Behavior Research Methods
https://doi.org/10.3758/s13428-021-01709-z

Level-specific residuals and diagnostic measures, plots, and tests for random effects selection in multilevel and mixed models

Sun-Joo Cho[1] · **Paul De Boeck**[2,3] · **Matthew Naveiras**[1] · **Hope Ervin**[4]

Accepted: 13 September 2021

Abstract

Multilevel data structures are often found in multiple substantive research areas, and multilevel models (MLMs) have been widely used to allow for such multilevel data structures. One important step when applying MLM is the selection of an optimal set of random effects to account for variability and heteroscedasticity in multilevel data. Literature reviews on current practices in applying MLM showed that diagnostic plots are only rarely used for model selection and for model checking. In this study, possible random effects and a generic description of the random effects were provided to guide researchers to select necessary random effects. In addition, based on extensive literature reviews, level-specific diagnostic plots were presented using various kinds of level-specific residuals, and diagnostic measures and statistical tests were suggested to select a set of random effects. Existing and newly proposed methods were illustrated using two data sets: a cross-sectional data set and a longitudinal data set. Along with the illustration, we discuss the methods and provide guidelines to select necessary random effects in model-building steps. R code was provided for the analyses.

Keywords Diagnostic plots · Level-specific residuals · Mixed-effects model · Multilevel model · Random effect selection

Introduction

In multiple substantive research areas, data are often collected from clusters (e.g., Raudenbush & Bryk, 2002; Snijders & Bosker 1999). As an example, a random sample of hospitals (clusters) is selected, and then patients (observations) from the selected hospitals are randomly sampled. Furthermore, it is common to have a longitudinal design in which individuals (clusters) are observed over time (observations). To account for between-cluster variation, a multilevel model (MLM, e.g., Goldstein, 2003) has been widely applied. An MLM for continuous outcomes is also referred to as a random effect model (Laird & Ware, 1982), a hierarchical linear model (e.g., Bryk & Raudenbush, 1992), a linear mixed-effects model (LMM, e.g., McCulloch et al. 2008), a random regression model (Bock, 1983), or a random coefficient model (e.g., de Leeuw & Kreft, 1986; Longford 1993).

In MLM specifications, the between-cluster variation is represented by *random effects* such as a random intercept and a random slope of a covariate. In the MLM literature, a quantity is considered random if it varies over clusters within a population, in which case the set of observed clusters can be interpreted as a random sample (e.g., Snijders & Bosker 1999, Section 4.2). The random intercept can be considered to model random variation across clusters, and the random slope can be used to model random variation of a covariate effect within the population of clusters. The primary interest of many MLM applications is in the estimation of fixed effects and their standard errors (Raudenbush & Bryk, 2002, p. 253), although we acknowledge that random effects (e.g., individual differences) can be of interest in other MLM applications. In many cases, the interest in random effects is auxiliary, to obtain accurate standard errors for the fixed effects. When necessary random effects are not included in MLM to model all sources of variability and heteroscedasticity in the data, standard errors of the fixed effects of interest are typically negative biased (see Longford 1993, pp. 53–56 for technical details). This bias leads to overestimating the statistical significance of the fixed effects.

✉ Sun-Joo Cho
sj.cho@vanderbilt.edu

[1] Vanderbilt University, Nashville, TN, USA

[2] The Ohio State University, Columbus, OH, USA

[3] KU Leuven, Leuven, Belgium

[4] Vanderbilt University Medical Center, Nashville, TN, USA

Inclusion of a random intercept is often justified by a sufficiently large intraclass correlation (ICC, Shrout & Fleiss, 1979) based on an unconditional MLM (i.e., a random intercept model without any covariates). After including the random intercept, the next step is to investigate whether covariate effects need to be included, with a fixed effect and possibly with random effects as well (i.e., random slopes). In addition to the random intercept and random slopes, other kinds of random effects have been discussed to model heteroscedasticity (as described in detail below), although these random effects are rarely considered in practice. It is common to compare candidate models with different random effects (e.g., random intercept vs. random intercept-slope model) and to select a model based on likelihood ratio tests (LRT).[1]

As a supplement to LRT in selecting random effects, diagnostic plots such as a residual plot, a scatter plot, and a normal probability plot can be used. These diagnostic plots can be used to explore missing random effects not captured by the model. For example, a scatter plot of residuals versus fitted values can be used to explore heteroscedasticity in residuals (variance changes within clusters). A non-random pattern in the plot such as a wedge-shaped pattern can be indicative of heteroscedasticity (e.g., Pinheiro & Bates, 2000, p. 341). In addition, the diagnostic plots can be used for assessing model assumptions in MLM such as homogeneity of residual variance, linearity, and normality of residuals. Model checking through diagnostic plots can be informative when selecting random effects. For instance, a diagnostic plot can be used to explore heteroscedasticity in residuals when considering adding a random effect to model heteroscedasticity. In the statistics literature for LMM, different kinds of residuals and diagnostic plots have been suggested for model selection and model checking (e.g., Galecki & Burzykowski, 2013, pp. 264–266, pp. 339–346; Pinheiro & Bates, 2000, Ch. 4). However, our literature review of current practices in using MLM shows that the diagnostic plots in selecting random effects and model checking are rarely used (also noted in Claeskens [2013, p. 442] and O'Connell et al. [2016, p. 99]).[2]

We identify the following problems in the current practices of using diagnostic plots mostly based on residuals for random effect selection in MLM applications in the social sciences. First, to the best of our knowledge, there are no publications in which an exhaustive list of random effects has been presented. It is not easy for substantive researchers to be aware of all possible random effects to be considered for model selection. Second, level-specific residuals have been developed in the statistics literature on LMM (Hilden-Minton, 1995; Loy & Hofmann, 2014; Pinheiro & Bates, 2000; Verbeke & Lesaffre, 1997) which have not been introduced in MLM textbooks for the social sciences (e.g., Goldstein, 2003; Hox, Moerbeek, & van de Schoot, 2018; Longford, 1993; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).[3] Substantive researchers are not always aware of the available range of options. Third, the scaling of residuals (standardized vs. unstandardized) and the definitions of conditional vs. marginal residuals are possible sources of confusion. For example, Snijders and Bosker (1999, p. 129) suggested using unstandardized (individual-level) residuals to check the linearity effect of (an individual-level) covariate, whereas Hox, Moerbeek, and van de Schoot (2018) used standardized (individual-level) residuals to check the linearity. As far as we know, it has not been discussed what kind of residuals (e.g., unstandardized vs. standardized; conditional vs. marginal) should be used when. Fourth, many of the model diagnostics are graphical in nature, and interpretations of patterns in diagnostic plots can be subjective (McCullagh & Nelder, 1989, pp. 392–393). To enhance the detection of visual patterns in the diagnostic plots, it has been suggested to consider including smoothing functions in the plot (e.g., Snijders & Berkhof, 2007). For instance, the scatter plot of (individual-level) residuals versus a covariate can be smoothed using spline functions (Snijders & Bosker, 1999). However, statistical tests for the patterns are rarely conducted. To summarize, unanswered questions are (a) what kind of residuals should be used for different kinds of diagnostic plots when selecting all necessary effects and checking model assumptions (as information to select a set of the random effects), and (b) how should the visual patterns in the diagnostic plots be tested.

Purpose of the current study

The purpose of the current study is to overcome the problems we listed above. Specifically, first, possible random effects which can be included in the model are presented and a generic description of those random effects is provided. Second, an extensive literature review on

[1]To review current practices of using diagnostic plots and model selection methods regarding random effects, 72 papers were randomly selected from nine APA journals through the PsychINFO database. We found that random effects were selected based on LRT (33%), Wald test (26%), goodness-of-fit statistics (13%), information criteria (2%), and pseudo R-square (2%). Twelve papers (17%) did not consider a model selection regarding random effects.

[2]Of the 72 papers we reviewed, there was one paper which presented a diagnostic plot to investigate autoregressive effects.

[3]Exceptionally, O'Connell, Yeomans-Maldonado, and McCoach (2016) listed conditional and marginal residuals for education researchers. In this study, we added one more kind of residual called independent residuals, based on extensive reviews in the statistics literature.

residuals and diagnostic plots in LMM and MLM literature is conducted for model selection regarding random effects and for model checking. The review is based on four LMM texts (Faraway, 2016; Galecki & Burzykowski, 2013; Pinheiro & Bates, 2000; Verbeke & Molenberghs, 2000) and 9 MLM texts, handbooks, edited books, and book chapters (Finch, Bolin, & Kelley, 2014; Goldstein, 2003; Hox et al., 2018; Longford, 1993; Raudenbush & Bryk, 2002; Singer & Willett, 2003; Snijders & Bosker, 1999; Snijders & Berkhof, 2007; O'Connell, Yeomans-Maldonado, & McCoach, 2016). Third, specific kinds of residuals and diagnostic plots are presented to explore random effects and to check model assumptions, and inference methods are also presented to test patterns in diagnostic plots. Finally, in addition to diagnostic plots, we will also propose diagnostic measures to select an optimal set of random effects. All these proposed methods are presented and illustrated for a two-level design involving individual- and cluster-level units. Hereafter, we refer to the individual level as *level 1* and the cluster level as *level 2* throughout this paper. Generalizability to other multilevel designs will be discussed. Parameter estimation of MLMs is conducted using the popular nlme R package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2020). We chose the nlme package because it allows all kinds of random effects we discussed in this paper to be modeled. The R code used in this paper is presented in Appendix A.

The rest of this paper is organized as follows. In "Different kinds of random effects in multilevel models", we describe MLMs for cross-sectional and longitudinal data, list all kinds of random effects in MLM, and describe model-building steps. In "Illustrative data sets", two empirical data sets are described for illustration. In "Level-specific residuals", the literature review on the types of residuals in MLM is presented. In "Diagnostic measures, diagnostic plots, and statistical tests", we suggest diagnostic measures, list diagnostic plots for random effect selection and model assumption checks, and discuss reasons for the kind of residuals to be used in diagnostic plots. In addition, we present statistical inference on patterns in the diagnostic plot. In "Illustration", we illustrate the proposed methods using two empirical data sets. Finally, we end with a summary and a discussion.

Different kinds of random effects in multilevel models

In this section, we describe MLMs for cross-sectional and longitudinal data, list all kinds of random effects in MLM, and describe the model-building steps to be used in the selection of a set of random effects.

