbf{u}}$ are the conditional modes of the random effects.

The $\mathbf{y}^*$ term is modeled by $\mathbf{L}\boldsymbol{\beta} + \mathbf{Z}\widetilde{\mathbf{u}}$. Since $\mathbf{L}\boldsymbol{\beta} = \beta_0\mathbf{1}$ (where $\mathbf{1}$ is a vector of ones), the *working residual* is

$$\mathbf{r} = \mathbf{y}^* - \left(\mathbf{L}\boldsymbol{\beta} + \mathbf{Z}\widetilde{\mathbf{u}}\right) = \mathbf{y} - \mathbf{o} - \beta_0\mathbf{1} - \mathbf{Z}\widetilde{\mathbf{u}}. \tag{14}$$

This setup corresponds to fitting a model where the fixed-effect component only includes the intercept β_0. The offset $\mathbf{o}$ captures the known predictions from XGBoost. The working residual $\mathbf{r}$ captures the variation remaining after subtracting the offset, the estimated intercept, and the predicted random effects.

2.2.2. The C-projection

The function $f(\mathbf{x})$, learned by XGBoost on working residuals, can inadvertently absorb group-mean structure already accounted for by the cross-classified random effects ($\mathbf{u}_1$ and $\mathbf{u}_2$). To enforce a clean separation between the random effects and the residual learner, we apply a weighted C-projection to the learner's prediction vector, removing the span of the combined group dummies.

In the LMM context with a Gaussian outcome, estimation is non-iterative and the IRLS precision weight is constant:

$$\omega_n = \frac{1}{\sigma^2} \quad \text{for all } n. \tag{15}$$

Because this scalar factor is common to all terms in the projection, we may set the weight matrix proportional to the identity and, without loss of generality, take $\mathbf{W} = \mathbf{I}$ for the projection step.

Let $\mathbf{Z}_1$ be the dummy (indicator) matrix for cluster j (random intercept u_{1j}), and $\mathbf{Z}_2$ be the dummy matrix for cluster k (random intercept u_{2k}). Define the combined group-dummy matrix

$$\mathbf{Z} = \left[\mathbf{Z}_1\ \mathbf{Z}_2\right]. \tag{16}$$

Let $\mathbf{X}_g$ collect all group-constant predictors whose signal we intend to preserve (i.e., predictors constant within cluster j or within cluster k).

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Define the residual-maker that partials out $\mathbf{X}_g$:

$$\mathbf{M}_g \;=\; \mathbf{I} \;-\; \mathbf{X}_g\big(\mathbf{X}_g^\top \mathbf{X}_g\big)^{+}\mathbf{X}_g^\top, \qquad \tilde{\mathbf{Z}} \;=\; \mathbf{M}_g\mathbf{Z}, \tag{17}$$

where $(\cdot)^+$ denotes the Moore–Penrose pseudoinverse.

Let $\tilde{\mathbf{f}}$ denote the raw XGBoost prediction vector (i.e., $\tilde{\mathbf{f}} = \mathrm{XG}(\mathbf{x})$ in (6)). The (weighted) C-projection in the Gaussian case reduces to the unweighted projection

$$\tilde{\mathbf{f}}_\perp \;=\; \Big(\mathbf{I} - \tilde{\mathbf{Z}}\big(\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}}\big)^{+}\tilde{\mathbf{Z}}^\top\Big)\tilde{\mathbf{f}}. \tag{18}$$

This construction makes $f(\mathbf{x})$ orthogonal (in the Euclidean sense) to the combined group-dummy space spanned by the columns of $\mathbf{Z}$ while preserving the variation conveyed by the designated group-constant predictors $\mathbf{X}_g$. As a result, the LMM's cross-classified structure is respected, and the effects of group-constant predictors are not inadvertently removed from the residual learner. A numerical illustration comparing no projection, naive Z-projection, and the proposed C-projection is provided in Appendix A.

2.2.3. Algorithm

We implement an "E/M-style" alternating procedure in which a nonparametric learner updates the offset (E-step), and a mixed model updates the variance components and fixed intercept (M-step).

Initialization: Set $f^{(0)}(\mathbf{x}) \equiv 0$, $\mathbf{o}^{(0)} = \mathbf{0}$. Fit an initial LMM with offset $\mathbf{o}^{(0)}$ to obtain $\hat{\beta}_0^{(0)}$, $\widetilde{\mathbf{u}}^{(0)}$, and variance components $\widehat{\boldsymbol{\theta}}^{(0)} = (\hat{\sigma}^2, \hat{\sigma}^2_{u1}, \hat{\sigma}^2_{u2})$. Set $t = 0$.

E-step (nonparametric offset update): Form the working residual under the current fit

$$\mathbf{r}^{(t)} \;=\; \mathbf{y} \;-\; \mathbf{o}^{(t)} \;-\; \hat{\beta}_0^{(t)}\mathbf{1} \;-\; \mathbf{Z}\widetilde{\mathbf{u}}^{(t)},$$

where the superscript (t) indicates the current estimates at iteration t. Train XGBoost (Gaussian loss) on $\{(\mathbf{x}_n, r_n^{(t)})\}$ to obtain predictions $\tilde{\mathbf{f}}^{(t)} = \tilde{f}^{(t)}(\mathbf{x})$ using the `xgboost` function in R implementation (Chen & Guestrin, 2016; Chen et al., 2025). In this function, the histogram-based tree construction algorithm is employed, which provides a scalable approximation to exact greedy split finding. We enforce orthogonality using the C-projection defined in Section 2.2.2

Globally center the learner to keep the overall intercept in the fixed part:

$$\bar{f}^{(t)} \;=\; \frac{1}{N}\mathbf{1}^\top\tilde{\mathbf{f}}_\perp^{(t)}, \qquad \mathbf{f}^{(t)} \;=\; \tilde{\mathbf{f}}_\perp^{(t)} - \bar{f}^{(t)}\mathbf{1}.$$

M-step (mixed-model update): With $\mathbf{y}^* = \mathbf{y} - \mathbf{o}^{(t+1)}$, refit the LMM to obtain $\widehat{\beta}_0^{(t+1)}$, $\widehat{\sigma}^{2(t+1)}_{u1}$, $\widehat{\sigma}^{2(t+1)}_{u2}$, and $\widehat{\sigma}^{2(t+1)}$. For the homoscedastic LMM, IRLS reduces to GLS on $\mathbf{y}^*$ with

$$\big(\mathbf{L}^\top\mathbf{V}^{-1}\mathbf{L}\big)\,\Delta\boldsymbol{\beta} \;=\; \mathbf{L}^\top\mathbf{V}^{-1}\big(\mathbf{y}^* - \mathbf{Z}\widehat{\mathbf{u}}\big), \quad \mathbf{L} = \mathbf{1}.$$

The M-step is implemented using the `lmer` function from the `lme4` package (Bates et al., 2015).

Learning rate α_L: At each iteration, the offset is updated by a convex combination of its previous value and the centered, C-projected learner increment:

$$\mathbf{o}^{(t+1)} \;=\; (1-\alpha_L)\,\mathbf{o}^{(t)} \;+\; \alpha_L\,\mathbf{f}^{(t)}, \qquad \alpha_L \in (0,1].$$

The parameter α_L acts as a step size controlling how aggressively the learner's contribution is assimilated. When $\alpha_L = 1$, the update fully incorporates the new increment (fast but potentially unstable); as $\alpha_L \to 0$, the update becomes more conservative, smoothing the sequence $\{\mathbf{o}^{(t)}\}$ and reducing overfitting from any single iteration. Unfolding the recursion shows that $\mathbf{o}^{(t)}$ is an exponentially weighted average of

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

past increments,

$$\mathbf{o}^{(t)} = (1-\alpha_L)^t \mathbf{o}^{(0)} + \alpha_L \sum_{s=0}^{t-1} (1-\alpha_L)^{t-1-s} \mathbf{f}^{(s)},$$

so recent increments receive higher weight. Note that α_L scales only the assimilation of the fitted learner into the offset; it does not change the working residuals used to train the learner at iteration t. In practice, α_L plays a role analogous to the shrinkage parameter in boosting and can be chosen alongside XGBoost's own learning-rate and early-stopping settings.

A possible alternative approach would be to first remove cluster effects using a fixed-effects model and then apply XGBoost to the residualized outcome. However, this strategy differs from the proposed LMM–XGBoost framework in several important respects. First, residualizing the outcome using fixed effects removes cluster-level variation but does not estimate variance components associated with random effects or allow conditional predictions based on cluster-specific effects. Second, such preprocessing treats the clustering adjustment as a one-step procedure, whereas the proposed algorithm alternates between updating the ML component and estimating random effects. This iterative structure allows the nonlinear predictor learned by XGBoost and the random-effect estimates to be refined jointly during estimation. Third, the proposed framework naturally accommodates cross-classified clustering structures, whereas fixed-effects preprocessing would require introducing a large number of dummy variables and may become cumbersome in such settings. For these reasons, the proposed method provides a more integrated framework for combining ML with multilevel modeling than residualization followed by standalone XGBoost.

Convergence: A simpler rule such as checking the absolute or relative change in consecutive log-likelihoods of the fitted LMM, $\ell^{(t)} - \ell^{(t-1)}$, against a tolerance ε is often used but can be unreliable in the LMM–XGBoost algorithm. This method may halt prematurely after a transient small change due to noise in the XGBoost fit during the E-step or approximation error in the M-step's variance component estimation, and it may fail to detect mild oscillations that indicate a lack of convergence. The iterative LMM–XGBoost algorithm employs a more robust method, checking for stability over a window of recent iterations, which is first applied after a user-configured starting iteration (e.g., the fourth iteration). The window size (e.g., the last few log-likelihood values) is a user-configurable parameter, as is the starting iteration for this check, to make the convergence criterion more robust to transient fluctuations. The algorithm terminates the iteration if the following condition holds: $t \geq 4$ and $\mathrm{SD}\left(\ell^{(t-4)}, \ldots, \ell^{(t)}\right) < \varepsilon$, where ε is a user-specified tolerance (e.g., the standard deviation [SD] of the last three log-likelihoods). Otherwise, set $t \leftarrow t+1$ and repeat.

2.3. Group-aware permutation importance for LMM–XGBoost

Permutation importance is defined as the loss inflation that occurs when the empirical link between one predictor and the fitted learner is destroyed while all other quantities are held fixed. Larger loss inflation implies a greater contribution of that predictor to predictive accuracy. Because the design is cross-classified, permutations respect grouping: (a) predictors that vary at the individual level are shuffled within each (Cluster1 × Cluster2 id cell; i.e., permuted among rows sharing the same school and neighborhood) and (b) predictors that are constant within a cluster (e.g., school- or neighborhood-constant features, including elements of $\mathbf{x}_g$) are permuted at the corresponding cluster level and the permuted values are copied to all rows in that cluster. This procedure breaks the predictor–outcome association without fabricating cross-level combinations, thereby avoiding the bias of naïve global (row-wise) shuffling.

The residual learner is trained on the working residuals (Equation (14)). Let $\tilde{f}(\mathbf{x})$ denote the final centered and C-projected learner. Baseline predictions and loss on the evaluation data are

$$\hat{r}_n^{(0)} = \tilde{f}(\mathbf{x}_n), \qquad M_0 = \mathrm{MSE}(\tilde{f}) = \frac{1}{N}\sum_{n=1}^{N} \left(r_n - \hat{r}_n^{(0)}\right)^2,$$

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

where MSE stands for mean squared error. During permutation, the fitted LMM–XGBoost (including C-projection and global centering applied to $\tilde{f}$) is kept fixed; only the chosen predictor is permuted.

For predictor x_q (where q indexes predictors), let π_q be a group-aware permutation operator applied to x_q on the evaluation data. The permuted predictions and loss are

$$\hat{r}_n^{(\pi_q)} = \tilde{f}\left(\mathbf{x}_n^{\pi_q}\right), \qquad M_q = \frac{1}{N}\sum_{n=1}^{N}\left(r_n - \hat{r}_n^{(\pi_q)}\right)^2.$$

Importance is the average loss increase over D independent permutations:

$$\widehat{\Delta}_q = \frac{1}{D}\sum_{d=1}^{D}\left(M_{q,d} - M_0\right).$$

Algorithm 1 Iterative LMM-XGBoost.

Require: Responses $\{y_n\}_{n=1}^N$ with predictors $\mathbf{x}_n$; groupings (j,k) with design matrices $\mathbf{Z}_1, \mathbf{Z}_2$ and $\mathbf{Z} = [\mathbf{Z}_1\ \mathbf{Z}_2]$; group-constant design $\mathbf{X}_g$ to be preserved; tolerance ε (SD of recent log-likelihoods); learning rate $\alpha_L \in (0,1]$; maximum iterations $T_{\max}$.

Ensure: Estimates $\left(\hat{\beta}_0, \hat{\sigma}^2, \hat{\sigma}^2_{u1}, \hat{\sigma}^2_{u2}\right)$ and a centred, C-projected offset $\mathbf{o}$.

Initialize $\mathbf{o}^{(0)} \leftarrow \mathbf{0}$; $t \leftarrow 0$.