Multilevel models

An MLM with design matrices as in LMM is written as

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\epsilon}_j, \tag{1}$$

where j is an index for (non-overlapping) clusters; $\mathbf{y}_j$ is a vector of continuous responses; $\mathbf{X}_j$ is the design matrix of the fixed effects; $\mathbf{Z}_j$ is the design matrix of the random effects; $\boldsymbol{\beta}$ is the vector of fixed effects; $\mathbf{b}_j$ is the vector of random effects; and $\boldsymbol{\epsilon}_j$ is the vector of random residuals. The random effects are assumed to follow a multivariate normal distribution, $\mathbf{b}_j \sim MVN(\mathbf{0}, \Sigma)$, where Σ is a variance-covariance matrix of the random effects. In addition, the random residuals are assumed to follow a multivariate normal distribution, $\boldsymbol{\epsilon}_j \sim MVN(0, \mathcal{R}_j)$. The residual variance-covariance matrices $\mathcal{R}_j$ can be decomposed into two independent components, a variance component (σ^2) and a correlation component (R_j):

$$\mathcal{R}_j = \sigma^2 R_j. \tag{2}$$

For the cross-sectional data, the residual variance-covariance matrices are assumed to have a homoscedastic conditional independent structure:

$$\mathcal{R}_j = \sigma^2 R_j = \sigma^2 \mathbf{I}_{n_j}, \tag{3}$$

where n_j is the cluster size (i.e., the number of level-1 units within a cluster j). However, for the longitudinal data in which outcomes are collected repeatedly from the same individuals (i.e., clusters), it is common to model correlated errors (Galecki & Burzykowski, 2013; Pinheiro & Bates, 2000, Section 5.3.1; Verbeke & Lesaffre, 1997):

$$\mathcal{R}_j = \sigma^2 R_j = \sigma^2 \Lambda_j \mathbf{C}_j \Lambda_j, \tag{4}$$

where Λ_j is a diagonal matrix with nonnegative diagonal elements and $\mathbf{C}_j$ is a correlation matrix. The Λ_j allows for heteroscedasticity of observations within individual (cluster) j and $\mathbf{C}_j$ allows for correlation between the observations within the individual (Galecki & Burzykowski, 2013, p. 179). Various kinds of correlation matrices $\mathbf{C}_j$ can be specified for longitudinal data such as uniform correlation and correlations with an autoregressive (AR) component of order p and a moving average (MA) component of order q (ARMA(p, q)), or a continuous-time autoregressive process (Pinheiro & Bates, 2000).

The conditional distribution, $f_{y|b}(\mathbf{y}_j|\mathbf{b}_j)$, of $\mathbf{y}_j$ given $\mathbf{b}_j$ is multivariate normal, with mean and variance written as:

$$E(\mathbf{y}_j|\mathbf{b}_j) = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j \tag{5}$$

and

$$Var(\mathbf{y}_j|\mathbf{b}_j) = \sigma^2 R_j. \tag{6}$$

The marginal distribution, $f_y(\mathbf{y}_j)$, of $\mathbf{y}_j$, is obtained by integrating out the random effects $\mathbf{b}_j$ from the joint distribution of $\mathbf{y}$ and $\mathbf{b}_j$:

$$f_y(\mathbf{y}_j) = \int f_{y|b}(\mathbf{y}_j|\mathbf{b}_j) f_b(\mathbf{b}_j) d\mathbf{b}_j, \tag{7}$$

where $f_{y|b}(\mathbf{y}_j|\mathbf{b}_j)$ is the conditional distribution and $f_b(\mathbf{b}_j)$ is the density of the unconditional distribution of $\mathbf{b}_j$. The marginal distribution is also multivariate normal, with mean and variance written as:

$$E(\mathbf{y}_j) = \mathbf{X}_j\boldsymbol{\beta} \tag{8}$$

and

$$Var(\mathbf{y}_j) = \sigma^2 V_j = \sigma^2 \mathbf{Z}_j D \mathbf{Z}_j' + \sigma^2 R_j, \tag{9}$$

where D is the variance-covariance matrix of random effects, $\mathbf{b}_j$.

Random effects in MLM

We will use different elements to describe different kinds of random effects: (a) data modes and (b) the unit of random effect, and (c) the unit of variation.

Data modes Each of the *levels* in a multilevel structure is considered as a source of variation (Longford, 1993, p. 18). Data modes refer to kinds of units which may be a source of variation (Coombs, 1964). For example, students and schools are data modes and at the same time they are the levels in the two-level design in which students are nested within schools.

Unit of random effect and unit of variation The unit of random effect (UR) is a covariate of a mode (e.g., students) which induces variation across the units of (mostly but not always) another mode (e.g., schools), denoted as unit of variation (UV). For example, the gender of students (a covariate of students) may induce an effect that is random across schools (another mode than students). For the two-level data, the UR can be (a) all data entries (i.e., the 1-constant), (b) a level-1 covariate ($x^{(1)}$), and (c) a level-2 covariate ($x^{(2)}$). The UV is the set of elements across which the random effect of the UR varies. The UV can be individuals (e.g., students) and clusters (e.g., schools) in cross-sectional two-level data, and time points (e.g., weeks or years) and individuals (e.g., students) in longitudinal two-level data. For example, age (UR) can vary across students (UV). Together, the paired notions of UR and UV define a random effect. For notation, we propose $UR|UV$, inspired by random effect specifications in the `lme4` (Bates, Maechler, Bolker, & Walker, 2015) and `nlme` `R` packages.

For two-level data as an example, there are four kinds of random effects:

- Random intercept: The effect of the 1-constant can vary across clusters at level 2 (e.g., schools in a cross-sectional design, individuals in a longitudinal design).
- Random slope: The effect of a level-1 covariate ($x^{(1)}$) varies across clusters at level 2.
- Random effects with different variances to model heteroscedasticity: In general, heteroscedasticity refers to the pattern in which the variability of a variable is unequal across the range of values of a second variable that explains or predicts it.
  - An effect of a level-1 covariate ($x^{(1)}$) can vary across units of level 1.

    Level-1 heteroscedasticity is heteroscedasticity of the random residuals ($\boldsymbol{\epsilon}_j$ in Eq. 1). For a continuous level-1 covariate ($x^{(1)}$), the number of levels in the level-1 covariate should be less than the number of level-1 observations. For a categorial level-1 covariate ($x^{(1)}$), the variances of random residuals (σ^2) are modeled depending on the level of the categorical level-1 covariate which allows for heteroscedasticity. For example, gender as a level-1 covariate can create heterogeneity in that the variance across individuals of one gender may be different from the variance across individuals of the other gender.
  - An effect of a level-2 covariate ($x^{(2)}$) can vary across units of level 2 (i.e., clusters). As an example of heteroscedasticity, the variance of schools may be different depending on public vs. private as categories of a level-2 covariate.

Table 1 shows a list of all possible random effects in the two-level data, using our proposed notation $UR|UV$ for random effects. For the two-level cross-sectional data, level 1 (observation level) refers to individuals (e.g., students) and level 2 (cluster level) refers to clusters (e.g., schools). As shown in Table 1 (top), the following kinds of random effects listed above are as follows:

- Random intercept: $1|clusters$
- Random slope: $x^{(1)}|clusters$
- Random effects with different variances to model heteroscedasticity: Heteroscedasticity means that the variance of the random effects is allowed to differ depending on the values of the covariate in question.
  - The effect of a covariate at level 1 varies across level-1 observations: $x^{(1)}|individuals$
  - The effect of a covariate at level 2 varies across level-2 clusters: $x^{(2)}|clusters$

For the two-level longitudinal data, the level-1 units are time points and the level-2 units are individuals. As

Table 1 Different kinds of random effects in cross-sectional two-level data (*top*) and in longitudinal two-level data (*bottom*)

Unit of	Unit of Variation (UV)	
Random Effect (UR)	Individuals	Clusters
1		$1\|clusters$
$x^{(1)}$	$x^{(1)}\|individuals$	$x^{(1)}\|clusters$
$x^{(2)}$	–	$x^{(2)}\|clusters$
Unit of	Unit of Variation (UV)	
Random Effect (UR)	time	Individuals
1		$1\|individuals$
$x^{(1)}$	$x^{(1)}\|time$	$x^{(1)}\|individuals$
$x^{(2)}$	–	$x^{(2)}\|individuals$

–Indicates that a random effect cannot be modeled

presented in Table 1 (bottom), the following kinds of random effects listed above are as follows:

- Random intercept: $1|individuals$
- Random slope: $x^{(1)}|individuals$
- Random effects with different variances to model heteroscedasticity:
  - The effect of a covariate at level 1 varies across level-1 observations: $x^{(1)}|time$
  - The effect of a covariate at level 2 varies across level-2 clusters: $x^{(2)}|individuals$

Model-building steps

In the literature, it has been discussed how to proceed to check residuals. Either one starts with level 1 and continues to level 2 (i.e., upward approach, Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002) or one starts from the highest level and continues with each subsequent lower level (i.e., downward approach, Langford & Lewis, 1998 for outlier detection; Verbeke & Molenberghs, 2000). Snijders and Berkhof (2007) supported the upward approach for model assumption checking because level-1 residuals can be studied unconfounded by the higher-level residuals while the reverse is not possible (as noted in Hilden-Minton, 1995). However, the authors noted that checking level-2 outliers first is more efficient than checking level-1 outliers. Our reasoning behind this higher efficiency is that the number of clusters is smaller than the number of observations. It is inefficient to evaluate which level-1 units are outliers within a cluster that itself may be identified as a level-2 outlier. In our perspective, the argument of Hilden-Minton (1995) to work in the upward direction for residuals and the argument of Snijders and Berkhof (2007) to work downwards for outlier detection are persuasive. Thus, we take the upward approach for residual checking and the downward approach for outlier checking.

In MLM applications, necessary random effects are often selected in model building steps (e.g., Hox et al., 2018). As mentioned earlier, model checking through diagnostic plots can be informative when selecting random effects. In the following model-building steps, we discuss how both diagnostic plots and various tests of residuals can be used. The goal of model-building is to develop a parsimonious model that describes the data adequately while remaining interpretable. In the model-building steps below, we take a mixed approach including both (a) a confirmatory hypothesis testing for covariate(s) related to research questions and (b) an exploratory approach for the other covariates not related to the research questions.

- Step 0: A preliminary descriptive analysis is conducted without any modeling.
- Step 1: Random intercepts for the clusters are introduced as the only model component.
- Step 2: Fixed effects of level-1 covariates of interest are added to the random intercept model, such as the fixed effect of time in the longitudinal model. When the level-1 covariates are added to the random intercept model, level-1 linearity and level-1 heteroscedasticity can be explored. For the longitudinal data, correlated errors across time points can be investigated.
- Step 3: Random effects of level-1 covariates (i.e., random slopes) are added.
- Step 4: Fixed effects of level-2 covariates are added as well as their random effects (i.e., random effects to model level-2 heteroscedasticity).
- Step 5: A model with random effects and fixed effects of other covariates is selected based on goodness-of-fit criteria.

While building the model, we will keep track of outliers, influential points, and normality at level 1 and level 2, mostly through diagnostic plots and in some cases by using diagnostic measures we will introduce.

We suggest the following strategies, while going through the consecutive steps:

- From Step 2 to Step 4, outliers, influential points, and normality will be assessed to determine whether fixed and random effects of the model need to be included in the model.
- Outlying observations or clusters will not be removed before Step 4, because the outlying nature of an observation or cluster may change during the model building process. However, because outliers may have consequences for further steps, we will return to earlier steps after removing outliers.

- Non-normality will also be followed up in each step without making definite assessments until Step 4 is reached.

Illustrative data sets

We will use two empirical data sets in the following sections: a two-level cross-sectional data set and a two-level longitudinal data set, to illustrate level-specific diagnostic plots.

Example 1: two-level cross-sectional data (`Math` data)

A two-level cross-sectional data set was chosen from a popular MLM textbook (Kreft & de Leeuw, 1998; see pp. 58–60 for data description). It can be freely downloaded from http://www.bristol.ac.uk/cmm/learning/mmsoftware/data-rev.html. The dataset includes 519 students (level 1) nested within 23 schools (level 2) and an average cluster size of approximately 23. ICC was .243 (= 26.124/(26.124 + 81.244)), based on the unconditional random intercept model, which indicates that 24.3% of the total variation in math scores was explained by between-school variation. Rights (2019) considered the parents' highest level of education (i.e., level-1 covariate $x_{ij}^{(1)}$) to predict math scores. As in Rights (2019), a goal of analysis in this paper is to predict math scores from parents' highest level of education (called `parentHED` hereafter as a primary covariate; ranging from 1 to 6). Rights (2019) applied MLMs to the same dataset using random intercepts, random slopes, and a random effect for level-2 heteroscedasticity. As in Rights (2019), we consider the school-mean-centered parent education as a level-2 covariate ($x_{.j}^{(2)}$) and deviations of parent education from the level-2 covariate as a level-1 covariate ($x_{ij}^{(1)} - x_{.j}^{(2)}$) for an unconflated random slope and level-2 heteroscedasticity. In addition to `parentHED` as a primary covariate, level-1 covariates include socioeconomic status of parents (`ses`), the number of hours of homework done per week (`homework`; ranging from 0 to 6), and a student ethnicity covariate (`white`; 1=white and 0=non-white) as level-1 control covariates. Additional level-2 covariates include education sector (`public`; 1=public and 0=private), the percentage of ethnic minority students in the school (`percmin`), and class size measured by the student-teacher ratio (`ratio`) as level-2 control covariates.

Example 2: Two-level longitudinal data (`Hamilton depression [HD]` data)

A longitudinal psychiatric data set was chosen, and the data set has been used to illustrate the application of MLM to longitudinal data (Hedeker, 2004). The data set can be freely downloaded from https://stats.idre.ucla.edu/r/examples/alda/r-applied-longitudinal-data-analysis-ch-7/. The data set is originally from a study described in Reisby et al. (1977). Reisby et al. (1977) investigated the longitudinal relationship between imipramine (commonly prescribed for the treatment of major depression) and desipramine plasma levels. The study included responses of the Hamilton depression (HD) rating scale (Hamilton, 1960) from 66 depressed inpatients. Lower HD scores indicate lower degrees of depression. Among the 66 depressed inpatients, 29 were diagnosed with a nonendogenous depression associated with tragic life events and 37 were diagnosed with an endogenous depression not associated with any specific event. This nonendogenous vs. endogenous group variable (`Endog`) is considered a level-2 covariate.

In the study of Reisby et al. (1977), patients received 225-mg/day doses of imipramine for four weeks, following 1 week with a placebo: week 0 (start of placebo week), week 1 (end of placebo week), week 2 (end of first drug treatment week), week 3 (end of second drug treatment week), week 4 (end of third drug treatment week), and week 5 (end of fourth drug treatment week). Patients were rated with the HD rating scale twice at week 0 (at the start and end of week 0) and at the end of each week during the four treatment weeks. In this longitudinal example, the level-1 covariate `Week` is a primary covariate and the level-2 covariate `Endog` is a control covariate. Forty-six of the 66 patients completed all responses at all time points, and the number of patients with complete responses at each week varied: 61 at week 0, 63 at week 1, 65 at week 2, 65 at week 3, 63 at week 4, and 58 at week 5. Patients with missing HD scores were omitted prior to analysis and only complete data were used. ICC due to clusters (i.e., patients) was 0.268 (= 13.929/(13.929 + 37.957)), based on an unconditional random intercept model (i.e., a random intercept model without any covariates). This indicates that 26.8% of repeated measures is explained by between-patient variation.

Level-specific residuals

A residual is defined as the difference between the observed value of the outcome variable and the fitted (or predicted) value: Residual = Observed - Fitted Value (or Predicted Value). A zero residual means that the selected model explains or predicts the observation exactly and non-zero residuals indicate model-data misfit. For MLM, residuals can be specified at each level of the multilevel data. Below, we describe various kinds of level-specific residuals.

Level-1 residuals

Ordinary least squares (OLS) regression for each cluster separately has been recommended for the analysis of level-1 residuals by Hilden-Minton (1995). In fitting separate OLS regression, random effects (random slopes) are treated as fixed effects so that the level-1 residuals can be inspected without being confounded by random effects and their underlying assumptions.

For the case in which one does not use OLS for each cluster, Hilden-Minton (1995) and Verbeke and Molenberghs (2000, pp. 151–152) defined two kinds of residuals in LMM:

- Conditional residuals: Conditional residuals are discrepancies between the observed and fitted values, and they indicate how much the observed values deviate from the predicted regression line for a cluster j
  $\tilde{\boldsymbol{\epsilon}}_{C.j} = \mathbf{y}_j - E(\mathbf{y}_j|\tilde{\mathbf{b}}_j) = \mathbf{y}_j - \mathbf{X}_j\widehat{\boldsymbol{\beta}} - \mathbf{Z}_j\tilde{\mathbf{b}}_j$
  Conditional residuals are obtained by conditioning on the random effects.
- Marginal residuals: Marginal residuals measure how a specific profile deviates from the estimated overall population mean, which means conditioning on the fixed effects only
  $\tilde{\boldsymbol{\epsilon}}_{M.j} = \mathbf{y}_j - E(\mathbf{y}_j) = \mathbf{y}_j - \mathbf{X}_j\widehat{\boldsymbol{\beta}} = \mathbf{Z}_j\tilde{\mathbf{b}}_j + \tilde{\boldsymbol{\epsilon}}_j$
  The marginal residuals include both the random effects and level-1 errors.

Hilden-Minton (1995) considers a residual to be *confounded* when it depends on errors other than the one that it is supposed to predict. Following this view, predicted conditional residuals ($\tilde{\boldsymbol{\epsilon}}_{C.j}$) are confounded because the conditional residuals are co-determined by the predicted random effects ($\tilde{\mathbf{b}}_j$) which themselves may deviate from the true random effects. That is, if the predicted random effects ($\tilde{\mathbf{b}}_j$) do not follow a normal distribution, the predicted conditional residuals ($\tilde{\boldsymbol{\epsilon}}_{C.j}$) may not follow a normal distribution even when the conditional residuals ($\boldsymbol{\epsilon}_{C.j}$) follow a normal distribution.

Residuals, whether conditional or marginal, can be raw residuals or transformed residuals. Based on all these distinctions, the following six types of residuals can be considered:

- Raw residuals
  - Conditional residuals: $\tilde{\boldsymbol{\epsilon}}_{C.j}$
  - Marginal residuals: $\tilde{\boldsymbol{\epsilon}}_{M.j}$
- Standardized (or Pearson or internally studentized) residuals: Scaling is implemented by using the estimated standard deviation of the corresponding residuals ($\widehat{\sigma}$).
  - Conditional residuals: $\frac{\tilde{\boldsymbol{\epsilon}}_{C.j}}{\widehat{\sigma}}$
  - Marginal residuals: $\frac{\tilde{\boldsymbol{\epsilon}}_{M.j}}{\widehat{\sigma}}$
- Independent residuals[4]: For LMM with correlations between residuals (level-1 error in longitudinal data), orthogonalization is suggested to obtain approximately independent residuals when the within-individual variance-covariance model can describe the level-1 error adequately (e.g., Galecki & Burzykowski, 2013, pp. 265–266). We assume that $\widehat{R}_j$ (the estimated correlation of the residuals) is an adequate description of the level-1 error. An adequate description is necessary to yield independent residuals (e.g., Galecki & Burzykowski, 2013, pp. 265–266).
  - Independent conditional residuals: The independent residuals can be calculated based on the Cholesky decomposition of the estimate of the residual variance-covariance matrix $\sigma^2 R_j$ (Pinheiro & Bates 2000, p. 239). They can be calculated as $\tilde{\boldsymbol{\epsilon}}^*_{C.j} = (\widehat{\sigma}\widehat{U}_{C.j})^{-1}\tilde{\boldsymbol{\epsilon}}_{C.j}$, where $\widehat{U}'_{C.j}\widehat{U}_{C.j} = \widehat{R}_j$.
  - Independent marginal residuals: The independent residuals can be calculated based on the Cholesky decomposition of the estimate of the marginal variance-covariance matrix $\sigma^2 V_j$ (Schabenberger, 2004). They can be calculated as $\tilde{\boldsymbol{\epsilon}}^*_{M.j} = (\widehat{\sigma}\widehat{U}_{M.j})^{-1}\tilde{\boldsymbol{\epsilon}}_{M.j}$, where $\widehat{U}'_{M.j}\widehat{U}_{M.j} = \widehat{V}_j$.

Standardization does not change the shape of the distribution (which is not necessarily normal), but the mean is transformed to a value of 0 and the standard deviation is transformed to a value of 1. We recommend using standardized residuals over unstandardized residuals because they are independent of the scale of the observations and are therefore easier to interpret.

For uncorrelated level-1 error models (e.g., the cross-sectional example data set), conditional standardized residuals are the same as conditional independent residuals, and marginal standardized residuals are the same as marginal independent residuals. For longitudinal data (e.g., the second example data set), level-1 errors are likely correlated because repeated measures are from the same individuals. For correlated level-1 error models mainly in longitudinal data analysis, standardized residuals are different from independent residuals. Because standard regression diagnostics are for independent residuals, we recommend using independent residuals for the correlated level-1 error models.

[4] In the statistics literature, independent residuals are also known as transformed or normalized residuals (e.g., Galecki and Burzykowski, 2013).

Random effects as level-2 residuals

The intercepts and slopes of level-1 covariates can vary across the clusters at level 2. These random coefficients are modeled as level-2 random effects and are considered level-2 residuals (e.g., Longford, 1993, pp. 60–61). The random effects (e.g., $\tilde{\mathbf{b}}_j = [\tilde{b}_{0j}, \tilde{b}_{1j}]'$, where $\tilde{b}_{0j}$ is the predicted random intercept and $\tilde{b}_{1j}$ is the predicted random slope) reflect how much specific profiles deviate from the population average profile (or discrepancies between expected values based on level-1 fixed effects and fitted values), $\tilde{\mathbf{b}}_j = E(\mathbf{y}_j|\tilde{\mathbf{b}}_j, \widehat{\boldsymbol{\beta}}) - E(\mathbf{y}_j|\widehat{\boldsymbol{\beta}})$.

The following kinds of level-2 residuals have been used:

- Empirical Bayes (EB) residuals: There are two main prediction methods for the random effects $\mathbf{b}$ (Snijders & Berkhof, 2007). The OLS method, which treats $\mathbf{b}$ as fixed effects, and the EB method, which estimates $\mathbf{b}$ as a conditional expectation given the data ($\mathbf{y}_j$) and parameter estimates ($\widehat{\boldsymbol{\beta}}$). The relation between the level-2 predicted data ($\tilde{\mathbf{y}}_j = \frac{\sum_i \tilde{y}_{ij}}{n_j}$, where $\tilde{y}_{ij}$ is the level-1 predicted data and n_j is the number of individuals for a cluster j) and EB residuals is $\tilde{b}_j = \frac{n_j \sigma_b^2}{(n_j \sigma_b^2 + \sigma^2)} \tilde{\mathbf{y}}_j$, where σ_b^2 is a variance of a random effect and σ^2 is a residual variance. The $\tilde{b}_j$ are called shrunken residuals because the EB ($\tilde{b}_j$) is shrunken with decreasing n_j (Goldstein, 2003, p. 22).
- Standardized EB residuals: Snijders and Berkhof (2007) define the standardized level-2 residuals as $\tilde{\mathbf{b}}' Cov(\tilde{\mathbf{b}})^{-1} \tilde{\mathbf{b}}$, where $Cov(\tilde{\mathbf{b}})$ is the marginal sampling variance-covariance matrix.

What patterns in residuals indicate a good model?

For a model to be considered adequate, the following patterns should be observed in the level-specific residuals and a scatter plot of the residuals vs. fitted values:

- There should be no systematic trend in residuals.
- No more than approximately 5% of standardized residuals should have magnitudes greater than 1.96 (assuming that standardized residuals follow a standard normal distribution for a large sample size).
- The residuals should be randomly scattered around zero.
- The level-specific residuals should be normally distributed.
- The level-1 residuals (independent residuals for correlated level-1 error models in longitudinal data) should be independent of one another and independent of the fitted (predicted) values, $E(\mathbf{y}_j; \widehat{\boldsymbol{\beta}}, \tilde{\mathbf{b}}_j)$.