Fit an **initial** Gaussian LMM with offset $\mathbf{o}^{(0)}$:

$$\mathbf{y} = \beta_0^{(0)}\mathbf{1} + \mathbf{Z}\widetilde{\mathbf{u}}^{(0)} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}),$$

to obtain $\widehat{\beta}_0^{(0)}, \widetilde{\mathbf{u}}^{(0)}$, and variance components $\widehat{\boldsymbol{\theta}}^{(0)} = (\hat{\sigma}^2, \hat{\sigma}^2_{u1}, \hat{\sigma}^2_{u2})$.

Compute the initial working residual

$$\mathbf{r}^{(0)} = \mathbf{y} - \widehat{\beta}_0^{(0)}\mathbf{1} - \mathbf{Z}\widetilde{\mathbf{u}}^{(0)}.$$

Tune XGBoost hyperparameters using combined K-fold CV with outcome $\mathbf{r}^{(0)}$.

while $t < T_{\max}$ **do**

E-step: update the offset with XGBoost

Form the working residuals

$$\mathbf{r}^{(t)} = \mathbf{y} - \mathbf{o}^{(t)} - \widehat{\beta}_0^{(t)}\mathbf{1} - \mathbf{Z}\widetilde{\mathbf{u}}^{(t)}.$$

Fit XGBoost (squared-error loss) to $\{(\mathbf{x}_n, r_n^{(t)})\}_{n=1}^N$; obtain raw predictions $\tilde{\mathbf{f}}^{(t)} = \left(\tilde{f}^{(t)}(\mathbf{x}_n)\right)_{n=1}^N$.

Compute the C-projection that preserves group-constant signal in $\mathbf{X}_g$:

$$\mathbf{M}_g = \mathbf{I} - \mathbf{X}_g(\mathbf{X}_g^\top\mathbf{X}_g)^+\mathbf{X}_g^\top, \quad \tilde{\mathbf{Z}} = \mathbf{M}_g\mathbf{Z}, \quad \tilde{\mathbf{f}}_\perp^{(t)} = \left(\mathbf{I} - \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}})^+\tilde{\mathbf{Z}}^\top\right)\tilde{\mathbf{f}}^{(t)} \quad \text{(cf. (18))}.$$

Global centering to keep the overall mean in the fixed intercept:

$$\tilde{f}^{(t)} = \tfrac{1}{N}\mathbf{1}^\top\tilde{\mathbf{f}}_\perp^{(t)}, \qquad \mathbf{f}^{(t)} = \tilde{\mathbf{f}}_\perp^{(t)} - \tilde{f}^{(t)}\mathbf{1}.$$

Offset update with learning rate (convex blend):

$$\mathbf{o}^{(t+1)} = (1-\alpha_L)\,\mathbf{o}^{(t)} + \alpha_L\,\mathbf{f}^{(t)}.$$

M-step: update LMM parameters with the new offset

Refit the Gaussian LMM with offset $\mathbf{o}^{(t+1)}$ (e.g., REML/GLS):

$$\mathbf{y} = \beta_0^{(t+1)}\mathbf{1} + \mathbf{Z}\widetilde{\mathbf{u}}^{(t+1)} + \text{offset}\left(\mathbf{o}^{(t+1)}\right) + \boldsymbol{\varepsilon}.$$

Record $\ell^{(t+1)} = \log L\left(\widehat{\beta}_0^{(t+1)}, \widetilde{\mathbf{u}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)}; \mathbf{o}^{(t+1)}\right)$.

Convergence check: if $t \geq 4$ and $\text{SD}\left(\ell^{(t-4)}, \ldots, \ell^{(t+1)}\right) < \varepsilon$ then break.

$t \leftarrow t+1$

end while

return $\left(\hat{\beta}_0, \hat{\sigma}^2, \hat{\sigma}^2_{u1}, \hat{\sigma}^2_{u2}, \mathbf{o}^{(t)}\right)$.

A nonnegative, normalized score is reported for readability:

$$\text{RelImp}_q = \frac{\max\left(\widehat{\Delta}_q, 0\right)}{\sum_j \max\left(\widehat{\Delta}_j, 0\right)}.$$

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Sampling variability across permutations is summarized by

$$\Delta_{q,\mathrm{sd}} = \sqrt{\frac{1}{D-1}\sum_{d=1}^{D}\left[(M_{q,d} - M_0) - \widehat{\Delta}_q\right]^2}.$$

In addition to full-sample permutation importance, generalization can be assessed with combined group K-fold CV. For each fold: (a) fit the full LMM–XGBoost pipeline on the training folds only (including tuning, iterative offset updates with learning rate α, C-projection, and global centering); (b) score the held-out fold to obtain $M_0^{(\mathrm{fold})}$; and (c) apply group-aware permutations of x_q *within the held-out fold* to compute $M_{q,d}^{(\mathrm{fold})}$ and $\Delta_{q,d}^{(\mathrm{fold})} = M_{q,d}^{(\mathrm{fold})} - M_0^{(\mathrm{fold})}$. The OOF importance aggregates across folds and permutations:

$$\widehat{\Delta}_q^{\mathrm{OOF}} = \frac{1}{FR}\sum_{\mathrm{fold}=1}^{F}\sum_{d=1}^{D}\Delta_{q,d}^{(\mathrm{fold})},$$

with the same truncation and normalization used for RelImp_q.

2.4. Combined group cross-validation

For tuning XGBoost in LMM–XGBoost, as well as for OOF prediction and permutation importance, the combined-group CV strategy was for cross-classified data structures, where observations are simultaneously nested within two higher-level clusters (e.g., schools and neighborhoods). Instead of grouping by only one clustering factor, this approach defines each unique pair of clusters—such as each specific school–neighborhood combination—as a distinct cross-classified cell and treats it as an indivisible unit during CV. All observations within the same cell are assigned to the same fold, ensuring that no cell contributes data to both the training and test sets. This blocking structure preserves dependencies arising from both clustering dimensions and prevents information leakage across either level.

3. Simulation study

In this section, simulation studies are presented to (a) evaluate the relative prediction accuracy of LMM–XGBoost, LMM, and XGBoost; (b) examine parameter recovery for LMM–XGBoost; and (c) assess the accuracy of group-aware permutation importance for LMM–XGBoost by comparing it with conventional permutation importance (naïve global shuffling) from XGBoost and LMM.

3.1. Simulation design and expected results

The varying simulation factors of variances of random effects and the number of clusters were designed to evaluate the relative performance of LMM–XGBoost versus LMM (linear effects only and no-interactions) and LMM–XGBoost versus XGBoost. Main and linear effects were considered for LMM because it serves as a baseline model for comparison. This choice does not imply that LMMs cannot handle interactions or nonlinearities; rather, it establishes a standard reference point to assess how data-driven methods, such as LMM–XGBoost, enhance explanatory performance over a traditionally specified LMM. Specifically, outcomes were generated with the following levels of the two varying simulation factors:

- Variances of random effects (ICCs): no, small, medium, and large.[3]
  For the no-variance level, the values of σ_{u1} and σ_{u2} were both set to 0. While this scenario may be uncommon in practice, it was included to illustrate the maximal performance of XGBoost and the potential issue of boundary solutions (0 variances) in random effects for LMM and

[3]The term "small" does not refer to an absolute size of residual variances. In practice, it may represent a medium size of variances. Both small and large variances should be interpreted relatively, rather than absolutely, within the context of a simulation condition.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

LMM–XGBoost. For the small variance level, $\sigma_{u1} = \sigma_{u2} = 0.130$. For the chosen $\sigma = 0.551$, this setting yields per-cluster ICCs $\frac{\sigma^2_{u1}}{\sigma^2_{u1}+\sigma^2_{u2}+\sigma^2} = \frac{\sigma^2_{u2}}{\sigma^2_{u1}+\sigma^2_{u2}+\sigma^2} = 0.05$. For the medium variance level, $\sigma_{u1} = \sigma_{u2} = 0.255$, yielding per-cluster ICCs of 0.15: $\frac{\sigma^2_{u1}}{\sigma^2_{u1}+\sigma^2_{u2}+\sigma^2} = \frac{\sigma^2_{u2}}{\sigma^2_{u1}+\sigma^2_{u2}+\sigma^2} = 0.15$. For the large variance level, $\sigma_{u1} = \sigma_{u2} = 0.390$, yielding per-cluster ICCs of 0.25: $\frac{\sigma^2_{u1}}{\sigma^2_{u1}+\sigma^2_{u2}+\sigma^2} = \frac{\sigma^2_{u2}}{\sigma^2_{u1}+\sigma^2_{u2}+\sigma^2} = 0.25$. These ICC magnitudes are consistent with values reported in applications of LMM or multilevel models (e.g., Snijders & Bosker, 2012).

- The number of clusters: small versus large.

  The simulation study conditions for the number of clusters (J and K) and observations per cell (I_{jk}) were selected based on established multilevel literature. The number of clusters for both factors was set at two values, $J = K = 50$ and $J = K = 100$. The former represents the minimum acceptable clusters often cited to mitigate bias in variance-component estimates (σ^2_{u1} and σ^2_{u2}) (Maas & Hox, 2004), whereas the latter serves as a benchmark for maximizing estimator stability and accuracy (Clarke & Wheaton, 2007). The number of observations per cell, I_{jk}, was fixed at a constant, moderate value of 5. This choice ensures an adequate Level-1 sample size to estimate the Level-1 residual variance (σ^2) across all cluster combinations and satisfies the identification requirement $I_{jk} > 1$.

The data-generating model is an LMM-tree (one "true" tree) under two varying simulation conditions. At Level-1, 10 continuous predictors were sampled from a multivariate normal distribution $\mathcal{MVN}(\mathbf{0}, \Sigma_{L1})$, where Σ_{L1} had unit variances and off-diagonal correlations of 0.2. A Cholesky decomposition of the correlation matrix was used to ensure positive definiteness. In addition, two Level-1 categorical predictors were generated via multinomial draws: one with two levels $(0.5, 0.5)$ and one with three levels $(0.2, 0.5, , 0.3)$. For the Cluster1-level predictors, three continuous variables were sampled from $\mathcal{MVN}(\mathbf{0}, \Sigma_{C1})$ with unit variances and pairwise correlations of 0.2, and one categorical predictor with two levels $(0.5, , 0.5)$ was generated via a multinomial draw. For the Cluster2-level predictors, three continuous variables were sampled from $\mathcal{MVN}(\mathbf{0}, \Sigma_{C2})$ with unit variances and pairwise correlations of 0.2, and one categorical predictor with three levels $(0.2, 0.5, , 0.3)$ was generated analogously. Cluster-level predictors were realized once per cluster and were identical for all observations sharing the corresponding cluster identifier (i.e., invariant within Cluster1 or Cluster2). The true tree was generated using 30 randomly selected variables drawn from the 20 continuous predictors and 37 $(= 49 - 12)$ dummy predictors (from 12 generated categorical predictors, except the first level of each categorical predictor) predictors were considered. The decision tree was defined as a deterministic function of the 30 selected predictors, using fixed cutpoints held constant simulation across conditions. The true tree structure is shown in Appendix B.

The total number of simulation conditions is 8, comprising four levels of variances of random effects and two levels of the number of clusters. For each simulation condition, 1,000 replications were conducted. The LLM with linear and main effects, XGBoost, and LLM–XGBoost models were fitted to the same generated data sets. That is, the total number of simulation runs is 24,000 (= 8 conditions × 1,000 replications × 3 models).

In generating outcomes for each simulation condition, the true tree and the generated predictors were set to the same across the replications, and the random effects were generated at each replication in the simulation condition. The same cluster identifiers for the two cross-classified clusters (e.g., schools and neighborhoods) were used across replications. In data generation, $\beta_0 = 0$ was specified, and standardized predicted values from the true tree structure were used to estimate β_0.

Evaluation measures were: (a) mean squared error (MSE) for relative model accuracy; (b) bias and root MSE (RMSE) for parameter recovery; and (c) Spearman's rank correlation between true and estimated permutation importance to evaluate the accuracy of group-aware permutation importance. For the MSE of LMM and LMM–XGBoost, four variants were considered for comparison: (a) full-sample conditional MSE (with random-effect predictions), (b) full-sample marginal MSE (without

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

random-effect predictions), (c) OOF conditional MSE, and (d) OOF marginal MSE. Conditional versus marginal MSE distinguishes model performance attributable to cluster-specific random effects from that explained by the fixed-effects component alone. Conditional predictions correspond to predicting new observations within clusters observed in the training data, whereas marginal predictions correspond to predicting outcomes for new clusters where cluster-specific random effects are unavailable. Full-sample MSE quantifies how well the fitted model matches the estimation data, whereas OOF MSE assesses generalization to held-out clusters or units, thereby reducing overfitting bias. The true permutation importance was computed from the true tree structure using the group-aware procedure described in Section 2.4. Group-aware permutation importance was estimated for LMM–XGBoost. Conventional (row-shuffling) permutation importance was computed for XGBoost and for LMM, reflecting common practice (e.g., Breiman, 2001; Fisher et al., 2019). For LMM, permutation scores were calculated using predictions from the fixed-effects component only. Because the true tree selects predictors from both Level-1 and cluster-level variables, the rank correlation between the true and estimated permutation importance implicitly evaluates the recovery of importance across predictors operating at different hierarchical levels. Permutation importance was evaluated both full-sample (to characterize the behavior of the fitted models with minimal resampling variance) and under CV (to assess generalizability).