Review of level-specific residuals in LMM and MLM literature

We reviewed 13 texts, handbooks, edited books, and book chapters in LMM and MLM literature to survey current practices of inspecting level-specific residuals. Table 2 presents a summary.[5] As shown in Table 2, for level-1 residuals, conditional and marginal raw residuals, conditional standardized residuals, and conditional independent residuals have all been used in the LMM literature. However, in the MLM literature, only conditional raw and standardized residuals have been used. As also shown in Table 2 for random effects, unstandardized EB has been used in the LMM literature, and unstandardized and standardized EB have been used in the MLM literature.

Diagnostic measures, diagnostic plots, and statistical tests

In this section, we present diagnostic measures to select a set of random effects in model-building steps, level-specific diagnostic plots based on literature reviews, and statistical tests for patterns in the diagnostic plots.

Diagnostic measures

As a measure of difference between observed data and model predicted values (i.e., absolute fit), the root mean squared error (RMSE) is considered:

$$\text{RMSE} = \sqrt{\frac{\sum_i \sum_j (\text{data}_{ij} - \text{fitted}_{ij})^2}{N}}, \tag{10}$$

where fitted values ($E(\mathbf{y}_j; \widehat{\boldsymbol{\beta}}, \tilde{\mathbf{b}})$) are calculated based on the parameter estimates and predicted random effects from a selected model and N is the total sample size (calculated as $N = n_j J = nJ$ for a balanced design and $N = \sum_{j=1}^{J} n_j$ for an unbalanced design). The RMSE is interpreted as the standard deviation of the part of the data unexplained by a model, ($\text{data}_{ij} - \text{fitted}_{ij}$). The normalized RMSE is the proportion of the RMSE related to the range of the outcome variable. Lower values of normalized RMSE indicate better model-data fit. Because it is easier to interpret, we suggest using the normalized RMSE to find an optimal set of fixed effects and to present a model-data fit measure for the selected model. The normalized RMSE can be obtained

[5] Although we mainly use the term of MLM instead of LMM throughout this paper, we divide the literature into MLM and LMM as far as inspecting residuals is concerned because MLM literature presents practices of inspecting level-specific residuals in the context of the social and behavioral sciences whereas LMM presents them in the context of statistics.

Table 2 Uses of different residuals in LMM and MLM

Level	Residuals	LMM	MLM
Level 1	Conditional raw	OLS:Faraway (2016), Pinheiro and Bates (2000), Verbeke and Molenberghs (2000)	OLS:Snijders and Bosker (1999), Snijders and Berkhof (2007)
			EB:Longford (1993)
	Marginal raw	OLS:Galecki and Burzykowski (2013)	–
	Conditional STD	OLS:Galecki and Burzykowski (2013), Pinheiro and Bates (2000)	OLS:Finch, Bolin, and Kelley (2014), Snijders and Bosker (1999)*; Snijders and Berkhof (2007) Goldstein (2003), Hox et al. (2018), Raudenbush and Bryk (2002)
	Marginal STD	–	–
	Conditional Independent	OLS:Galecki and Burzykowski (2013)	–
	Marginal Independent	–	–
Level 2	UnSTD EB	Faraway (2016),Galecki and Burzykowski (2013)	Goldstein (2003),Hox et al. (2018), Longford (1993)
		Pinheiro and Bates (2000),Verbeke and Molenberghs (2000)	Raudenbush and Bryk (2002), Snijders and Berkhof (2007)
	STD EB		Snijders and Berkhof (2007)**

- indicates that a specific kind of residuals has not been used; STD indicates standardized residuals; *Snijders and Bosker (1999) used squared STD residuals; **Snijders and Berkhof (2007) used squared STD EB

using the `rmse(., normalized = TRUE)` function from the `performance` package in R (Lüdecke, 2020).

In addition to the RMSE for the unexplained variance across the whole data set, we propose two measures for the level-1 unexplained variability and for exploring variability across clusters. They are based on the conditional (standardized) residuals per cluster:

- The median within-cluster semi-interquartile range of the residuals (median SIQR) across clusters. The smaller the median is, the better the model captures level-1 variability in the data.

- The SIQR of the within-cluster SIQRs (SIQR(SIQR)) across clusters. The smaller the SIQR(SIQR) is, the smaller the heteroscedasticity is.
  The median SIQR represents the unexplained variability within clusters and is robust against outliers, while the SIQR(SIQR) is a measure of heteroscedasticity and is robust against outlying within-cluster unexplained variability. The median and SIQR are used because they are less sensitive to outliers than the mean and standard deviation (SD).

Fixed effects of level-1 covariates are the major explanatory factors of within-cluster variation. However, when the linearity of the effects of level-1 covariates is violated, one can further reduce the variation of the residuals by adjusting for the effects for non-linear components. Thus, a substantial reduction of the median SIQR is a useful index for the inclusion of fixed effects of level-1 covariates and for investigating any non-linearity of such effects.

Random effects of level-1 covariates (i.e., random slopes) are an explanatory factor of differences in the variance within clusters (i.e., level-1 heteroscedasticity). A steeper slope can explain why a within-cluster variance is larger. If a covariate has a steeper slope in a cluster, the covariate has a stronger effect in that cluster, which implies more variation in the outcome variable (unless compensated by a smaller variation of the covariate). Random effects of level-2 covariates can explain differences in between-cluster variance (i.e., level-2 heteroscedasticity). Therefore, a substantial reduction of the SIQR(SIQR) is a useful index for the inclusion of random effects for level-1 and level-2 heteroscedasticity.

In addition to the median SIQR of the conditional (standardized) residuals per cluster, median SIQR can be calculated based on outcome variables to explore variability in the outcome variable. RMSE is a global measure obtained from various sources of variation, and therefore may be less sensitive to one specific source of variation. However, RMSE, which is calculated based on the variance of (data

- fitted) may be oversensitive to outliers, whereas SIQR-based measures obtained with median and interquartile range are not. Therefore, we recommend the joint use of all three: RMSE, median SIQR, and SIQR(SIQR).

Diagnostic plots

Below, different kinds of diagnostic plots are discussed and organized by the model-building steps we introduced earlier. For each kind of diagnosis, we included reviews of the use of diagnostic plots in the LMM and MLM literature, summarized in Tables 3 and 4, respectively. Based on the literature reviews, commonly used diagnostic plots are selected in the model-building steps (Step 1 to Step 4). We will explain which residuals are most appropriate for each diagnostic plot when selecting random effects and which model assumptions can be checked. We use the `Math` data to illustrate diagnostic plots in *each* of the model-building steps (Step 0 to Step 4), except for a diagnostic plot of level-1 errors which is only used for the longitudinal data. For this diagnostic plot of the level-1 error, the `HD` data were used for illustration. In this section we present the different kinds of plots without making any model selection decisions for the `Math` data. As mentioned earlier, checking level-specific outliers, influential points, and normality can be implemented in Step 2 to Step 4 (which will be illustrated in the subsequent section.) However, we will wait with diagnostic plots for checking outliers, influential points, and normality until after other plots are presented from Step 0 to Step 4.

Diagnostic plots in Step 0: A preliminary descriptive analysis without any modeling In Step 0, we consider two plots: (a) a scatter plot of an outcome variable vs. a primary covariate (related to a research question) and (b) a scatter plot of an outcome variable vs. median SIQR to explore variability in the outcome variable across clusters.

A main research question for the `Math` data is the relationship between the math scores and the parents' highest level of education (`parentHED`). Thus, the scatter plot of the math scores vs. `parentHED` is considered. As shown in Fig. 1 (Step 0 (a)), there appears to be an approximately linear relationship between the math scores and `parentHED`. To explore variability across clusters (i.e., schools) in the math scores, SIQR of the conditional standardized residuals was calculated for each of the clusters, and a scatter plot of the math scores vs. SIQR was made.[6] As presented in Fig. 1 (Step 0 (b)), the SIQR varies with the math score, suggesting that the variability of math scores should be modeled as a function of the level of math scores in a cluster.

Diagnostic plots in Step 1: Random intercept for the clusters To investigate the necessity of including a random intercept, box plots of conditional raw residuals have been considered by Pinheiro and Bates (2000, p. 138). In the box plots, we suggest using conditional standardized residuals to aid interpretability of the scale because the estimated standard deviation can be different depending on the scale of the covariates. Grouping of residuals by cluster can be indicative of a random intercept because they indicate between-cluster differences and thus within-cluster dependency.

Using the `Math` data, the following two models with and without a random intercept were fit:

$$y_{ij} = \beta_0 + \epsilon_{ij} \tag{11}$$

and

$$y_{ij} = \beta_0 + b_{0j} + \epsilon_{ij}. \tag{12}$$

Standardized residuals of the null model (11) and conditional standardized residuals of the random intercept model (12) were calculated. In Fig. 1 (Step 1 (a)), the residuals for the same cluster tend to have the same sign, showing dependency within clusters. After including a random intercept, the mean of the residuals for clusters tends to be closer to 0 (presented by the horizontal line in Step 1 (b) of Fig. 1) than before. As a way to quantify dependency in outcomes due to clusters, the ICC was also calculated using the random intercept model. The ICC value of .243 confirms the dependency and validates the inclusion of a random intercept.

Diagnostic plots in Step 2: Fixed effects of level-1 covariates. When the level-1 covariates are added to the random intercept model, level-1 linearity and level-1 heteroscedasticity can be explored using diagnostic plots. We checked the linearity of level-1 covariate effects prior to investigating heteroscedasticity to meet the assumption that the expected value of the residuals was 0 (Snijders & Berkhof, 2007, pp. 148–149). If the assumption is not valid, the interpretation of heteroscedasticity may be incorrect.

Level-1 linearity A scatter plot of level-1 residuals vs. level-1 covariate is commonly used to explore the level-1 linearity. The level-1 covariate has been plotted against different kinds of the level-1 residuals in the literature: marginal unstandardized (raw) residuals (Galecki & Burzykowski, 2013), conditional unstandardized residuals (Snijders & Berkhof, 2007; Snijders & Bosker, 1999), and conditional or marginal residuals (O'Connell et al., 2016).[7] We recommend using standardized residuals to aid interpretability of

[6] Because there are 23 schools in the `Math` data, there are 23 SIQR scores. The points represent the math scores of the individual students from the cluster with a SIQR value indicated on the x-axis.

[7] O'Connell et al. (2016) did not mention whether standardized or unstandardized residuals were used.

Table 3 Uses of diagnostic plots using residuals in LMM texts

Reference	Level 1		Level 2	
	Diagnostic for	Plot (y-axis vs. x-axis)	Diagnostic for	Plot (y-axis vs. x-axis)
Faraway (2016)	Homoscedasticity	Con. raw residuals vs. fitted	Normality	QQ plot of unSTD EB
	Normality	QQ plot of con. raw residuals		
	Outlier	QQ plot of con. raw residuals		
Galecki and Burzykowski (2013)	Uncorrelated residuals	Scatter plot of con. and independent STD residuals	Normality	QQ plot of unSTD EB (intercept)
	Linearity	Scatter plot of mar. raw residuals vs. level-1 cov.	Outlier	Cluster ids vs. unSTD EB (intercept of slope)
	Normality	-QQ plot of con. STD residuals		
		-Scatter plot of con. STD residuals vs. level-1 cov. by group		
	Outlier	-Scatter plot of con. STD residuals vs. fitted	Missing level-2 cov.	Histogram of unSTD EB
		-Stripplots and box plots of con. STD by level-1 cov.		
		-Cook's D vs. individual ids		
Pinheiro and Bates (2000)	Homoscedasticity	-Con. STD residuals vs. fitted by a discrete level-1 cov.	Homoscedasticity	-UnSTD EB* of random slope vs. EB of random intercept
		-Quantiles of standard normal using con. STD residuals	of random effects	-Scatter plot of unSTD EB (intercept or slope)
		by a discrete level-1 cov.	Normality	QQ plot of unSTD EB (intercept of slope)
	Level-1 error	-Autocorrelation function of the con. STD residuals		
	Normality	QQ plot of con. raw residuals	Outlier	-QQ plot of unSTD EB (intercept of slope)
	Outlier	-Box-plots of con. raw residuals by clusters		-Scatter plot of unSTD EB (intercept or slope)
		-Con. STD residuals vs. fitted by a discrete level-1 cov.		
	Random effect	-Box-plots of con. raw residuals by clusters		
		-Con. raw residuals vs. level-1 cov.		
Verbeke and Molenberghs (2000)	Homoscedasticity	Smoothed average trend of squared OLS residuals vs. level-1 cov.	Normality	Histogram of unSTD EB
	Random effect	OLS residuals vs. level-1 cov.		

Residuals under level-1 are all level-1 residuals; con. indicates a conditional residual; mar. indicates a marginal residual; UnSTD indicates raw residuals; STD indicates standardized; cov. indicates a covariate; EB indicates empirical Bayes estimator;*Pinheiro and Bates (2000) called EB the estimated BLUPs of the random effects; The following texts in statistics were reviewed, but they did not include topics on level-specific residuals and/or diagnostic plots: Demidenko (2004) and McCulloch, Searle, and Neuhaus (2008)

Table 4 Uses of diagnostic plots using residuals in MLM texts, handbooks, edited books, and book chapters

Reference	Level 1		Level 2	
	Diagnostic for	Plot (y-axis vs. x-axis)	Diagnostic for	Plot (y-axis vs. x-axis)
Finch et al. (2014)	Normality	Histogram and QQ plot of con. STD residuals	–	
Goldstein (2003)	Homoscedasticity	Con. STD residuals vs. fitted	Normality	QQ plot of UnSTD EB (intercept or slope)
	Normality	QQ plot of con. STD residuals		
Hox et al. (2018)	Linearity	Con. STD residuals vs. fitted	Linearity	UnSTD EB (intercept or slope) vs. fitted
	Homoscedasticity	Con. STD residuals vs. fitted	Homoscedasticity	UnSTD EB (intercept or slope) vs. fitted
	Normality	Con. STD residuals vs. fitted; QQ plot of con. STD residuals	Normality	UnSTD EB (intercept or slope) vs. fitted
			Outliers	Error bar plot of UnSTD EB
Longford (1993)	Normality	Histogram and QQ plot of EB con. unSTD residuals*	Normality	QQ plot of unSTD EB (intercept or slope)
	Outlier	EB con. raw residuals* vs. level-1 cov.	Missing level-2 cov.	UnSTD EB (intercept) vs. level-2 cov.
Raudenbush and Bryk (2002)	Normality	QQ plot and stem-and-leaf of con. STD residuals	Normality	Mahalanobis plot of unSTD EB
			Homoscedasticity	UnSTD EB of random slope vs. SD of y for each cluster
				UnSTD EB of random slope vs. level-2 cov.

Table 4 (continued)

Reference	Level 1		Level 2	
	Diagnostic for	Plot (y-axis vs. x-axis)	Diagnostic for	Plot (y-axis vs. x-axis)
Singer and Willett (2003)	Linearity	Con. raw residual vs. level-1 cov.	OLS of random effects vs. level-2 cov.	
	Homoscedasticity	UnSTD residual vs. level-1 cov.	OLS random effects vs. level-2 cov.	
	Normality	QQ plot of con. raw residuals	QQ plot of unSTD OLS random effects	
	Outlier		STD OLS random effects vs. cluster id	
Snijders and Bosker (1999)	Linearity	Con. raw residuals vs. level-1 cov.	Missing level-2 cov.	UnSTD EB (intercept or slope) vs. level-2 cov.
		Mean con. raw residuals with error bars vs. level-1 cov.	Normality	QQ plot of unSTD EB
	Normality	QQ plot of con. STD residuals		
Snijders and Berkhof (2007)	Linearity	Con. raw residuals vs. level-1 cov.	Linearity	UnSTD EB vs. level-2 cov.
	Homoscedasticity	Squared con. semi-STD residuals vs. level-1 cov.	Homoscedasticity	Squared STD EB vs. level-2 cov.
			Normality	QQ plot of STD EB
O'Connell et al. (2016)	Linearity	Con. or mar. residuals vs. level-1 cov.	Normality	Histogram of unSTD EB
	Homoscedasticity	Box plots of con. or mar. residuals vs. cluster ids	Influential points	Cook's D vs. cluster ids
	Normality	QQ plot of con. or mar. residuals		
	Outlier	Studentized residuals vs. fitted		

UnSTD indicates raw residuals; STD indicates standardized; cov. indicates a covariate; EB indicates empirical Bayes estimator; SD of y indicates the standard deviation of the outcome variable; - indicates that information was not provided; Snijders and Berkhof (2007) described multivariate residuals and Studentized residuals, but they did not use them in the diagnostic plots; Kreft and de Leeuw (1998) did not describe any kinds of residuals in their book;*Longford (1993) did not mention that he used EB. However, we think it is EB because conditional distribution of the level-1 residuals is calculated based on maximum likelihood estimates.; In O'Connell et al. (2016), the authors did not specific whether they used unSTD or STD residuals

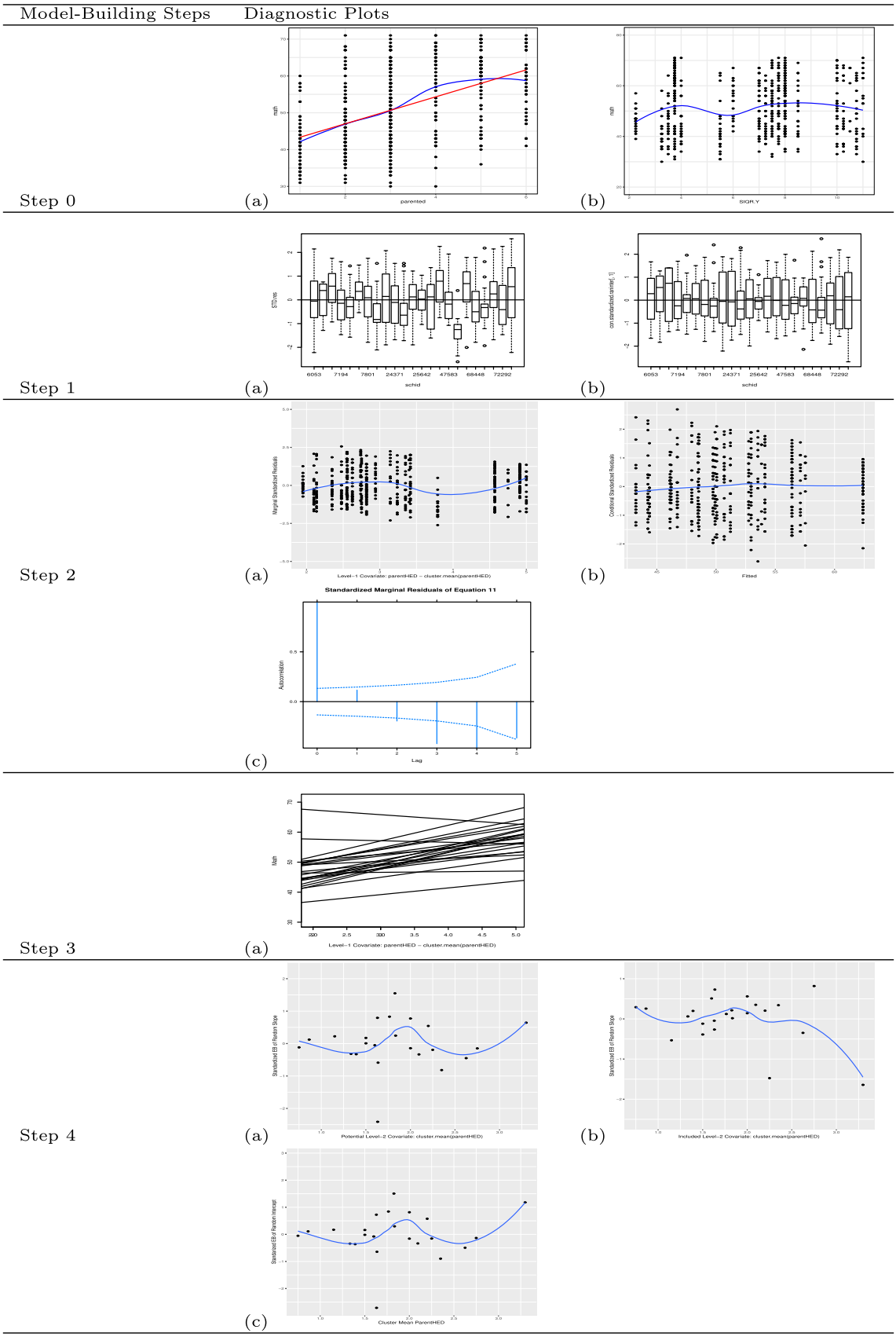

Fig. 1 Diagnostic plots for random effects selection in model-building steps

the scale. In addition, we suggest using marginal level-1 residuals instead of conditional level-1 residuals because the marginal level-1 residuals include all sources of variability (random effects and level-1 errors) for the relation between the level-1 covariate and outcomes (note that the marginal level-1 residuals are residuals obtained after only removing the fixed effects rather than after removing the random effects as well) (Santos Nobre & da Motta Singer, 2007). When the assumption of level-1 linearity holds, the average of the marginal standardized level-1 residuals is close to 0 and no systematic patterns in the residuals are found.

For `parentHED` ($x_{ij}^{(1)}$) as a continuous covariate in `Math` data, level-1 linearity was investigated based on standardized marginal residuals for a model with a random intercept and a linear level-1 covariate ($x_{ij}^{(1)} - x_{.j}^{(2)}$) effect:

$$y_{ij} = \beta_0 + \beta_1(x_{ij}^{(1)} - x_{.j}^{(2)}) + b_{0j} + \epsilon_{ij}. \quad (13)$$

Marginal standardized residuals were calculated and plotted against the level-1 covariate. As shown in Fig. 1 (Step 2 (a)), it appears that there was a slight cubic polynomial pattern (which will be tested using statistical tests in the illustration section).

Level-1 heteroscedasticity The most commonly used plot to explore level-1 heteroscedasticity is a scatter plot of residuals vs. fitted values ($E(\mathbf{y}_j; \widehat{\boldsymbol{\beta}}, \tilde{\mathbf{b}})$). Examples include conditional unstandardized (raw) residuals vs. fitted values (Faraway, 2016) and conditional standardized residuals vs. fitted values (Hox et al., 2018; Goldstein, 2003; Pinheiro & Bates, 2000). We recommend using standardized residuals for interpretability. In addition, we recommend using conditional level-1 residuals because they include only the unexplained variance and level-1 heteroscedasticity would show as unexplained variance. To check for level-1 heteroscedasticity, we explore whether the average of the conditional standardized level-1 residuals is close to 0 ($E(\tilde{\boldsymbol{\epsilon}}_{C.j}/\widehat{\sigma}) = 0$) and whether there is a constant variance across clusters ($Var(\tilde{\boldsymbol{\epsilon}}_{C.j}/\widehat{\sigma}) = \sigma^2$).