For the selected simulation levels of the two varying conditions, the following patterns are expected. First, the LMM–XGBoost model yields the smallest conditional MSE when the random-effect variances are non-zero because the predictions condition on the estimated random effects, leaving primarily residual variation unexplained. The marginal MSE from LMM–XGBoost is larger because the marginal predictions average over the random effects and therefore do not include the realized cluster-specific deviations present in the observed outcomes. In contrast, the XGBoost model is trained directly on the observed responses and may partially capture between-cluster structure through predictors whose distributions vary across clusters. Consequently, its MSE is expected to lie between the conditional and marginal MSEs from LMM–XGBoost. Second, the marginal MSE from LMM and LMM–XGBoost is expected to increase as the ICC increases because the marginal predictions do not account for cluster-specific random effects. Third, for parameter recovery, the bias and RMSE of the fixed intercept are expected to increase as ICC increases and the number of clusters decreases, because the empirical mean of the random effects deviates more from zero when between-cluster variability is larger and fewer clusters are available to average out this variability. The random-effects SDs are expected to show decreasing bias as ICC and the number of clusters increase, whereas their RMSE increases with ICC (because they are estimated on a larger absolute scale) and decreases with the number of clusters. For the residual SD, both bias and RMSE are expected to show little dependence on ICC and to decrease as the number of clusters increases. Fourth, group-aware permutation importance from LMM–XGBoost is expected to outperform conventional (row-shuffling) permutation importance from XGBoost and LMM because it preserves the clustered (and cross-classified) dependence structure during permutation. At the no-variance level, the accuracy of the permutation importance measure is expected to be similar between LMM–XGBoost and XGBoost. In the presence of ICCs (i.e., non-zero variances), the accuracy of permutation importance under LMM–XGBoost is expected to exceed that of XGBoost, with the performance gap widening as ICC increases.

3.2. Analyses

In the residual-learning stage of XGBoost and the XGBoost component of the LMM–XGBoost, hyperparameters were selected using a grid search to balance capacity and stability under the simulation design. The learning rate (`eta`) was tuned as $\in 0.02, 0.03, 0.05$ for trees with depth $d \leq 5$ and restricted to $0.02, 0.03$ for deeper trees ($d = 6$). The maximum tree depth (`max depth`) was set to $d \in 3,4,5$ for $N \approx 1.25 \times 10^4$ and $d \in 3,4,5,6$ for $N \approx 5.0 \times 10^4$, constraining interaction order. The expected observations per terminal node were approximately $N/2^d$, yielding $\approx 1.6 \times 10^3$, 7.8×10^2, and 3.9×10^2 for $d = 3,4,5$ when $N = 1.25 \times 10^4$, and $\approx 6.3 \times 10^3$, 3.1×10^3, 1.6×10^3, and 7.8×10^2 for $d = 3,4,5,6$ when $N = 5.0 \times 10^4$. Because unit sample weights were used, the leaf curvature H equals the number of training

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

cases in each leaf under the squared-error objective (`reg:squarederror`); thus, the minimum child weight (`min child weight`) sets a lower bound on the minimum leaf size. The grid was adjusted by ICC level: `min child weight` $\in 25,50,75$ for none or small ICCs (0.00–0.05), $50,100$ for medium ICC (0.15), and $75,150$ for large ICC (0.25); for $d = 6$ (available only when $N = 5.0 \times 10^4$), an additional value 200 was included. ℓ_2 regularization (`lambda`) explored $\lambda \in 1,3,10,30$ to shrink leaf values and stabilize small partitions. For the ℓ_1 penalty, `alpha` was set to $0,0.5,1,3$. Under XGBoost's second-order objective, the optimal leaf value is defined as $\omega^* = -(H+\lambda)^{-1}\,\mathrm{sgn}(G)\,\max\big(|G|-\alpha,0\big)$ (Chen & Guestrin, 2016), where `alpha` = α applies soft-thresholding to the summed gradient G, and `lambda` = λ inflates curvature H (ridge shrinkage). On this scale, $\alpha \in [0.5,3]$ induces mild-to-moderate sparsity, zeroing leaves whose signal is near noise while preserving substantive effects; this complements the smoother ℓ_2 shrinkage from $\lambda \in 1,3,10,30$. Including $\alpha = 0$ ensures a baseline without ℓ_1 regularization, $\alpha = 0.5$–1 provides light sparsity to reduce variance when residual signal is diffuse across predictors, and $\alpha = 3$ offers stronger regularization for sensitivity analyses or deeper trees with many small leaves. Larger α values were avoided to prevent over-sparsification given moderate H. The search permitted up to `num boost round` = 1,200 with early stopping and used RMSE as the tuning criterion. Hyperparameter selection employed a combined five-fold CV, with subsampling ratios fixed at `colsample bytree` = 1.0 and `subsample` = 1.0 to prevent cross-cluster leakage.

In LMM–XGBoost, the tuning constants were set to $\varepsilon = 0.4$ and $\alpha_L = 0.5$. Empirically, with $\varepsilon = 0.4$, the estimates in LMM–XGBoost $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2_{u1}, \widehat{\sigma}^2_{u2}, \widehat{\sigma})$ differed only in the fourth or fifth decimal place between successive iterations. For permutation-based importance, 1,000 permutations were used for both the group-aware permutation importance from LMM–XGBoost and the conventional (global shuffle, unconditional) permutation importance from XGBoost. For comparison purposes, both the full-sample and the OOF importance estimates were obtained. LMM–XGBoost exhibited no convergence problems in any condition. The number of iterations required for convergence ranged from 10 to 45 across 8,000 runs (8 conditions × 1,000 replications).

3.3. Results

Table 2 reports the simulation results averaged over 1,000 replications for the full sample and the OOF values. Overall, the expected patterns were observed, with a few exceptions. Monte Carlo standard errors were computed to assess simulation uncertainty. With 1,000 replications, the resulting Monte Carlo standard errors were typically below 0.01 for the reported summaries and therefore negligible relative to the differences observed across methods.

Table 2, Panel A, presents the MSE under LMM, XGBoost, and LMM–XGBoost for the full sample and OOF. Across both the full sample and OOF, the MSE from LMM was the largest, indicating that ignoring interactions and nonlinearity in LMM led to poorer predictive performance. In addition, the marginal MSEs from LMM and LMM–XGBoost increased as the ICC increased in both the full sample and OOF. With OOF, the conditional MSE from LMM–XGBoost was the smallest, and, as expected, it fell between the conditional and marginal MSEs from LMM–XGBoost when the random-effect variances were non-zero. However, with the full sample, the MSE from XGBoost was smaller than or equal to the conditional MSE from LMM–XGBoost under non-zero random-effect variances. This may be due to the downward bias of full-sample MSE for XGBoost, which can partially absorb random-effect variation and overfit idiosyncratic patterns when predictions are evaluated on the same data used for training.

Table 2, Panel B, shows bias and RMSE of parameter estimates from LMM–XGBoost. Across all simulation conditions, the fixed intercept exhibited increasing bias in absolute value and larger RMSE as the variance of the random effects (or ICC) increased, with both bias and RMSE reduced when the number of clusters was larger. For the random-effects SDs (σ_{u1} and σ_{u2}), bias decreased as the random-effect variance (or ICC) increased and as the number of clusters increased, indicating improved centering of these estimators when between-cluster variability was stronger and more clusters were available. Their RMSE, however, increased with larger random-effect variance (reflecting greater

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Table 2. Simulation results: Panel A (prediction accuracy), Panel B (parameter recovery), and Panel C (accuracy of permutation importance)

Table 2a: Panel A: Simulation results for prediction accuracy (MSE)

	Conditions	Full Sample					OOF				
		LMM		XGBoost	LMM–XGBoost		LMM		XGBoost	LMM–XGBoost	
Var. of random effects	Sample size	Con.	Mar.		Con.	Mar.	Con.	Mar.		Con.	Mar.
No	Small	0.828	0.836	0.379	0.407	0.423	0.842	0.845	0.458	0.416	0.423
Small	Small	0.831	0.872	0.380	0.355	0.452	0.849	0.882	0.468	0.394	0.482
Medium	Small	0.831	0.954	0.365	0.383	0.574	0.849	0.965	0.474	0.395	0.575
Large	Small	0.831	1.097	0.384	0.384	0.783	0.850	1.108	0.481	0.395	0.758
No	Large	0.817	0.824	0.374	0.408	0.426	0.823	0.827	0.403	0.414	0.426
Small	Large	0.819	0.854	0.375	0.397	0.454	0.827	0.857	0.408	0.402	0.454
Medium	Large	0.818	0.933	0.384	0.396	0.555	0.827	0.936	0.415	0.401	0.555
Large	Large	0.818	1.073	0.389	0.396	0.734	0.827	1.075	0.420	0.402	0.734

Table 2b: Panel B: Simulation results for parameter recovery of LMM–XGBoost

	Conditions	$\hat{\beta}_0$		σ_{u1}		σ_{u2}		σ	
Var. of random effects	Sample size	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
No	Small	−0.007	0.019	0.082	0.083	0.086	0.086	0.089	0.094
Small	Small	−0.023	0.044	0.053	0.056	0.066	0.058	0.080	0.082
Medium	Small	−0.040	0.074	0.050	0.057	0.055	0.060	0.079	0.082
Large	Small	−0.058	0.107	0.049	0.070	0.054	0.081	0.080	0.082
No	Large	−0.004	0.014	0.069	0.070	0.107	0.107	0.089	0.091
Small	Large	0.011	0.026	0.032	0.035	0.043	0.033	0.070	0.077
Medium	Large	0.027	0.048	0.028	0.035	0.025	0.034	0.078	0.078
Large	Large	0.043	0.072	0.027	0.040	0.017	0.037	0.071	0.077

Table 2c: Panel C: Accuracy of permutation importance (rank correlations)

	Conditions	Fall Sample			OOF		
Var. of random effects	Sample size	LMM	XGBoost	LMM–XGBoost	LMM	XGBoost	LMM–XGBoost
No	Small	0.805	0.935	0.993	0.877	0.956	0.986
Small	Small	0.720	0.791	0.991	0.784	0.827	0.980
Medium	Small	0.679	0.764	0.991	0.676	0.750	0.980
Large	Small	0.686	0.695	0.991	0.656	0.663	0.976
No	Large	0.841	0.978	0.993	0.843	0.993	0.993
Small	Large	0.798	0.890	0.993	0.801	0.885	0.994
Medium	Large	0.753	0.794	0.993	0.732	0.794	0.994
Large	Large	0.743	0.710	0.995	0.750	0.704	0.994

absolute estimation error on a larger scale) and decreased with more clusters. In contrast, the residual SD (σ) showed relatively stable bias and RMSE across different levels of random-effect variance, with both measures modestly smaller when more clusters were available. Under the zero-variance conditions, LMM–XGBoost tended to overestimate the random-effect SD and exhibited larger sampling variability, rather than showing convergence issues due to boundary solutions.

Table 2, Panel C, reports the rank correlations between the true importance measures and the estimated permutation importance from LMM, XGBoost, and LMM–XGBoost for the full sample and OOF. Across all conditions and methods, the full-sample and OOF importance estimates were comparable. Across all conditions, LMM–XGBoost yielded the highest rank correlations, which were uniformly high, ranging from 0.976 to 0.995. When the random-effect variances were zero, the accuracy of LMM–XGBoost and XGBoost was similar, particularly in the large-sample OOF condition where their rank correlations were identical, although LMM–XGBoost still outperformed XGBoost in the small-sample conditions. In the presence of non-zero variances, the rank correlations for XGBoost and LMM deteriorated as ICC increased, whereas those for LMM–XGBoost remained high, so that the performance gap between LMM–XGBoost and XGBoost widened with increasing ICC.

4. Applications

In this section, the method described in Section 2 is applied to the Add Health data to take an advantage of rich data on depressive symptoms. The objective of the analysis is to comprehensively investigate the multi-domain factors of adolescent depressive symptoms, aiming to inform more effective and targeted mental health interventions.

4.1. Samples

Add Health is a nationally representative sample from the United States, comprising over 20,000 adolescents in grades 7–12 during the 1994–1995 school year (Harris et al., 2019). These participants have been tracked for 25 years, spanning into early midlife, through five waves of interviews. In this study, Wave 1 data (interviewed 1994–1995; $N = 20{,}745$) were selected. Add Health restricted-use data contract and Institutional Review Board approval were obtained.

Add Health data used 132-school representative of youth in the United States (52 middle schools and 80 high schools; `PSUSCID` variable) (Harris, 2013). Cases with missing sampling weights were deleted. After deletion, the total number of participants was 18,924. There were 2,344 neighborhood identifiers defined as the 1990 U.S. census tracts (`W1NHOOD1` variable), and 170 missing cases of neighborhood identifiers. Thus, 18,754 participants (18,924 − 170 missing cases of neighborhood identifiers) who were cross-classified by 132 schools and 2,344 neighborhoods were considered for analyses.