Using the `Math` data, conditional standardized residuals were calculated based on the random intercept model with a level-1 covariate (13). Fig. 1 (Step 2 (b)) presents possible level-1 heteroscedasticity. In the figure, the means of the conditional standardized residuals appear to be close to 0. However, the variance of the conditional standardized residuals looks different across the range of fitted values.

Level-1 error for longitudinal data For longitudinal data, AR and MA can be explored using an autocorrelation function of the conditional standardized residuals from a fitted model (Pinheiro & Bates, 2000, p. 242[8]). Use of the marginal residuals was advocated by Lesaffre and Verbeke (1998) to investigate a within-person variance-covariance matrix ($Var(\mathbf{y}_j) = \sigma^2 V_j = \sigma^2 \mathbf{Z}_j D \mathbf{Z}_j' + \sigma^2 R_j$ [where D is the variance-covariance matrix of random effects, $\mathbf{b}_j$]). We also recommend using marginal residuals because they include the random effects necessary to investigate whether the assumed covariance structure of the data ($Var(\mathbf{y}_j)$) does indeed fit the data. In addition, we suggest using standardized residuals for interpretability. Autocorrelations will be non-zero only in the presence of MA in the autocorrelation function (e.g., Chatfield, 2004). Fig. 1 (Step 2 (c)) presents the autocorrelation function of the marginal standardized residuals using the `HD` data.

A model can be selected among candidate models with differing $\mathbf{C}_j$ in Eq. 4 based on model selection methods, the Akaike information criterion (AIC, Akaike, 1974) and the Bayesian information criterion (BIC, Schwarz, 1978). When the correlated level-1 error model is selected, conditional or marginal independent residuals are recommended in the following steps to have approximately independent residuals, as residuals corrected for correlated level-1 errors. For example, after modeling the level-1 error regarding AR and MA, we recommend presenting an autocorrelation function of the marginal independent residuals to check whether there are noticeable patterns in the plot.

Diagnostic plot in Step 3: Random effects of level-1 covariates (i.e., random slopes) are added The most common diagnostic plot to explore random slopes is OLS regression coefficients per cluster (Hox et al., 2018; Kreft & de Leeuw, 1998; Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002; Snijders & Berkhof, 2007). In the plot, cluster-to-cluster variability in the OLS intercepts across clusters is indicative of a random intercept and cluster-to-cluster variability in the OLS slope of a level-1 covariate across clusters is indicative of a random slope. Figure 1 (Step 3 (a)) shows 23 OLS regression lines (one for each school) in `Math` data, which suggests that the slope (and intercept) differs across schools.

Diagnostic plots in Step 4: Fixed effects of level-2 covariates are added In Step 4, the potential inclusion of level-2 covariates, level-2 linearity, and level-2 heteroscedasticity can be explored. In all plots listed below, standardized EB residuals are recommended for interpretability.

Potential inclusion of level-2 covariate A scatter plot of unstandardized EB of random slope vs. a potential level-2 covariate (which is not included in the model yet) has been used to identify the functional form of the relationship between the potential level-2 covariate and the variable of interest (Raudenbush & Bryk, 2002, p. 269; Snijders

[8] Pinheiro and Bates (2000, p. 245) also presented an autocorrelation function of the conditional independent residuals to assess the adequacy of a model with the level-1 error.

& Berkhof, 2007, p. 133). Systematic patterns in the plot support the inclusion of the level-2 covariate in the model.

To illustrate this scatter plot, standardized EB of random slope was calculated based on the following model for the `Math` data with the cluster-mean-centered `parentHED` as the level-1 covariate:

$$y_{ij} = \beta_0 + \beta_1(x_{ij}^{(1)} - x_{.j}^{(2)}) + b_{0j} + b_{1j}(x_{ij}^{(1)} - x_{.j}^{(2)}) + \epsilon_{ij}. \quad (14)$$

Standardized EB of the random slope was plotted against the *potential* level-2 covariate, $x_{.j}^{(2)}$ (the cluster mean of `parentHED`), as shown in Fig. 1 (Step 4 (a)). In the figure, the standardized EBs tended to be large in the middle range of $x_{.j}^{(2)}$, which may support the inclusion of $x_{.j}^{(2)}$.

Level-2 linearity A scatter plot of unstandardized EB of random slope vs. level-2 covariates has been used to check the adequacy of the structure of those level-2 covariates (Raudenbush & Bryk, 2002, pp. 269–270). When the linear relationship between a level-2 covariate and the slope holds, the EB of the level-2 random slope should be randomly dispersed around 0 along the full range of the level-2 covariate.

To illustrate this scatter plot, standardized EB of random slope was calculated based on the following model:

$$y_{ij} = \beta_0 + \beta_1(x_{ij}^{(1)} - x_{.j}^{(2)}) + \beta_2 x_{.j}^{(2)} + b_{0j} + b_{1j}(x_{ij}^{(1)} - x_{.j}^{(2)}) + \epsilon_{ij}. \quad (15)$$

The standardized EB of the random slope was plotted against the *included* level-2 covariate, $x_{.j}^{(2)}$. As shown in Fig. 1 (Step 4 (b)), the standardized EB does not seems to be random around 0, which may indicate that the level-2 linearity assumption may not hold.

Level-2 heteroscedasticity A scatter plot of unstandardized EB of the random intercept (i.e., the level-2 residuals) vs. the level-2 covariate has been used to investigate the level-2 heteroscedasticity (Rights, 2019; Pinheiro & Bates, 2000, p. 189). In the plot, level-2 heteroscedasticity is checked by exploring whether the between-group variance depends on the level-2 covariate. Differences in the variance as a function of the level-2 covariate indicates heteroscedasticity. Standardized EB of the random intercept was calculated based on Eq. 15. In Fig. 1 (Step 4 (c)), it can be observed that variability differed depending on the level of the level-2 covariate, indicating the existence of level-2 heteroscedasticity.

For the following plots illustrating level-specific outliers, influence points, and normality, level-1 residuals and standardized EB were calculated based on a random intercept model (Equation 12) using `Math` data:

Diagnostic plots for outliers There are two categories of outlier detection methods for LMM. The first category is a set of univariate outlier detection methods such as detection based on z-scores of the outcome variable and the IQR at each level of multilevel data. The second category is a multivariate method such as Mahalanobis distance (Mahalanobis, 1936). As reviewed in Table 5, Mahalanobis distance has mostly been used at level 2. In this paper, we use the univariate outlier detection method because of its simple calculation using level-specific residuals.

Level-1 outliers The following plots can be used to detect outliers at level 1: (a) residuals vs. fitted values based on a selected model (e.g., O'Connell et al., 2016) and (b) box plot of conditional unstandardized residuals. For the plots (a) and (b), we recommend using conditional standardized residuals for uncorrelated level-1 error models and using conditional independent residuals for correlated level-1 error models (in longitudinal data). In plot (a), dispersed points in the plot can be identified as outliers. In plot (b), outliers can be detected based on the IQR. For the `Math` data, level-1 outliers were not detected as shown in Fig. 2 (outlier, Level 1) because there were no points outside of the whiskers.

Level-2 outliers The following plots can be used to detect outliers at level 2: (a) A normal probability plot of unstandardized EB for random effects (Galecki & Burzykowski, 2013, p. 344; Longford, 1993). In the normal probability plot, the data are plotted against a theoretical normal distribution. In the plot, clusters which deviate from the straight line indicate outliers. Similar to the boxplots used for level-1 outlier detection, (b) box plots of standardized EB can also be used to detect Level-2 outliers. In the box plot, outliers can be detected for clusters outside of the whiskers. Again, we recommend using standardized EB for interpretability of the plots (a) and (b). For the `Math` data, there was one cluster at the lower end which deviates extremely from the straight line in the normal probability plot as shown in Fig. 2 (outlier, Level 2 (a)). In addition, the same cluster was outside of the whiskers in the box plot, Fig. 2 (outlier, Level 2 (b)).

Diagnostic plots for influential points The Cook's distance for each observation (Cook, 1977) is often calculated to detect influential data points. Cook's distance of an observation is defined as the squared standardized difference between the estimates obtained with and without the observation in question, with large values suggesting possible influential data points. Demidenko and Stukel (2005) presented a Cook's distance for LMM.

Level-1 influential points The influence of an observation on parameter estimates is examined by leaving out each level-1 observation in turn and by recomputing parameter estimates. Because Cook's distance is in the metric of

Table 5 Diagnostic plots

Diagnostic for		Diagnostic Plots
Random Effects		-Box-plots of con. raw residuals by clusters (Pinheiro & Bates, 2000)
		-Con. raw residuals vs. level-1 cov.(Pinheiro & Bates, 2000),OLS residuals vs. level-1 cov. (Verbeke & Molenberghs, 2000)
		-OLS regression coefficients by clusters (Hox et al., 2018; Kreft & de Leeuw, 1998; Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002; Snijders & Berkhof, 2007)
Heteroskedasticity	Level 1	-Con. raw residuals vs. fitted (Faraway, 2016)
		-Con. STD residuals vs. fitted by a discrete level-1 cov. (Pinheiro & Bates, 2000)
		-Con. STD residuals vs. fitted (Hox et al., 2018; Goldstein, 2003)
		-Squared con. semi-STD residuals vs. level-1 cov. (Snijders & Berkhof, 2007)
		-Con. raw residuals vs. level-1 cov. (Singer & Willett, 2003)
		-Smoothed average trend of squared OLS residuals vs. level-1 cov. (Verbeke & Molenberghs, 2000)
		-Normal probability plot using con. STD residuals (Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002)
		-Stem-and-leaf plot of con. STD residuals (Raudenbush & Bryk, 2002)
		-Box plots of con. or mar. residuals vs. cluster ids (O'Connell et al., 2016)
	Level 2	-Scatter plot of unSTD EB (intercept) by level-2 cov. (Pinheiro & Bates, 2000)
		-OLS EB vs. level-2 cov. (Singer & Willett, 2003)
		-UnSTD EB (intercept or slope) vs. fitted (Hox et al., 2018)
Linearity	Level 1	-Scatter plot of mar. raw residuals vs. level-1 cov. (Galecki & Burzykowski, 2013);
		Scatter plot of con. or mar. residuals vs. level-1 cov. (O'Connell et al., 2016)
		-Con. raw residuals vs. level-1 cov. (Singer & Willett, 2003; Snijders & Berkhof, 2007; Snijders & Bosker, 1999)
		-Mean con. raw residuals with error bars vs. level-1 cov. (Snijders & Bosker, 1999)
		-Con. STD residuals vs. fitted (Hox et al., 2018)
	Level 2	-Scatter plot of unSTD EB of random slope vs. included level-2 covariate (Raudenbush & Bryk, 2002)
		-Scatter plot of OLS EB vs. level-2 cov.(Singer & Willett, 2003)
		-Scatter plot of unSTD EB (intercept or slope) vs. fitted (Hox et al., 2018)

Table 5 (continued)