4.2. Outcomes and predictors

4.2.1. Outcomes

Depressive symptoms at Wave 1 were assessed using a modified version of the Center for Epidemiologic Studies Depression Scale (CES-D) (Perreira et al., 2005; Radloff, 1977). The modified 19-item version used in this study demonstrated good reliability in prior work (Perreira et al., 2005). Responses were recorded as 0 ("never or rarely"), 1 ("sometimes"), 2 ("a lot of the time"), and 3 ("most or all of the time or all of the time"). In this article, the sum scores of 19 items (H1FS1 through H1FS19 from the Wave I in-home interview) were used, with items H1FS4, H1FS8, H1FS11, and H1FS15 reverse-coded, and responses of "refuse" (response of 6), "don't know" (response of 8), and "not applicable" (response of 9) treated as missing data. Higher sum scores represent greater depressive symptoms. The distribution of the sum scores exhibited moderate right skewness and lighter tails (skewness = 1.08, kurtosis = 1.57). Therefore, the sum scores were square-root transformed to approximate normality, as

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

was done with CES-D data in previous work with the Add Health dataset (e.g., Dawson et al., 2019). After the square-root transformation, only two observations (participants) were identified as extreme points that deviated significantly from the normal distribution. However, because the deviation was small, the two observations were retained for analyses.

For 18,754 participants cross-classified by 132 schools and 2,344 neighborhoods, the median number of students per school is 123 (range = [1,1703]; SIQR = 43), and the median number of students per neighborhood is 2 (range = [1,275]; SIQR = 2). Each neighborhood includes one or multiple schools, ranging from 1 to 3. In addition, each school draws students from one or multiple neighborhoods, ranging from 1 to 224. Based on the unconditional LMM, the intraclass correlation (ICC) for schools and neighborhoods was calculated as 0.043, indicating that 4.3% of the total variance in students' depressive symptoms is attributable to differences between the schools they attended and neighborhoods they lived in.

Below, predictors are described. An extensive literature review was conducted to select these predictors from the Add Health data (see Appendix C for paper selection and the OSF, https://osf.io/jkctg/ for review results). In this study, all available and identifiable predictors from a prior study on depressive symptoms using Add Health, except for predictors related to romantic relationships (which were not part of the contract with Add Health), were selected.

4.2.2. Predictors

Based on the literature review, 36 individual-level predictors were extracted from the Wave I in-home interview and Wave I in-school questionnaire data. The domains[4] for these predictors (listed in parentheses) include: (a) demographic characteristics (age [`age`], sex/gender [`gender`], race [`race`], ethnicity [`ethnicity`], attending religious activities [`relactivities`], religiosity [`religiosity`]), (b) psychological well-being & cognition (self-esteem scores [`selfesteem`]), (c) education (grade in school [`grade`], school connectedness [`connectedness`], trouble getting along with teachers [`alongwithtch`], GPA [`overall GPA`]), (d) family (parent–adolescent relationship with mom and dad, respectively [`PArelation.mom`, `PArelation.dad`], parental education [`Peducation`], family/household structure [`structure`], family income [`familyincome`], parent occupation [`Poccupation`], language spoken at home [`homelang`], parent age [`parentage`], marital status of parents [`Pmarital`], parental attachment [`attachment`], the number of household members [`numhousemem`]), (e) friends and social network (social support [`socialsupport`], feeling cared about by teachers [`feelingT`], participation in sports [`partinsports`], participation in non-sport extracurriculars [`nonsportec`]), (f) individual-level contextual factors (whether the schools were a reason for living in the neighborhood [`neighsch`], neighborhood interaction [`neighinteraction`], parental-perceived neighborhood collective efficacy [`neighborefficacy`], parental-perceived neighborhood disorder [`neighbordisorder`], perceived neighborhood social cohesion [`neighborcohesion`], whether the reason for living in the neighborhood is safety [`neighsafety`]), (g) medication & substance use and abuse (alcohol problems [`alcohol`]), (h) risk behavior (delinquent behaviors [`delinquent`]), (i) physical health (migraine status of adolescent [`migraine`]), and (j) socio-economic status, labor market, & occupation (employment for adolescent [`Aemployment`]). Descriptive statistics of these predictors from multiple domains is provided in the table of Appendix D.

When there are 36 predictors, there are many possible linear- and nonlinear- (e.g., quadratic and cubic) main and interaction effects, leading to high dimensionality for outcomes from 18,754 participants: 36 linear main effects, 630 linear two-way interactions, 7,140 linear three-way interactions,

[4]The domains were defined using the list of topics in the Codebook Explorer on the Add Health website: https://addhealth.cpc.unc.edu/documentation/codebook-explorer/#/. If predictors did not clearly fall into a specific domain (for instance, family income could be classified under either the Family domain or the SES domain) or were not previously assigned to a domain, the authors in the current study classified the predictor accordingly.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

36 quadratic main effects, 630 quadratic two-way interactions, 7,140 quadratic three-way interactions, 36 cubic main effects, 630 cubic two-way interactions, and 7,140 cubic three-way interactions, resulting in a total of 22,418 terms.

4.3. Pre-processing and analyses

There were two near-zero-variance predictors (`race`, `structure`) and two highly correlated predictors (absolute correlation > |0.8|; `relactivities`, `connectedness`). These four variables may cause problems in fitting LMM or XGBoost. Thus, they were excluded from the analyses. Categorical predictors were coded as dummy variables with the first category as the reference, and all dummy variables except the reference were included in the analyses. Continuous predictors were standardized when fitting LMM, XGBoost, and LMM–XGBoost.

For comparison purposes, LMM, XGBoost, and LMM–XGBoost were fit to the same empirical dataset. In fitting the LMM, linear and no-interaction effects of 32 predictors (36 predictors minus 4 predictors with near-zero variance or high correlations) were considered, which is a common practice in applications. The comparison of LMM with LMM–XGBoost underscores the inclusion of potential nonlinearity and interaction effects in LMM–XGBoost. In addition, the comparison between XGBoost and LMM–XGBoost highlights the modeling of crossed random effects. The lmer function from the lme4 package and the xgboost function were used to fit LMM and XGBoost, respectively. The estimation method, as outlined in Section 2, was employed to fit LMM–XGBoost.

For the Add Health analysis, hyperparameters tuning for XGBoost and the XGBoost component of the LMM–XGBoost were designed to balance flexibility and stability under a large cross-classified structure. The analytic sample comprised $N = 18{,}754$ respondents nested within 132 schools and 2,344 neighborhoods. The predictor set included 20 continuous variables and 12 categorical variables, which expanded to 36 dummy indicators after treatment coding, yielding a marginal design matrix with 56 predictors. To control model capacity while allowing nonlinearities and interaction effects, the learning rate (`eta`) was tuned over $\{0.03, 0.05, 0.08\}$. Tree depth was restricted to `max_depth` $\in \{3,4,5\}$, which limits interactions to modest order and ensures adequate effective sample size per terminal node. With $N = 18{,}754$, the expected number of observations per leaf is approximately $N/2^d$, yielding roughly 2.34×10^3, 1.17×10^3, and 5.86×10^2 cases for depths $d = 3,4,5$, respectively—sufficient for stable leaf estimates under the squared-error objective. Because all observations carry unit weights, the leaf curvature H equals the number of cases in each leaf, so the minimum child weight (`min_child_weight` $\in \{10,20,40\}$) imposes a lower bound on leaf size to reduce variance in small partitions. Regularization hyperparameters were tuned to stabilize boosting in the presence of moderate collinearity across demographic and psychosocial predictors. Ridge shrinkage (ℓ_2, `lambda` $\in \{1,3,10\}$) controlled the magnitude of leaf values, whereas the ℓ_1 penalty (`alpha` $\in \{0,0.1,0.5\}$) induced mild sparsity via soft-thresholding. Under the second-order approximation, XGBoost's optimal leaf value is $\omega^* = -(H+\lambda)^{-1}\,\mathrm{sgn}(G)\max\left(|G| - \alpha, 0\right)$ so $\alpha \in [0.1, 0.5]$ provides light regularization appropriate for dense data without overshrinking meaningful signals. Subsampling parameters were fixed at `subsample` = 1.0 and `colsample_bytree` = 1.0 to prevent cross-cluster leakage during combined five-fold CV, where folds were defined on whole (school × neighborhood) cells. Up to 2,000 boosting iterations with early stopping were allowed, and RMSE served as the tuning criterion.

In LMM–XGBoost, the tuning constants were fixed at $\varepsilon = 0.4$ and $\alpha_L = 0.5$ as in the simulation study. With $\varepsilon = 0.4$, the LMM–XGBoost estimates $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2_{u1}, \widehat{\sigma}^2_{u2}, \widehat{\sigma})$ changed only in the fourth or fifth decimal place between successive iterations. For the permutation-based importance measures, 1,000 permutations were used for both the group-aware permutation importance from LMM–XGBoost and the conventional (global shuffle, unconditional) permutation importance from XGBoost, and for comparison purposes, importance was computed using both full-sample and OOF estimates. The number of iterations to convergence was 25.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Table 3. Empirical study: Prediction accuracy of LMM, XGBoost, and LMM–XGBoost (MSE)

Models	Modeling components		MSE	
	Nonlinearity and interactions	Random effects	Full sample	OOF
LMM conditional		√	0.899	0.920
LMM marginal		√	0.922	0.930
XGBoost	√		0.826	0.842
LMM–XGBoost conditional	√	√	0.646	0.721
LMM–XGBoost marginal	√	√	0.870	0.890

Computation time was recorded using the elapsed time from `system.time()` in `R` and, for GLMM–XGBoost, the stage-wise timing log produced by the implementation. The baseline LMM required 770.78 seconds (12.85 minutes). For standalone XGBoost, hyperparameter tuning (grid search with CV) plus the final refit required 9,290.28 seconds (154.84 minutes), and the global permutation importance analysis with 1,000 replications required 1,307.72 seconds (21.80 minutes), yielding a combined total of 176.64 minutes (2.94 hours) for these two components. For LMM–XGBoost, the XGBoost tuning step required 9,101.22 seconds (151.69 minutes), the EM-style iterative estimation required 19,369.5 seconds (322.83 minutes), and the full-sample permutation importance procedure with 1,000 replications required an additional 1,266.66 seconds (21.11 minutes), yielding a combined total of 495.63 minutes (8.26 hours) for these components. All computations were conducted on the Add Health Secure Research Workspace (SRW).

4.4. Results

Table 3 reports four variants of the MSE to differentiate model fit, generalization performance, and the contribution of random effects. As expected, conditional MSE was lower than marginal MSE, and full-sample MSE was lower than OOF MSE. In addition, the MSE for XGBoost lay between the conditional and marginal MSEs from LMM–XGBoost. This pattern was expected because XGBoost captured nonlinear and interaction effects among predictors and may have absorbed some variability arising from unmodeled group-level structure, but it did not explicitly model crossed random effects; in contrast, the conditional LMM–XGBoost predictions exploited both random effects and complex predictor effects (yielding the lowest MSE), whereas the marginal LMM–XGBoost predictions integrated over random effects (yielding higher MSE). The LMM with linear, no-interaction effects and a conditional MSE provided a baseline that accounted for crossed random effects but did not capture nonlinear or interaction effects among the predictors. When both nonlinearity and interactions were modeled in LMM–XGBoost, the conditional OOF MSE decreased from 0.920 to 0.721, indicating improved accuracy and suggesting that LMM–XGBoost was better able to capture the complexity of the relationship between the predictors and depression than the linear, no-interaction LMM. In addition, XGBoost attained an OOF MSE of 0.842, which was higher than the OOF conditional MSE of 0.721 for LMM–XGBoost, suggesting that the inclusion of crossed random effects further enhanced predictive accuracy beyond what was achieved by XGBoost alone.

For comparison with LMM–XGBoost, the table of Appendix E[5] shows the fixed-effect estimates of LMMs. As shown for significance of fixed effects in the table of Appendix E, key predictors (factors) associated with increased depression scores include older age, alcohol use, delinquency, poorer social support, and individuals' perception of their neighborhood factors like disorder and lower neighborhood efficacy. Conversely, higher GPA, parental education, and social support are protective factors,

[5]In the table of Appendix E, estimates (EST) in bold indicate significance at alpha=0.05; values in brackets indicate standardized values, and a dash indicates that an effect is not modeled.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Table 4. Empirical study: Random-effect estimates of the LMMs and LMM-RF

	Unconditional LMM	Conditional LMM	LMM–XGBoost
	EST	EST	EST
SD of $u_{1j}[\sigma_{u1}$;PSUSCID]	0.212	0.118	0.150
SD of $u_{2k}[\sigma_{u2}$;W1NHOOD1]	0.125	0.045	0.059
SD of $\varepsilon_{ijk}[\sigma]$	1.155	0.953	0.765

leading to lower depression scores. These results highlight the complex interplay between individual, family, and individuals' perception of their neighborhood factors in understanding depression.

The intercept was estimated as 3.105 ($\widehat{\beta}_0$ = 3.105, SE = 0.025). In addition, Table 4 presents the SDs of the random effects and random residuals. Although the SDs of the crossed random effects were comparable between the conditional LMM (with predictors) and LMM–XGBoost, the SD of the residual variance was reduced from 0.953 to 0.765 in LMM–XGBoost compared to LMM. To investigate potential unmodeled nonlinear interactions in LLM, specifically two-way interactions, we examined whether the fitted values of the approximation, as a function of a pair of predictors, deviate from the linear combination of the two predictors (Elith et al., 2008). Departures from additivity can be assessed by computing the fitted values for any pair of predictors over a grid of all possible levels for the two predictors. For continuous predictors, 100 sample values were taken. The fitted values were then regressed onto the grid. Large residuals from this model indicate that the fitted values are not a linear combination of the predictors, suggesting nonlinearity or a possible interaction. As shown in Appendix F for selected predictors, there are unmodeled nonlinear two-way interactions in residuals of LLM with linear, no-interaction effects, indicated by the non-ignorable residuals.

Given that LMM–XGBoost exhibited the highest predictive accuracy among LMM, XGBoost, and LMM–XGBoost, its results are further interpreted below using the group-aware permutation importance. Because permutation-based importance is widely used in applications of XGBoost, the conventional importance measure is also reported in Table 5 for comparison. As the full-sample and OOF importance estimates were comparable, Table 5 presents the OOF-based values. Although dummy-coded indicators for categorical predictors were used in LMM–XGBoost, the variable-level OOF importance values were averaged across folds, aggregated to the level of the original predictors by summing over all dummy features derived from the same predictor, and finally rescaled to sum to 1 to obtain interpretable, predictor-level relative importance (RelImp) reported in Table 5.

As shown in Table 5, the RelImp ranks for the LMM–XGBoost model highlight a strong reliance on individual and contextual factors, with a particularly strong emphasis on indicators of social and relational well-being. The most influential predictors are `selfesteem` (Rank 1, RelImp = 0.270) and `socialsupport` (Rank 2, RelImp = 0.159). Both LMM–XGBoost and XGBoost identify these two variables as the top predictors, indicating highly stable importance across modeling frameworks. Beyond these leading predictors, a comparison of the full ranking reveals systematic differences between the two approaches. LMM–XGBoost assigns substantially greater importance to contextual variables reflecting home and school environments. For example, `numhousemem` (number of household members) receives Rank 3 under LMM–XGBoost but Rank 12.5 under XGBoost, while `alongwithtch` (trouble getting along with teachers) receives Rank 4 versus Rank 7 under XGBoost. These discrepancies reflect the influence of the mixed-effects structure, which captures clustering and adjusts for between-cluster variability. Conversely, LMM–XGBoost sharply de-prioritizes static demographic features. For example, `homelang` and `ethnicity` are ranked 29.5 and 31, respectively, whereas XGBoost ranks them notably higher (Ranks 12.5 and 22.5). These differences suggest that once school and neighborhood clustering and random effects are properly accounted for, demographic indicators contribute little incremental predictive value beyond more proximal psychological and contextual characteristics.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Table 5. Empirical study: Permutation importance comparisons between LMM–XGBoost and XGBoost (averaged ranks for ties)

Feature	LMM–XGBoost				XGBoost				Rank Diff
	Δ	Δ_{sd}	RelImp	Rank	Δ	Δ_{sd}	RelImp	Rank	(XGB – LMM-XGB)
selfesteem	1.061505e-3	9.562157e-6	0.270	1.0	3.872163e-1	1.947406e-2	0.585	1.0	0.0
socialsupport	6.264107e-4	8.914141e-6	0.159	2.0	1.454871e-1	5.596827e-3	0.220	2.0	0.0
delinquent	1.800853e-4	2.731383e-6	0.046	5.0	2.324967e-2	4.747986e-3	0.035	3.0	−2.0
gender	1.474684e-4	2.056467e-6	0.037	8.5	2.051238e-2	4.762596e-3	0.031	4.0	−4.5
overall_GPA	1.753712e-4	2.638773e-6	0.045	6.0	1.647767e-2	1.513699e-3	0.025	5.0	−1.0
feelingT	1.438357e-4	2.034181e-6	0.037	8.5	1.261554e-2	1.652907e-3	0.019	6.0	−2.5
alongwithtch	1.838346e-4	2.979503e-6	0.047	4.0	1.067708e-2	4.490256e-3	0.016	7.0	3.0
alcohol	8.445119e-5	1.625680e-6	0.021	13.0	8.168393e-3	3.028416e-3	0.012	8.0	−5.0
familyincome	1.586544e-4	2.333816e-6	0.040	7.0	7.536789e-3	4.886450e-3	0.011	9.0	2.0
age	6.313651e-5	9.653786e-7	0.016	16.0	6.059558e-3	1.663870e-3	0.009	10.0	−6.0
PArelation.mom	1.239922e-4	1.841065e-6	0.031	10.0	4.371630e-3	1.506191e-3	0.007	11.0	1.0
numhousemem	2.055719e-4	2.628732e-6	0.052	3.0	2.505770e-3	6.938410e-4	0.004	12.5	9.5
homelang	6.548942e-6	1.565954e-7	0.002	29.5	2.747692e-3	6.510980e-4	0.004	12.5	−17.0
neighborefficacy	3.975466e-5	6.318288e-7	0.010	20.0	1.706087e-3	1.249336e-3	0.003	15.0	−5.0
Aemployment	3.236073e-5	6.506731e-7	0.008	22.0	1.594188e-3	4.848590e-4	0.002	18.0	−4.0
Poccupation	9.133491e-5	1.145070e-6	0.023	12.0	2.208218e-3	9.924560e-4	0.003	15.0	3.0
religiosity	5.082890e-5	8.204799e-7	0.013	17.0	1.011321e-3	6.561730e-4	0.002	18.0	1.0
attachment	4.152325e-5	8.594322e-7	0.011	19.0	1.001145e-3	1.226145e-3	0.002	18.0	−1.0
ethnicity	5.619312e-6	2.003656e-7	0.001	31.0	9.699120e-4	6.391220e-4	0.001	22.5	−8.5
Peducation	7.239364e-5	8.427196e-7	0.018	15.0	1.708095e-3	7.211870e-4	0.003	15.0	0.0
nonsportec	7.915246e-5	1.537733e-6	0.020	14.0	8.635560e-4	9.916990e-4	0.001	22.5	8.5

(Continued)

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Table 5. Continued

Feature	LMM–XGBoost				XGBoost				Rank Diff
	Δ	Δ_{sd}	RelImp	Rank	Δ	Δ_{sd}	RelImp	Rank	(XGB – LMM-XGB)
migraine	1.380225e-5	4.160324e-7	0.004	25.5	6.466520e-4	6.421240e-4	0.001	22.5	−3.0
parentage	1.163021e-4	1.637514e-6	0.030	11.0	6.265370e-4	1.107772e-3	0.001	22.5	11.5
grade	7.293576e-6	1.992323e-7	0.002	29.5	3.938010e-4	6.102560e-4	0.001	22.5	−7.0
neighborcohesion	4.550288e-5	7.170917e-7	0.012	18.0	2.861630e-4	4.209550e-4	0.000	28.5	10.5
Pmarital	1.649119e-5	2.661199e-7	0.004	25.5	3.653590e-4	2.720530e-4	0.001	22.5	−3.0
neighbordisorder	2.746794e-5	6.286515e-7	0.007	23.0	2.325510e-4	3.959630e-4	0.000	28.5	5.5
neighinteraction	1.155581e-5	2.710271e-7	0.003	27.5	1.484090e-4	2.748010e-4	0.000	28.5	1.0
neighsch	1.127299e-5	3.549574e-7	0.003	27.5	−2.612390e-4	3.163560e-4	0.000	28.5	1.0
neighsafety	2.530546e-5	9.332399e-7	0.006	24.0	−8.890000e-5	3.243740e-4	0.000	28.5	4.5
partinsports	3.606456e-5	5.950321e-7	0.009	21.0	−2.350760e-4	4.108840e-4	0.000	28.5	7.5

Note: Rank Diff = $\text{Rank}_{\text{XGB}} - \text{Rank}_{\text{LMM-XGB}}$. Ranks use averaged ranks for ties based on full-precision relative importance.

This divergence illustrates how the multilevel structure inherent in LMM–XGBoost reshapes variable importance by adjusting for nested and cross-classified dependencies in the data.

Figure 1 presents an accumulated local effect (ALE) plot (Apley & Zhu, 2020) to visualize the functional form of the effects of the two most important predictors, self-esteem and social support, for illustrative purposes. A scatter plot of depression versus a predictor is also presented to show how the relationship in the data is adequately modeled in the LLM–XGBoost.

Figure 1 (top left) displays the ALE plot of standardized self-esteem scores, where lower scores indicate higher self-esteem.[6] As shown in the scatter plot in Figure 1 (top right, with a fitted smooth function in red and a fitted linear function as a blue dotted line), the ALE plot reveals a nonlinear relationship between self-esteem and depression scores. Specifically, there is a sharp increase in depression scores at moderate-low range self-esteem values (between −1 and 2), followed by a more gradual increase at lower self-esteem levels. This suggests that individuals with moderate self-esteem experience a stronger increase in predicted depression scores, while the effect continues but at a slower rate for individuals with lower self-esteem.

Figure 1 (middle left) provides the ALE plot of standardized social support scores, illustrating the nonlinear relationship between social support and depression scores. The ALE plot reveals a flat or slightly increasing trend at very low levels of social support (around −6 to −2). However, as social support increases beyond this point (around −2 and higher), there is a sharp decrease in depression scores, reflecting the strong protective effect of higher social support in reducing depression. As seen in the scatter plot in Figure 1 (middle right; with a fitted smooth function in red and a fitted linear function in a blue dotted line), the ALE plot captures the same nonlinear pattern, where depression scores rise slightly at extremely low levels of social support and then significantly decline as social support increases.

Figure 1 (bottom left), the heatmap in the provided ALE plot, represents the nonlinear interaction effects between standardized self-esteem scores (x-axis) and standardized social support scores (y-axis) on depression scores. The color gradient indicates different levels of depression, with darker shades representing lower depression scores and lighter shades indicating higher depression scores. The interaction between self-esteem and social support is evident in the varying color intensity across the grid. The combined effect of both variables appears nonlinear. At lower levels of social support, the effect of increasing self-esteem scores is more gradual. However, as social support increases, the effect of self-esteem on reducing depression becomes more pronounced, reflected in the steeper color gradient at moderate self-esteem levels. These patterns in the ALE plot for the nonlinear interactions were generally aligned with the interaction plot of the standardized self-esteem scores (x-axis) and standardized social support scores (binned based on centiles) on depression scores, as shown in Figure 1 (bottom right).[7] In Figure 1 (bottom right), a more substantial impact was observed at moderate to low levels of standardized self-esteem scores, depending on varying levels of social support.

5. Summary and discussion

This study was motivated by the problem of selecting important predictors for multilevel cross-classified continuous outcomes commonly found in the social and behavioral sciences. The LMM–XGBoost specification, an iterative estimation method for the model, a group-aware permutation importance

[6] In the Add Health study, the self-esteem items ask respondents to indicate their agreement or disagreement with statements such as, "You have a lot to be proud of." The response options for each item range from 1 (strongly agree) to 5 (strongly disagree). Lower scores indicate higher self-esteem because agreeing or strongly agreeing with positive self-assessment statements, such as having "a lot to be proud of," suggests a more favorable self-view, which aligns with higher self-esteem. In contrast, higher scores (e.g., disagreeing) reflect lower self-esteem, as they indicate a less positive self-perception.