Diagnostic for		Diagnostic Plots
Normality	Level 1	-Normal probability plot of con. raw residuals (Faraway, 2016; Pinheiro & Bates, 2000; Singer & Willett, 2003)
		-Normal probability plot of EB con. raw residuals (Longford, 1993)
		-Normal probability plot of con. STD residuals (Finch et al., 2014; Galecki & Burzykowski, 2013; Goldstein, 2003; Hox et al., 2018; Snijders & Bosker, 1999)
		-Normal probability plot of independent con. residuals (Galecki & Burzykowski, 2013)
		-Scatter plot of con. STD residuals vs. level-1 cov. by group (Galecki & Burzykowski, 2013)
		-Histogram of con. STD residuals (Finch et al., 2014)
		-Histogram of con. STD or mar. residuals (O'Connell et al., 2016)
	Level 2	-Normal probability plot of unSTD EB (Faraway, 2016; Galecki & Burzykowski, 2013; Goldstein, 2003; Longford, 1993; Pinheiro & Bates, 2000; Snijders & Bosker, 1999)
		-Normal probability plot of unSTD OLS (Singer & Willett, 2003)
		-Normal probability plot of STD EB (Snijders & Berkhof, 2007)
		-Histogram of unSTD EB (Verbeke & Molenberghs, 2000)
		-Mahalanobis plot of unSTD EB (Raudenbush & Bryk, 2002)
Outlier	Level 1	-Normal probability plot of con. raw residuals (Faraway, 2016)
		-Stripplots and box plots of con. STD by level-1 cov. (Galecki & Burzykowski, 2013)
		-Cook's D vs. individual ids (Galecki & Burzykowski, 2013)
		-Studentized vs. fitted (O'Connell et al., 2016)
	Level 2	-Normal probability plot of unSTD EB (intercept of slope) (Galecki & Burzykowski, 2013; Longford, 1993)
		-Scatter plot of unSTD OLS vs. cluster id (Singer & Willett, 2003)
		-Error bar plot of unSTD EB (Hox et al., 2018)
		-Mahalanobis plot of unSTD EB (Raudenbush & Bryk, 2002)
		-Cook's D vs. cluster ids (O'Connell et al., 2016)

STD indicates standardized; cov. indicates a covariate; EB indicates empirical Bayes estimator; con. indicates a conditional residual; mar. indicates a marginal residual

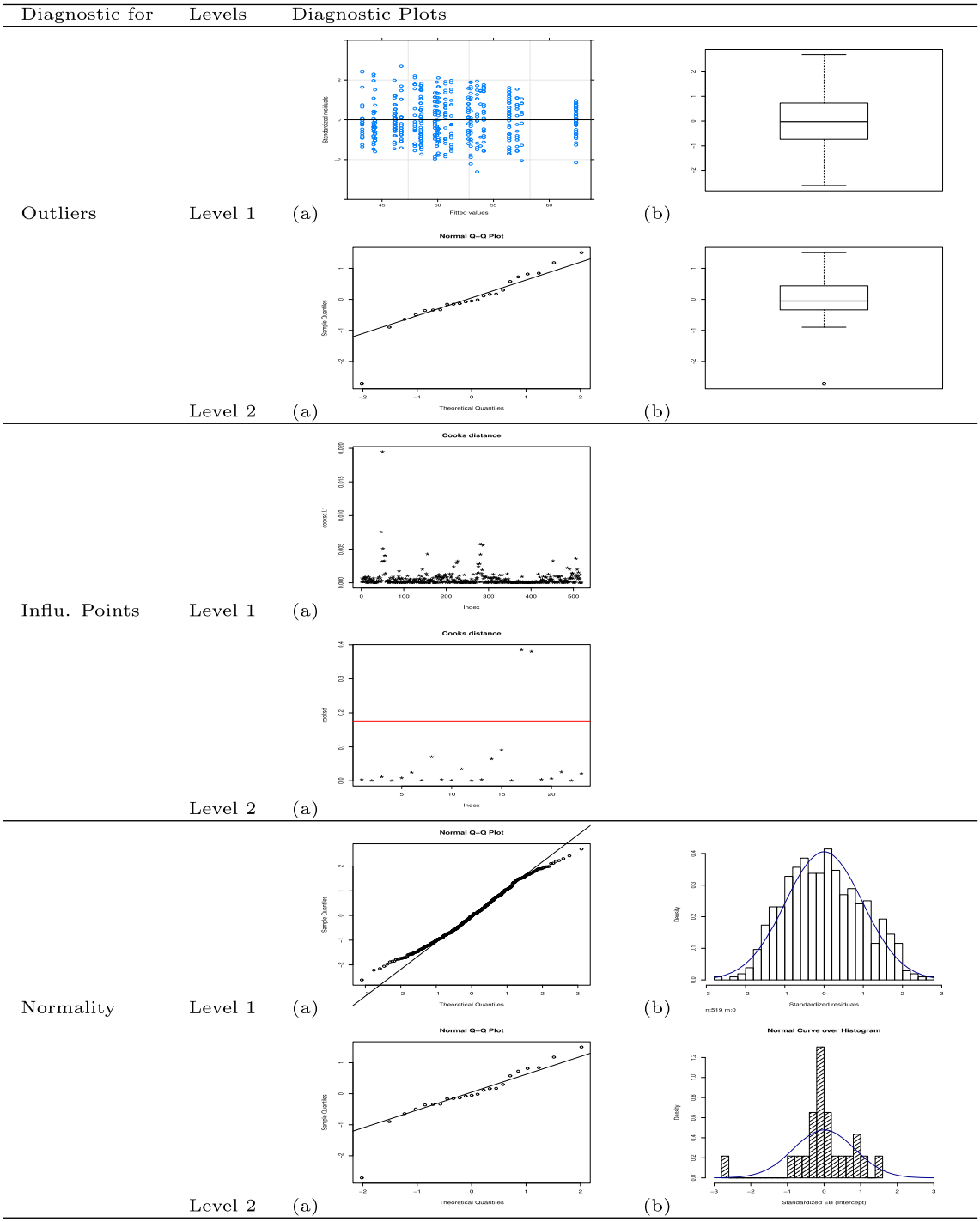

Fig. 2 Diagnostic plots for outliers, influential points, and normality checks

an $F(p, N - p)$ distribution (where p is the number of regression parameters excluding the intercept and N is the number of observations), the median point, $F_{0.5}(p, N - p)$, is used as a cut-off value to detect influential points (e.g., Bollen & Jackman, 1990). As another cut-off value, level-1 observations can be considered as highly influential points when the level-1 Cook's distance is larger than 1 for a large sample size (Cook & Weisberg, 1982). In this study, we use the cut-off value of 1 for the level-1 Cook's distance because the number of observations is often large in multilevel data. For the `Math` data, there were no influential points at level 1 because there were no points with a Cook's distance larger than the cut-off value of 1 (see Fig. 2 [influ. points, Level 1 (a)]).

Level-2 influential points At level 2, the influence of a cluster on parameter estimates is examined by leaving out each cluster in turn and by recomputing parameter estimates. To our knowledge, a theoretical justification of a cut-off value has not been proposed for the level-2 Cook's distance. In practice, a cut-off value of 4 divided by the number of clusters has been used to identify level-1 influence points if the sample size is not very large (e.g., 4059 individuals in Loy & Hofmann, 2014). We also use the cut-off value of 4 divided by the number of clusters. For the `Math` data, there were two influential points at level 2 (clusters), based on a cut-off value of .17 (= 4/23) (see Fig. 2 [influ. points, Level 2 (a)]).

Diagnostic plots for normality

Level-1 normality The following approaches have been used to check normality of level-1 residuals: (a) normal probability plots for various types of residuals such as conditional unstandardized residuals (Faraway, 2016; Pinheiro & Bates, 2000), unstandardized EB conditional residuals (Longford, 1993), conditional standardized residuals (Finch et al., 2014; Galecki & Burzykowski, 2013; Goldstein, 2003; Hox et al., 2018; Snijders & Bosker, 1999), and conditional independent residuals (Galecki & Burzykowski, 2013); (b) a scatter plot of conditional standardized residuals vs. level-1 covariate by group with a limited number of categories (Galecki & Burzykowski, 2013, p. 231); and (c) histograms overlaid with a curve based on conditional standardized residuals (Finch et al., 2014), conditional unstandardized residuals (Longford, 1993) and conditional or marginal standardized residuals (O'Connell et al., 2016). The normal probability plot (plot (a)) is created with an *independent* residual assumption. Thus, for level-1 correlated errors, we recommend using conditional independent level-1 residuals to obtain approximately independent residuals for the normal probability plot. However, it is not necessary to use the independent residuals for plots (b) and (c) because of their descriptive purpose. In the plots (b) and (c), standardized residuals are recommended for interpretability. For level-1 uncorrelated errors, standardized residuals are the same as independent residuals. Conditional residuals can be used in all three kinds of plots, and they are preferred over marginal residuals because in the conditional residuals both fixed and random effects of the model are accounted for. In the plot (a), straight lines indicate normality. In the plot (b), the normality assumption seems reasonable when there are no conditional standardized residuals (presented on the y-axis) smaller than the 1st percentile of the standard normal distribution (-2.33) or larger than the 99th percentile of the standard normal distribution (2.33) for a level-1 covariate (on the x-axis) by groups. In the plot (c), normality can be assumed when the shape of the distribution in the histogram looks like the overlaid normal (or bell-shape) distribution.

The plots (a) and (c) are illustrated using the `Math` data. The plot (b) is not applicable to the data because there are too many levels of the level-1 covariate. As shown in Fig. 2 (normality, Level 1), small deviations from normality were observed in the middle and toward the ends of the distributions of the conditional standardized residuals.

Level-2 normality The following plots have been used for checking normality of random effects: (a) normal probability plots of unstandardized EB (Faraway, 2016; Galecki & Burzykowski, 2013; Goldstein, 2003; Longford, 1993; Pinheiro & Bates, 2000; Snijders & Bosker, 1999) or standardized EB residuals (Snijders & Berkhof, 2007) and (b) histograms of unstandardized EB residuals (O'Connell et al., 2016; Verbeke & Molenberghs, 2000). Mainly unstandardized EB has been used in the plots, except in one case where a normal probability plot of standardized EB is used in Snijders and Berkhof (2007). We recommend using standardized EB for interpretability. For the `Math` data, deviations from normality were observed at the ends of distributions of the standardized EB in both plots, as shown in Fig. 2 (normality, Level 2).

Statistical tests

Interpreting patterns in diagnostic plots is subjective in nature. Thus, in this subsection, we provide statistical tests for a more objective interpretation.

Testing for randomness in residuals Bartels (1982) proposed a rank version of the von Neumann's (1941) ratio test to test the null hypothesis that there is randomness in data against the alternative hypothesis that there is trend in the data. Bartels ratio test statistic is defined as

$$T = \frac{\sum_{i=1}^{I-1}(r[i] - r[i+1])^2}{\sum_{i=1}^{I}(r[i] - \bar{r})^2}, \tag{16}$$

where i is an index for level-1 observations ($i = 1, \ldots, I$), $r[1], \ldots, r[I]$ are ranks of the level-1 residuals $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_I$ in the diagnostic plots, and $\bar{r}$ is the average rank based on the number of residuals, $(I+1)/2$. The Bartels ratio test can be used to test whether there is trend in the residuals of a selected model.

Testing for autocorrelations in residuals After modeling level-1 correlated errors in longitudinal data, the Durbin-Watson test (Durbin & Watson, 1950) can be used to test the null hypothesis of independent level-1 residuals against

first-order serially correlated errors. The Durbin-Watson test statistic is defined as

$$DW = \frac{\sum_{i=2}^{I}(\tilde{\epsilon}_i - \tilde{\epsilon}_{i-1})^2}{\sum_{i=1}^{I}\tilde{\epsilon}_i^2}, \tag{17}$$

where $\tilde{\epsilon}_i$ is a calculated residual based on data, parameter estimates, and predicted random effects.