[7] In Figure 1 (bottom right), note that the steep decreasing pattern in the smooth line for the 0.25–0.5 centile group of standardized social support scores is due to one extreme value.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

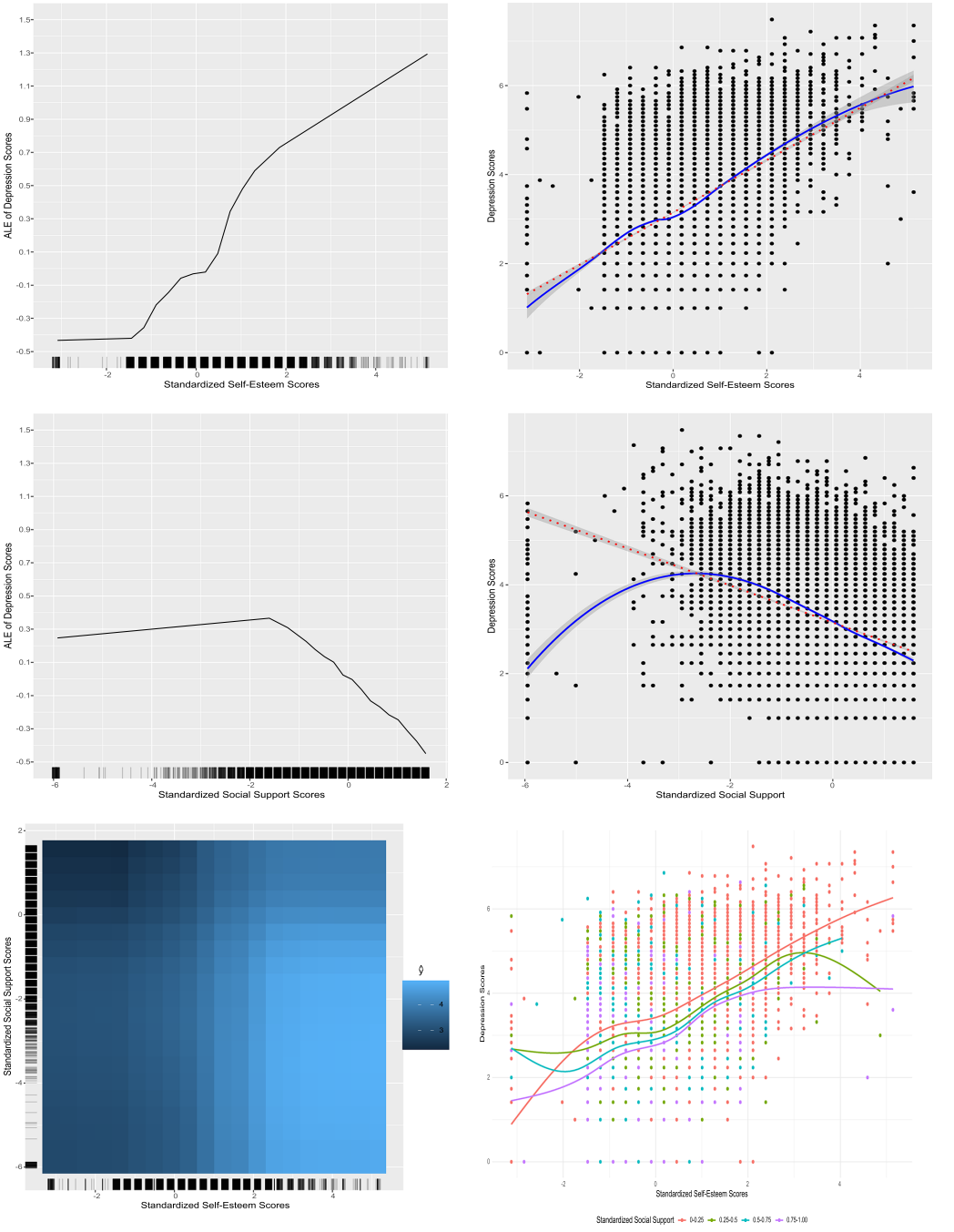


Figure 1. Empirical study: ALE plots and scatter plots of selected predictors—Standardized self-esteem scores (top), standardized social support scores (middle), and the interaction between the two (bottom).
Note: Each tick mark in the *x*-axis or the *y*-axis represents values of a continuous predictor; The smooth function is plotted in blue, and the fitted linear function is shown as a red dotted line.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

measure, and a combined-group CV method were developed, all tailored to multilevel cross-classified continuous outcomes.

The simulation study shows that LMM–XGBoost achieved the lowest conditional MSE with OOF and the highest rank correlations with the true variable importance across all conditions, with its performance advantage over XGBoost widening as ICC increased. Parameter estimates were generally well recovered, although random-effect SDs were mildly overestimated under zero-variance conditions. These findings indicate that LMM–XGBoost effectively captures nonlinear and cross-classified structure while maintaining stable estimation properties. Across all conditions, OOF prediction using combined-group CV produced more trustworthy evaluations of model performance than full-sample estimates, which tended to underestimate error and obscure differences between methods. The simulations further demonstrate that permutation importance must account for the cross-classified grouping to avoid distortions in variable rankings, with the proposed group-aware measure showing uniformly superior alignment with the true importance. These results affirm that proper handling of cross-classified structure is essential for both accurate prediction and valid variable-importance inference.

The proposed methods were applied to the Add Health data. The practical implications of the findings in the empirical study are significant for interventions targeting adolescent well-being and mental health. The results highlight that Psychological Well-Being & Cognition and Friends & Social Network are critical domains in shaping adolescent outcomes, with self-esteem and social support being the most influential predictors. This underscores the importance of developing programs aimed at improving adolescents' self-esteem and strengthening their social relationships to promote positive mental health outcomes. In addition, the nonlinear relationship, such as the sharp changes at specific ranges of these predictors, can provide valuable insights into where interventions might be most effective. For instance, enhancing self-esteem interventions could lead to significant mental health improvements for individuals with moderate-low levels of self-esteem. Similarly, increasing social support could lead to significant mental health benefits, particularly when an individual's level of social support is high.

5.1. Limitations and future research directions

Several limitations of the proposed LMM–XGBoost framework should be acknowledged, and they suggest directions for further methodological development. These limitations relate to model specification, estimation procedures, and potential extensions of the framework to other data structures.

The current specification of LMM–XGBoost models cluster-level dependence through random intercepts while allowing the ML component to capture complex predictor–outcome relationships. In conventional LMMs, heterogeneity in predictor effects across clusters is often modeled through parametric random slopes, which represent unexplained cluster-specific deviations in the linear effects of predictors. In contrast, the proposed framework assumes that part of the apparent slope heterogeneity arises from systematic nonlinearities, interactions, or higher-order relationships among predictors, which can be flexibly captured by the XGBoost component. Nevertheless, further investigation is needed to assess the performance of the proposed method when the data-generating process includes parametric random slopes. In such cases, random slopes could in principle be incorporated into the hybrid LMM–XGBoost framework. However, doing so would require substantial extensions to the estimation algorithm, particularly modifications to the projection step used to separate fixed-effect and random-effect components in order to disentangle nonlinear fixed effects learned by XGBoost from parametric random slopes. Developing an extended C-projection method that accommodates random slopes within the iterative LMM–XGBoost algorithm is an important direction for future research.

A limitation of the proposed framework is that cluster-level heterogeneity can be represented only to the extent that it is explained by observed predictors. The ML component captures nonlinear relationships among the observed predictors, while the mixed-effects component accounts for remaining cluster-level dependence through random intercepts. As in many observational modeling settings, this approach implicitly assumes that the relevant cluster-level predictors explaining systematic heterogeneity are observed. When important cluster-level predictors are omitted, unexplained heterogeneity

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

may remain in the random effects, and the separation between the nonlinear fixed component and the random effects may be imperfect. In such cases, the *C*-projection used to orthogonalize the ML predictions with respect to the cluster design may not fully isolate the nonlinear fixed component from the random-effect structure. Future research is needed to develop formal diagnostic procedures for detecting omitted cluster-level predictors in this framework.

The proposed algorithm alternates between updating the nonlinear component using XGBoost and estimating random-effect variance components using the LMM. Because these components are estimated sequentially rather than jointly within a single optimization step, the separation between the nonlinear fixed component and the random effects may be imperfect in finite samples. This may lead to some inflation of variance-component estimates, particularly under conditions with small ICC values and a limited number of clusters, where variance components are more weakly identifiable. The simulation results support this interpretation, as the bias in the estimated random-effects variances decreases with increasing ICC and with a larger number of clusters. These findings suggest that the observed variance inflation reflects a finite-sample estimation issue associated with the alternating estimation framework rather than a fundamental limitation of the proposed method.

In the current implementation, the XGBoost hyperparameters were tuned once prior to the iterative procedure and then held fixed across iterations to reduce computational burden. Although this approach performed well in the simulation study, it is possible that the optimal hyperparameters may change as the projected outcome is updated during the iterative algorithm. Developing adaptive re-tuning strategies within the iterative framework represents a potential direction for future research.

For tuning XGBoost in LMM–XGBoost, as well as for OOF prediction and group-aware permutation importance, CV was blocked on Cluster1 × Cluster2 intersections to prevent leakage across the cross-classified structure and to target generalization to unseen cell combinations. This strategy can be inappropriate when the cross-classification is sparse or highly imbalanced (e.g., few non-empty cells relative to *K*, very small per-cell), which yields unstable or degenerate folds; in such cases, blocking on a single dominant cluster is preferable. The `R` function developed in the present study implements both options via an argument that toggles between combined (Cluster1 × Cluster2) blocking and Cluster-only blocking, allowing the analyst to align the validation scheme with the data structure and the intended generalization target. In sensitivity analyses, simulation results were substantively unchanged across these CV choices, indicating that our conclusions were robust to the CV scheme. However, these findings are constrained by the simulation conditions examined in the current study, and additional simulation studies under alternative data-generating mechanisms and imbalance patterns are warranted.

Another direction for future research concerns the development of variable-importance measures that more explicitly reflect heterogeneity across clusters. In the present study, permutation-based importance measures summarize the overall contribution of predictors across the full sample as well as under grouped CV, while accounting for clustered dependence through the group-aware permutation procedure. However, in multilevel settings, the relevance of predictors may vary across clusters due to differences in contextual conditions or cluster-level characteristics. Extending the current framework to evaluate predictor importance within subsets of clusters defined by specific values of cluster-level predictors could provide additional insight into such heterogeneity. Developing cluster-conditional or subgroup-specific importance measures within the LMM–XGBoost framework may therefore provide a useful way to characterize how predictor relevance varies across clusters.

The present study focuses on cross-sectional multilevel data with cross-classified clustering structures, and the proposed estimation framework is developed for this setting. Consequently, the current formulation is not directly applicable to other types of data structures. For example, intensive longitudinal or repeated-measures settings typically involve additional sources of dependence, such as temporal correlation and subject-specific trajectories, which are commonly modeled using different random-effect structures. Adapting the proposed LMM–XGBoost framework to such settings would therefore require extensions to accommodate longitudinal dependence and simulation designs tailored to features, such as the number of time points and within-subject correlation patterns. Extending

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

the proposed approach to intensive longitudinal or repeated-measures data therefore represents an important direction for future methodological research.

When contextual or level-specific effects are of interest, Level-1 predictors can be cluster-mean centered, and the corresponding cluster means can be included as additional predictors when training the XGBoost component in LMM–XGBoost. This specification separates within-cluster and between-cluster variation in the predictors, allowing XGBoost to learn potentially nonlinear relationships for each component. In this way, the framework can distinguish observation-level effects from contextual effects associated with clusters. Approaches for disaggregating level-specific effects in cross-classified multilevel models are discussed by (Guo et al. (2024).

Despite these limitations, the present study demonstrates the applicability of a hybrid approach that combines LMM with an ML method (XGBoost) and variable-importance measures for cross-sectional continuous outcomes in cross-classified multilevel structures. This hybrid approach can be extended to other types of data and can incorporate different ML methods. For example, it can be adapted to cross-classified multilevel longitudinal data structures by modeling dependence arising from both between-person clusters and cross-classified clusters (e.g., schools and neighborhoods) using random effects, while accounting for within-person correlations through random residuals. In addition, within this hybrid framework, XGBoost can be replaced with alternative GB algorithms (e.g., LightGBM; Ke et al., 2017) or deep learning methods (e.g., Goodfellow et al., 2016). Although these extensions are not explored in this article, the present work may serve as an initial step toward applying the hybrid approach in these broader contexts.

Data availability statement. The example data and R code are available on the Open Science Framework at https://osf.io/jkctg/.

Acknowledgments. This research uses data from Add Health, funded by grant P01 HD31921 (Harris) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), with cooperative funding from 23 other federal agencies and foundations. Add Health is currently directed by Robert A. Hummer and funded by the National Institute on Aging cooperative agreements U01 AG071448 (Hummer) and U01AG071450 (Aiello and Hummer) at the University of North Carolina at Chapel Hill. Add Health was designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill.

Funding statement.. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

Disclosure of artificial intelligence (AI) use. The first author used ChatGPT-5 (OpenAI, 2025) to assist in checking grammar and creating tables in LaTeX format. All text and tables were subsequently reviewed, edited, and verified by the author to ensure accuracy, consistency, and appropriateness for scientific publication. The author takes full responsibility for the content of this manuscript.

References

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*, 1059–1086. https://doi.org/10.1111/rssb.12377

Barker, K. M., Dunn, E. C., Richmond, T. K., Ahmed, S., Hawrilenko, M., & Evans, C. R. (2020). Cross-classified multilevel models (CCMM) in health research: A systematic review of published empirical studies and recommendations for best practices. *SSM-Population Health*, *12*, 100661. https://doi.org/10.1016/j.ssmph.2020.100661

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

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. https://doi.org/10.1201/9781315139470.

Capitaine, L., Genuer, R., & Thiébaut, R. (2021). Random forests for high-dimensional longitudinal data. *Statistical Methods in Medical Research*, *30*, 166–184. https://doi.org/10.1177/0962280220946080

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., & Yuan, J. (2025). XGBoost: Extreme gradient boosting. (Version 1.7.11.1) [R package]. CRAN. https://doi.org/10.32614/CRAN.package.xgboost

Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research: Using cluster analysis to create synthetic neighborhoods. *Sociological Methods & Research*, *35*, 311–351. https://doi.org/10.1177/0049124106292362

Dawson, C. T., Wu, W., Fennie, K. P., Ibañez, G., Cano, M. Á., Pettit, J. W., & Trepka, M. J. (2019). Perceived neighborhood social cohesion moderates the relationship between neighborhood structural disadvantage and adolescent depressive symptoms. *Health & Place*, *56*, 88–98. https://doi.org/10.1016/j.healthplace.2019.01.001

De'Ath, G. (2002). Multivariate regression trees: A new technique for modeling species–environment relationships. *Ecology*, *83*, 1105–1117. https://doi.org/10.1890/0012-9658

Duong, J., & Bradshaw, C. (2014). Associations between bullying and engaging in aggressive and suicidal behaviors among sexual minority youth: The moderating role of connectedness. *Journal of School Health*, *84*, 636–645. https://doi.org/10.1111/josh.12196

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*, 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Eo, S.-H., & Cho, H. (2014). Tree-structured mixed-effects regression modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, *23*, 740–760. https://doi.org/10.1080/10618600.2013.794732

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models. *Journal of Machine Learning Research*, *20*, 1–81.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment–subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods*, *50*, 2016–2034. https://doi.org/10.3758/s13428-017-0971-x

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*, 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*, 916–954. https://doi.org/10.1214/07-AOAS148

Fu, W., & Simonoff, J. S. (2015). Unbiased regression trees for longitudinal and clustered data. *Computational Statistics & Data Analysis*, *88*, 53–74. https://doi.org/10.1016/j.csda.2015.02.004

Galimberti, G., & Montanari, A. (2002). Regression trees for longitudinal data with time-dependent covariates. In K. Jajuga, A. Sokołowski, & H. H. Bock (Eds.), *Classification, clustering, and data analysis: Recent advances and applications* (pp. 391–398). Springer. https://doi.org/10.1007/978-3-642-56181-8_43

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research*, *22*, 364–375. https://doi.org/10.1177/0049124194022003005

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Guo, Y., Dhaliwal, J., & Rights, J. D. (2024). Disaggregating level-specific effects in cross-classified multilevel models. *Behavior Research Methods*, *56*, 3023–3057. https://doi.org/10.3758/s13428-023-02238-7

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, *81*, 451–459. https://doi.org/10.1016/j.spl.2010.12.003

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*, 1313–1328. https://doi.org/10.1080/00949655.2012.741599

Harris, K. M. (2013). *The add health study: Design and accomplishments*. Carolina Population Center, University of North Carolina at Chapel Hill. https://doi.org/10.17615/C6TW87

Harris, K. M., & Halpern, C. T. (2022). The National Longitudinal Study of adolescent to adult health (add health): An underused resource for developmental science. *Annual Review of Developmental Psychology*, *4*, 297–318. https://doi.org/10.1146/annurev-devpsych-120920-042217

Harris, K. M., Halpern, C. T., Whitsel, E. A., Hussey, J. M., Killeya-Jones, L., Tabor, J., & Dean, S. C. (2019). Cohort profile: The National Longitudinal Study of adolescent to adult health (add health). *International Journal of Epidemiology*, *48*, 1415–1425. https://doi.org/10.1093/ije/dyz115

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*, 297–310. https://doi.org/10.1214/ss/1177013604

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. (2nd ed.) Springer.

Jacobucci, R., Grimm, K. J., & Zhang, Z. (2023). *Machine learning for social and behavioral research*. Guilford Press.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 3146–3154).

Kilian, P., Ye, S., & Kelava, A. (2023). Mixed effects in machine learning—A flexible mixedML framework to add random effects to supervised machine learning regression. *Transactions on Machine Learning Research*. https://openreview.net/pdf?id=rQBXj8cAAAAJ

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Krennmair, P., & Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *71*, 1865–1894. https://doi.org/10.1111/rssc.12600

Kuhn, M., & Johnson, K. (2018). *Applied predictive modeling*. (2nd ed.) Springer.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963–974.

Maas, C. J. M., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, *46*, 427–440. https://doi.org/10.1016/j.csda.2003.08.006

Molnar, C. (2019). *Interpretable machine learning*. Lulu.com.

Perreira, K. M., Deeb-Sossa, N., Harris, K. M., & Bollen, K. (2005). What are we measuring? An evaluation of the CES-D across race/ethnicity and immigrant generation. *Social Forces*, *83*, 1567–1601. https://doi.org/10.1353/sof.2005.0077

R Core Team. (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*, 385–401. https://doi.org/10.1177/014662167700100306

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. (2nd ed.) SAGE Publications.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*, 913–929. https://doi.org/10.1111/ecog.02881

Salditt, M., Humberg, S., & Nestler, S. (2023). Gradient tree boosting for hierarchical data. *Multivariate Behavioral Research*, *58*, 911–937. https://doi.org/10.1080/00273171.2022.2146638

Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, *87*, 407–418. https://doi.org/10.1080/01621459.1992.10475220

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, *86*, 169–207. https://doi.org/10.1007/s10994-011-5258-3

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. (2nd ed.) Sage.

Speiser, J. L., Wolf, B. J., Chung, D., Karvellas, C. J., Koch, D. G., & Durkalski, V. L. (2019). BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomes. *Chemometrics and Intelligent Laboratory Systems*, *185*, 122–134. https://doi.org/10.1016/j.chemolab.2019.01.002

Wang, J., Gamazon, E. R., Pierce, B. L., Stranger, B. E., Im, H. K., Gibbons, R. D., Cox, N. J., Nicolae, D. L., & Chen, L. S. (2016). Imputing gene expression in uncollected tissues within and beyond GTEx. *The American Journal of Human Genetics*, *98*, 697–708. https://doi.org/10.1016/j.ajhg.2016.02.020

Appendix A: A numerical illustration of the *C*-projection

This example illustrates why the *C*-projection is needed in the iterative LMM–XGBoost algorithm.

Consider $N = 4$ observations cross-classified by two schools (A, B) and two neighborhoods $(1, 2)$:

Obs n	1	2	3	4
School	A	A	B	B
Neighborhood	1	2	1	2

Suppose a school-level predictor to be preserved in the learner is

$$w = (0,0,1,1)^{\top},$$

which is constant within schools but varies between schools. The preserved design matrix therefore includes an intercept and this school-level predictor:

$$X_g = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

The random-effect design matrices encode cluster membership. The school random-effect design matrix is

$$Z_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},$$

which indicates whether an observation belongs to school A or B.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

The neighborhood random-effect design matrix is

$$Z_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

which indicates whether an observation belongs to neighborhood 1 or 2. Thus, Z_2 encodes neighborhood cluster membership, whereas w is a school-level predictor.

The combined random-effect design matrix is

$$Z = [\, Z_1 \; Z_2 \,].$$

Now suppose that, at a given iteration, XGBoost fitted to the working residuals produces the raw prediction

$$\tilde{f} = \begin{bmatrix} 1 \\ -1 \\ 3 \\ 1 \end{bmatrix}.$$

This can be decomposed as

$$\tilde{f} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ 2 \\ 2 \end{bmatrix}}_{\text{school-level predictor effect}} + \underbrace{\begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}}_{\text{pure neighborhood pattern}}.$$

Thus, the learner has captured two components: (a) a desired effect of the school-level predictor w and (b) a cluster-level neighborhood pattern that should instead be represented by the random effects.

Without projection: If $\tilde{f}$ is inserted directly into the offset, then the learner absorbs the neighborhood-specific pattern

$$(1, -1, 1, -1)^\top,$$

which is aligned with the neighborhood clustering. In this case, part of the between-neighborhood structure that should be attributed to the random effects is instead absorbed by the machine-learning component.

Direct projection onto the orthogonal complement of Z: A natural but overly aggressive alternative is

$$\tilde{f}_\perp^{\text{naive}} = (I - P_Z)\tilde{f}, \qquad P_Z = Z(Z^\top Z)^+ Z^\top.$$

However, in this example, both the pure neighborhood pattern and the school-level predictor effect lie in the span of Z. Hence,

$$(I - P_Z)\tilde{f} = 0.$$

This removes not only the undesired cluster effect but also the desired school-level predictor effect. Therefore, direct orthogonalization against Z discards valid group-constant signal.

The C-projection: The C-projection preserves the signal in X_g and removes only the part aligned with the residualized cluster design. Define

$$M_g = I - X_g(X_g^\top X_g)^+ X_g^\top, \qquad \tilde{Z} = M_g Z.$$

Because Z is first residualized with respect to X_g, the school-level predictor effect is protected. In this example, the residualized cluster space is spanned by the neighborhood contrast

$$(1, -1, 1, -1)^\top.$$

Therefore,

$$\tilde{f}_\perp = \left(I - \tilde{Z}(\tilde{Z}^\top \tilde{Z})^+ \tilde{Z}^\top\right)\tilde{f} = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 2 \end{bmatrix}.$$

Hence, the C-projection removes the pure cluster-level neighborhood pattern while preserving the effect of the school-level predictor w.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Why the residualized cluster space reduces to the neighborhood contrast: The columns of Z_1 (the school indicator matrix) are already contained in $\mathrm{col}(X_g)$, since both Z_1 and X_g encode the same school-level grouping. Consequently, $M_g Z_1 = 0$, and the residualized design $\tilde{Z} = M_g Z = M_g[\, Z_1 \; Z_2 \,]$ reduces to $[\, 0 \; M_g Z_2 \,]$.

The matrix $M_g Z_2$ has rank 1 and its column space is spanned by the neighborhood contrast $(1, -1, 1, -1)^\top$, which alternates sign by neighborhood while being constant within each school. This is precisely the component of the neighborhood structure that is not already explained by the school-level predictor, and it is this component alone that the C-projection removes from $\tilde{f}$.

This example illustrates two competing goals: (a) preventing the residual learner from reconstructing cluster effects already represented by the random effects and (b) preserving valid signal from group-constant predictors intentionally included in the learner.

Without projection, the learner can absorb cluster-level structure. With naive projection onto Z, valid group-constant predictor effects can also be removed. The C-projection achieves the desired compromise by preserving the span of X_g while orthogonalizing against the residualized cluster design.

Appendix B: Simulation study: True tree structure

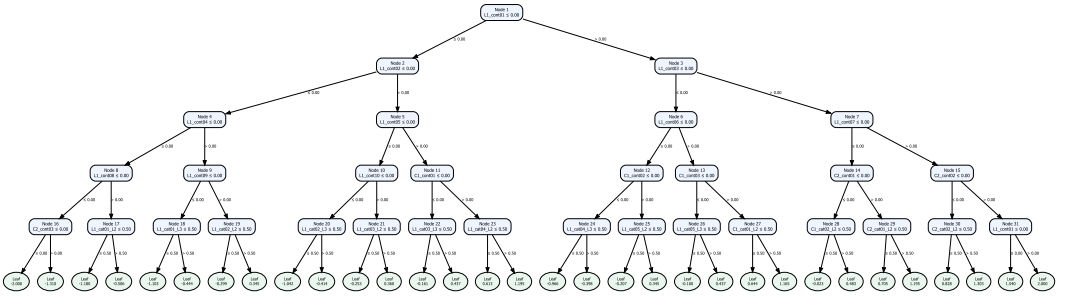


Figure B1. True tree structure in the simulation study.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Appendix C: Paper selection for literature review on using Add Health data

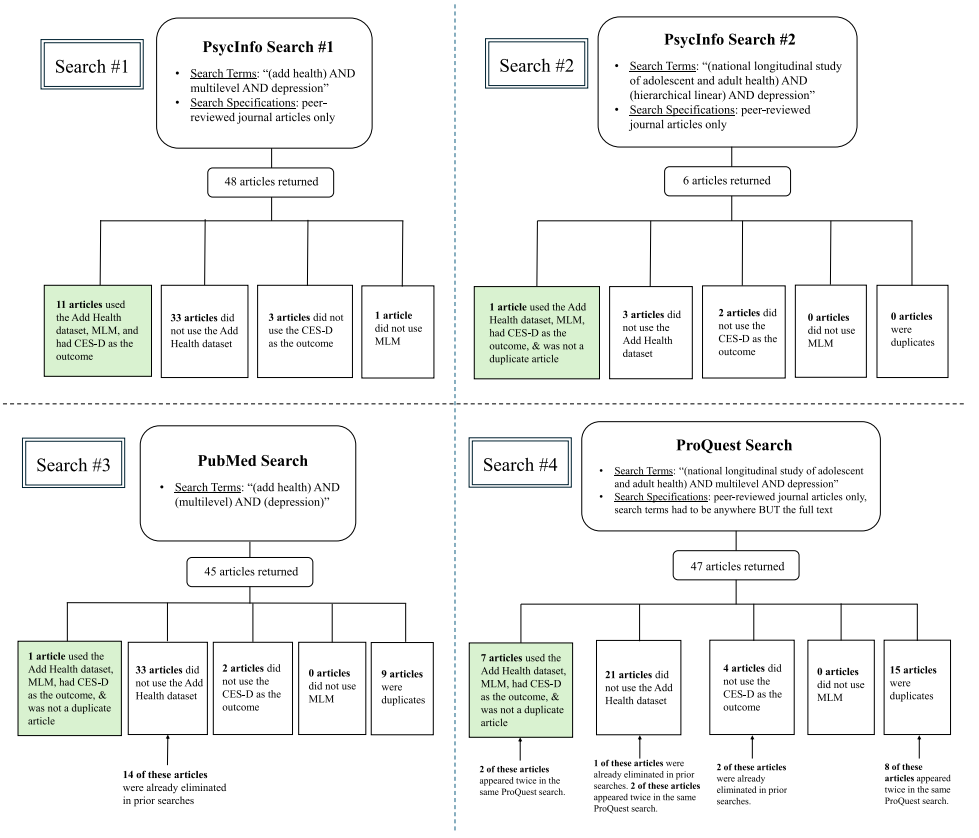


Figure C1. Paper selection for literature review on using Add Health data.
The literature review results are posted on the Open Science Framework, https://osf.io/jkctg/.

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Appendix D: Empirical study: Descriptive statistics of predictors

Table D1. Descriptive statistics of predictors in the empirical study.

Variable		Frequency (%)	Median (SIQR)
Demographic characteristics			
1. Age (`age`)			16(1)
2. Gender (`gender`)	Male	6,645(36.40)	
	Female	7,168(39.26)	
	Unavailable	5(0.03)	
3. Race (`race`)	White	374(2.05)	
	Black or African American(A)	255(1.40)	
	A Indian or Native A	77(0.42)	
	Asian or Pacific Islander	94(0.51)	
	Other	81(0.44)	
	Unavailable	17,375(95.17)	
4. Ethnicity (Hispanic or Latino origin) (`ethnicity`)	No	15,122(82.83)	
	Yes	3,075(16.84)	
	Unavailable	59(0.32)	
5. Attending religious activities* (`relactivities`)			2.5(1)
6. Religiosity (`religiosity`)			2(1)
Psychological well-being & cognition			
7. Self-esteem scores* (`selfesteem`)			12(2)
Education			
8. Grade in school (`grade`)	Grades 7 and 8	4,710(25.80)	
	Grades 9–12	13,098(71.75)	
	Unavailable	448(2.45)	

(Continued)

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Table D1. Continued.

Variable		Frequency (%)	Median (SIQR)
9. School connectedness* (`connectedness`)			10(14)
10. Trouble getting along with teachers (`alongwithtch`)			1(2)
11. GPA* (`overall GPA`)			3(1.25)
Family			
12. Parent–adolescent relationship with mom* (`PArelation.mom`)			12(1)
13. Parent–adolescent relationship with dad* (`PArelation.dad`)			13(8)
14. Parental education (`Peducation`)	College graduate	3,581(19.62)	
	High school graduate	4,557(24.96)	
	Less than high school	2,767(15.16)	
	Some college	4,566(25.01)	
	unavailable (refused)	15(0.08)	
15. Family/household structure (`structure`)	No	558(3.06)	
	Yes	168(0.92)	
	Unavailable	12,800(70.11)	
16. Family income (`familyincome`)			38(19.5)
17. Parent occupation (`Poccupation`)	Professional/managerial	5,939(32.53)	
	Other professional	953(5.22)	
	Sales, service, and administration	4,795(26.27)	
	Manual/blue collar	3,343(18.31)	
	Other (unspecified)	1,618(8.86)	
	Not working	1,192(6.53)	
	Unavailable	416(2.28)	

(Continued)

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Table D1. Continued.

Variable		Frequency (%)	Median (SIQR)
18. Language spoken at home (homelang)	English	16,182(88.64)	
	Spanish	1,433(7.85)	
	Other	631(3.46)	
	Unavailable	10(0.05)	
19. Parent age (parentage)			41(4.5)
20. Marital status of parents (Pmarital)	Single, never married	905(4.96)	
	Married	109,68(60.08)	
	Widowed	562(3.08)	
	Divorced	2,308(12.64)	
	Separated	810(4.44)	
	Unavailable	11(0.06)	
21. Parental attachment (attachment)			5(0)
22. The number of household members (numhousemem)			4(2)
Friends and social network			
23. Social support* (socialsupport)			32(3)
24. Feeling cared about by teachers (feelingT)			4(0.5)
25. Participation in sports (partinsports)			0(1)
26. Participation in non-sport extracurriculars (nonsportec)			0(1)
Individual-level contextual factors			
27. Reason for living in the neighborhood (school) (neighsch)			
	No	8,048(44.08)	
	Yes	7,234(39.63)	
	Unavailable	81(0.44)	

(Continued)

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Table D1. Continued.

Variable		Frequency (%)	Median (SIQR)
28. Neighborhood interaction (neighinteraction)	True	14,183(77.69)	
	False	3,968(21.74)	
	Unavailable		
29. Parental-perceived neighborhood collective efficacy* (neighborefficacy)			3(1)
30. Parental-perceived neighborhood disorder* (neighbordisorder)			4(1)
31. Perceived neighborhood social cohesion* (neighborcohesion)			3(0.5)
32. Perceived neighborhood safety (neighsafety)	No	6,065(33.22)	
	Yes	9,226(50.54)	
	Unavailable	81(0.44)	
Medication & substance use and abuse			
33. Alcohol problems* (alcohol)			0(0.5)
Risk behavior			
34. Delinquent behaviors* (delinquent)			0(0.5)
Physical health			
35. Migraine status of adolescent (migraine)	No	13,676(74.91)	
	Yes	1,772(9.71)	
	Unavailable	184(1.01)	
Socio-economic status, labor market, & occupation			
36. Employment for adolescent (Aemployment)	No	7,928(43.43)	
	Yes	10,234(56.06)	
	Unavailable	94(0.51)	

Note: Totals and percentages may not add up to the total sample size and 100, respectively, due to missing data and rounding. For continuous predictors, responses of "multiple response," "does not apply," "refused," "don't know," "legitimate skip," and "not applicable" are coded as "NA" (missing) when calculating sum scores. For categorical predictors, responses of "multiple response," "does not apply," "refused," "don't know," "legitimate skip," and "not applicable" are coded as "unavailable." Predictors with the asterisk indicate sum scores of multiple predictors based on the literature.

Appendix E: Empirical study: Fixed-effect estimates of the LMMs

Table E1. Fixed-effect estimates of the LMMs in the empirical study.

	Unconditional LMM		Conditional LMM	
	EST	SE	EST	SE
intercept	**3.085**	0.022	**3.365**[0.18]	0.055
alongwithtch	-		**0.059**[0.05]	0.007
partinsports	-		−0.003[−2.48e-03]	0.008
nonsportec	-		−0.016[−0.01]	0.008
numhousemem	-		**0.032**[0.03]	0.011
overall GPA	-		−**0.097**[−0.08]	0.008
age	-		**0.046**[0.04]	0.011
PArelation.mom	-		**0.018**[0.02]	0.008
PArelation.dad	-		−0.005[−3.94e-03]	0.011
feelingT	-		**0.052**[0.04]	0.009
selfesteem	-		**0.425**[0.36]	0.008
familyincome	-		−**0.017**[−0.01]	0.008
socialsupport	-		−**0.314**[−0.26]	0.012
neighborefficacy	-		−**0.033**[−0.03]	0.009
neighbordisorder	-		**0.033**[0.03]	0.009
neighborcohesion	-		**0.045**[0.04]	0.011
parentage	-		0.009[7.323 – 03]	0.009
alcohol	-		**0.072**[0.06]	0.008
delinquent	-		**0.068**[0.06]	0.008
religiosity	-		−0.008[−6.59e-03]	0.007
attachment	-		**0.041**[0.03]	0.009
ethnicity1	-		0.013[0.01]	0.026
ethnicity9	-		−0.218[−0.18]	0.131
grade2	-		−**0.063**[−0.05]	0.028
grade9	-		−0.003[−2.58e-03]	0.054
Peducation2	-		−0.050[−0.04]	0.038
Peducation3	-		−**0.137**[−0.12]	0.036
Peducation4	-		−**0.156**[−0.13]	0.037
Peducation5	-		−**0.192**[−0.16]	0.039
Peducation6	-		−0.947[−0.80]	1.005
Peducation9(refused)	-		−0.100[−0.08]	0.253
Aemployment1	-		−**0.037**[−0.03]	0.015
Aemployment9	-		−**1.534**[−1.30]	0.183
neighinteraction2	-		−**0.088**[−0.07]	0.025

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Table E1. Continued.

	Unconditional LMM		Conditional LMM	
	EST	SE	EST	SE
neighinteraction9	-		−**0.422**[−0.36]	0.174
Poccupation2	-		−0.022[−0.02]	0.034
Poccupation3	-		**0.051**[0.04]	0.021
Poccupation4	-		**0.100** [0.08]	0.024
Poccupation5	-		0.038[0.03]	0.029
Poccupation6	-		**0.077**[0.06]	0.035
Poccupation9	-		−0.111[−0.09]	0.060
homelang2	-		−0.015[−0.01]	0.039
homelang3	-		**0.202**[0.17]	0.041
homelang9	-		−**1.280**[−1.08]	0.331
Pmarital2	-		−**0.161**[−0.14]	0.035
Pmarital3	-		−0.054[−0.05]	0.051
Pmarital4	-		−**0.118**[−0.10]	0.037
Pmarital5	-		−0.059[−0.05]	0.045
Pmarital6	-		0.330[0.28]	0.302
neighsch1	-		0.011[9.01e-03]	0.016
neighsch9	-		−0.116[−0.10]	0.114
neighsafety1	-		0.005[3.97e-03]	0.017
neighsafety9	-		**0.251**[0.21]	0.113
migraine1	-		**0.086**[0.07]	0.024
migraine9	-		**0.144**[0.12]	0.071
gender2	-		**0.177**[0.15]	0.015
gender9	-		0.250[0.21]	0.430

Note: The levels of categorical predictors are based on information available on the Add Health website and in the codebooks. Responses of "multiple response," "does not apply," "refused," "don't know," "legitimate skip," and "not applicable" are coded as "unavailable (9)."

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press

Appendix F: Empirical study: Residual analyses of LMM with linear, no-interaction effects

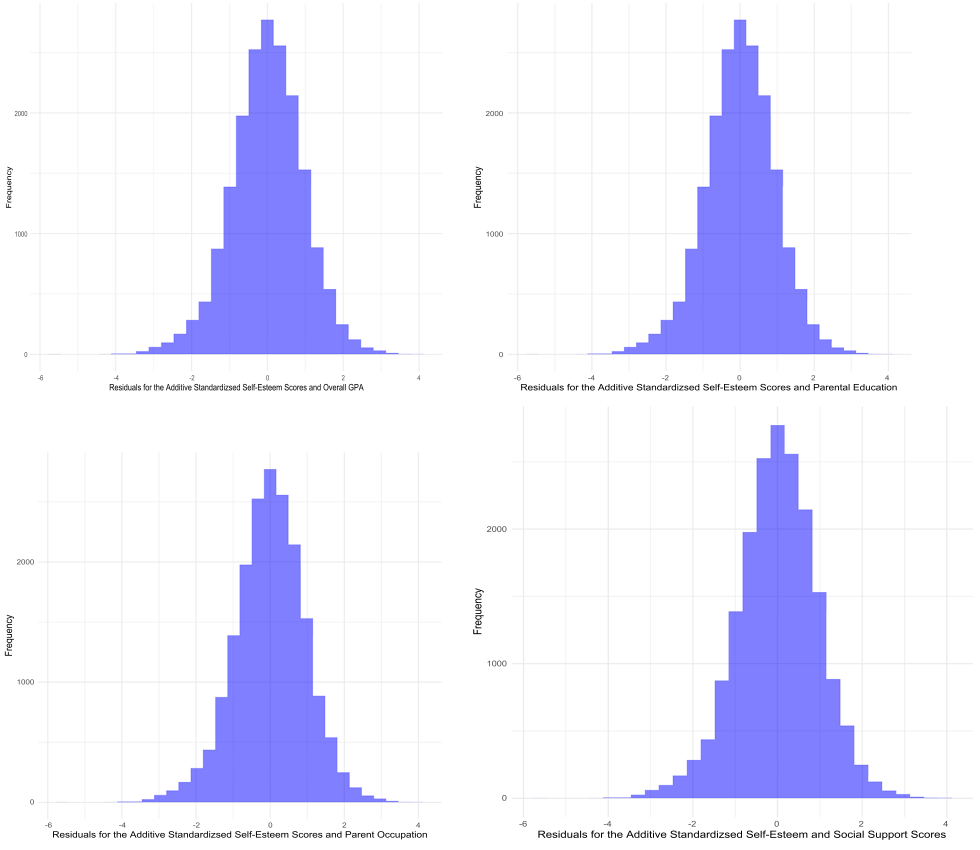


Figure F1. Residual analyses of LMM.

Cite this article: Cho, S.-J. and Mueller, S. (2026). Mixed-Effects XGBoost with Group-Aware Permutation Importance and Cross-Validation for Multilevel Cross-Classified Continuous Outcomes. *Psychometrika*, 1151–1190. https://doi.org/10.1017/psy.2026.10108

https://doi.org/10.1017/psy.2026.10108 Published online by Cambridge University Press