Testing for homogeneity of variance across groups There are various tests to test the homogeneity of variance in residuals across groups defined by one or more factors as in an analysis of variance (ANOVA) (see Wang et al., 2016 for reviews). In this study, Levene's test (Levene et al., 1960) was selected as a commonly used test in social and behavioral sciences (e.g., SPSS software, which uses Levene's test as the default). Wang et al. (2016) showed via simulation studies that the Levene's test maintained adequate Type I error rates and power in various conditions. When the number of levels for the level-1 and level-2 covariates is small, Levene's test can be used to test level-1 and level-2 homogeneity, respectively. In addition, Levene's test can be used to test whether the variance of residuals differs across clusters to confirm the necessity of including a random intercept.

Testing for smooth functions in the diagnostic plots Smooth functions can be plotted to observe patterns in the diagnostic plots, such as plotting level-1 (marginal standardized) residuals vs. level-1 covariate for testing level-1 linearity (Fig. 1 [Step 2 (a)]), level-1 (conditional standardized) residuals vs. fitted values for testing level-1 heteroscedasticity (Fig. 1 [Step 2 (b)]), standardized EB of the random intercept vs. level-2 covariate for testing level-2 linearity (Fig. 1 [Step 4 (b)]), and standardized EB of the random slope vs. level-2 covariate for testing level-2 heteroscedasticity (Fig. 1 [Step 4 (c)]).

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

$$f_h(x) = \sum_{k=1}^{K}\gamma_{hk}b_{hk}(x_h), \tag{18}$$

where k is an index for a basis function ($k = 1, \ldots, K$), x_h is a covariate for a smooth function h, γ_{hk} is a basis coefficient, and $b_{hk}(x)$ is the kth basis function for smooth function h. Because the $f_h(x)$ can be confounded with the intercept, a model is estimated with an identification constraint that the sum of the function f_h over the observed covariate values is 0 (i.e., $\sum_v f_h(x_{hv}) = 0$ for each h with v as a subscript for observations). For the univariate smooth function ($f_h(x)$), a cubic regression spline (CRS, Wood, 2017) and a thin plate regression spline (TPRS; Wood, 2017, 5.5.1) are commonly used splines that can be implemented using the mgcv R package (Wood, 2019).

To test whether a smooth function $f_h(x)$ is distinguishable from zero, the following null hypothesis can be tested: $H_0 : f_h(x) = 0$ for all x in the range of interest. A test statistic for $\mathbf{f}_h$ is

$$T_r = \widehat{\mathbf{f}}_h' \mathbf{V}_{f_h}^{-} \widehat{\mathbf{f}}_h, \tag{19}$$

where r is the rounded effective degrees of freedom (edf) of $\mathbf{f}_h$ and $\mathbf{V}_{f_h}^{-}$ is a rank r pseudo-inverse of $\mathbf{V}_{f_h}$ calculated as $XV_{\boldsymbol{\gamma}}X'$ (where X are basis functions and $V_{\boldsymbol{\gamma}}$ is the variance-covariance matrix for $\widehat{\boldsymbol{\gamma}}$). Each $\widehat{\mathbf{f}}_h$ is approximately multivariate normal,

$$\widehat{\mathbf{f}}_h \sim MVN(\mathbf{f}_h, \mathbf{V}_{f_h}), \tag{20}$$

where $\mathbf{f}_h$ is the vector of $f_h(x)$ evaluated at the observed covariate values. Under H_0, the test statistic T_r follows a Chi-square distribution ($T_r \sim \chi_r^2$) with $r = edf$ (Wood, 2012). When H_0 is rejected, one can conclude that there is a pattern (linear or nonlinear) in the data or residuals. The edf can be referred to when investigating whether the relation between a covariate and the outcome (e.g., residuals) is linear or nonlinear (Wood, 2017). The higher the edf, the wigglier the estimated smooth function is. An edf of 1 indicates a linear effect of a covariate on the outcome, an edf of 2 indicates an approximately quadratic effect of a covariate on the outcome, and an edf of 3 indicates an approximately cubic effect of a covariate on the outcome. Smooth functions have confidence intervals, which are obtained by taking the quantiles from the posterior distribution of the $\mathbf{f}_h$ (Marra & Wood, 2012).

Normality The normality assumption of level-1 residuals and univariate EB (in our case, EB of the random intercept) can be tested using the Shapiro–Wilk normality test. When a selected model includes more than one random effect (e.g., random intercept and random slope), the multivariate normality of the random effects can be tested. A multivariate normality test such as Mardia's test can be considered to test the multivariate normality assumption of the random effects (e.g., see Farrell, Salibian-Barrera, & Naczk, 2007; von Eye & Bogat, 2004, for the details of the test).

Illustration

In this section, uses of diagnostic plots based on level-specific residuals, diagnostic measures, and statistical tests of the patterns in the diagnostic plots are illustrated in a model-building strategy using cross-sectional and longitudinal empirical data sets. R code is provided for each step in Appendix A.

Example 1: Two-level cross-sectional data (`Math` data)

Steps 0 and 1 (A Preliminary Descriptive Analysis and Random Intercepts for the Clusters) and their results are discussed and reported earlier (see the Diagnostic Plots subsection). Below, Steps 2–5 are illustrated. Table 6 presents a summary of analyses and results.

Step 2. Fixed effects of the level-1 covariate of interest

As mentioned earlier, a goal of analysis using the `math` data set is to predict math scores from parents' highest level of education (`parentHED`). In Step 2, the fixed effect of the level-1 covariate of interest, the cluster-mean-centered `parentHED` ($x_{ij}^{(1)} - x_{\cdot j}^{(2)}$), is added to create Model 1:

$$y_{ij} = \beta_0 + b_{0j} + \beta_1(x_{ij}^{(1)} - x_{\cdot j}^{(2)}) + \epsilon_{ij}, \tag{21}$$

where β_1 is the fixed effect of the cluster-mean-centered `parentHED`. The addition of the fixed effect of the level-1 covariate lowered the median SIQR from 0.790 in the Null Model Random to 0.679 in Model 1. This result indicates that Model 1 better captured the level-1 variability in the data than Null Model Random.

Level-1 linearity Cluster-mean-centered `parentHED` was plotted against the marginal standardized residuals obtained from Model 1 to examine whether the relationship between cluster-mean-centered `parentHED` and math scores is strictly linear. As shown in Fig. 3 (Step 2 (a)), there is a nonlinear relationship between the cluster-mean-centered `parentHED` and the marginal standardized residuals at the extreme values of the cluster-mean-centered `parentHED`, indicating that ($x_{ij}^{(1)} - x_{\cdot j}^{(2)}$) may have a non-linear (square and/or cubic) relationship with math scores. To test these higher-degree effects of ($x_{ij}^{(1)} - x_{\cdot j}^{(2)}$) on math scores, an alternative version of Model 1 including the square and cubic effects of ($x_{ij}^{(1)} - x_{\cdot j}^{(2)}$) called Model 1a, was tested:

$$\begin{aligned} y_{ij} = {} & \beta_{0j} + b_{0j} + \beta_1(x_{ij}^{(1)} - x_{\cdot j}^{(2)}) + \beta_2(x_{ij}^{(1)} - x_{\cdot j}^{(2)})^2 \\ & + \beta_3(x_{ij}^{(1)} - x_{\cdot j}^{(2)})^3 + \epsilon_{ij}, \end{aligned} \tag{22}$$

where β_1, β_2, and β_3 are the linear, square, and cubic (respectively) fixed effects of the cluster-centered `parentHED`. Neither of the higher order (square and cubic) terms in Model 1a were found to be statistically significant, having p values of .2397 and .4529, respectively. In addition, a smooth curve fitted to predict marginal standardized residuals of Model 1 as a function of cluster-centered `parentHED` (using the `mgcv` package in `R`) showed that a smooth curve is not needed ($F = 1.627$, $edf = 10.6$, p value$= .077$). Based on these results, linearity was assumed, and Model 1 was used instead of Model 1a, with only the linear term for ($x_{ij}^{(1)} - x_{\cdot j}^{(2)}$) included.

Level-1 heteroscedasticity The fitted values of Model 1 were plotted against the conditional standardized residuals to explore the level-1 heteroscedasticity, as presented in Fig. 3 (Step 2 (b)). The conditional standardized residuals were distributed around 0 along the continuum of fitted values, meaning that homoscedasticity can be assumed. In addition, a Levene's test showed that the conditional standardized residuals were not significantly heteroscedastic (p value$= .098$).

Level-2 outliers If any level-1 or level-2 units are detected during the model building process (up to Step 4) as being both outlying and influential, these level-1 and/or level-2 units will be removed from the data, as they are expected to influence the resulting parameter estimates in a way that disagrees with the rest of the data (Hilden-Minton, 1995; Langford & Lewis, 1998). To detect level-2 outliers, a normal Q-Q plot of the standardized EB of the intercept for Model 1 was plotted against a theoretical normal distribution in Fig. 3 (Step 2 (c)). The standardized EB of the intercept were largely normal, with no standardized EB falling outside of the 95% confidence bands. The standardized EB of the intercept for all level-2 units ranged from -1.665 to 2.253. Based on these results, no level-2 units were considered to be outliers.

Level-2 influential points There were two level-2 influential points, having Cook's distances exceeding the cutoff of $0.1739 = 4/23$ for a sample size of 23 schools, as shown in Fig. 3 (Step 2 (d)).

Level-1 outliers To detect level-1 outliers, the fitted values from Model 1 were plotted against the conditional standardized residuals (see Fig. 3 [Step 2 (e)]). No outliers were detected as having unusually high conditional standardized residuals. The largest observed conditional standardized residual was 2.751, which although large is not unexpected given the large number of level-1 units (519).

Level-1 influential points No level-1 influential points were detected, as no point had a Cook's distance greater than the cut-off value of 1 in Fig. 3 (Step 2 (f)). The highest Cook's distance detected was 0.0195.

Level-1 normality A normal Q-Q plot as presented in Fig. 3 (Step 2 (g)) was generated to examine whether the conditional standardized residuals of Model 1 were normally distributed. The Q-Q plot shows that the conditional standardized residuals were mostly normal, with

Table 6 Summary of analyses and results for `Math` data

Steps	Diagnostic Measures, Plots, and Tests	Results
Step 0	(a) A plot of the math scores vs. `parentHED` (Fig. 2 [Step 0 (a)])	Approximately linear relationship between the math scores and `parentHED`
	(b) A plot of the math scores vs. SIQR (Fig. 2 [Step 0 (b)])	Different math scores depending on the levels of SIQR across clusters
Step 1	(a) Box plots of con. STD residuals for Null Model Fixed (Eq. 11) for each school (Fig. 2 [Step 1 (a)])	The residuals distributed around 0 when adding a random intercept
	(b) Box plots of con. STD residuals for Null Model Random (Eq. 12) for each school (Fig. 2 [Step 1 (b)])	⇒Null Model Random is selected.
Step 2	SIQR comparisons between Null Model Random and Model 1 (Eq. 21)	Reduced SIQR
Level-1 Linearity	(a) A plot of the mar. STD for Model 1 (Eq. 21) vs. cluster-mean-centered `parentHED` (Fig. 3 [Step 2 (a)])	A smooth curve not needed and
	with a test of a smooth function	insignificance of the square and cubic effects
	(b) Testing the square and cubic effects using Model 1a (Eq. 22) (Model 1 + the square and cubic effects)	⇒Model 1 is retained.
Level-1 Hetero.	(a) A plot of con. STD residuals for Model 1 vs. Fitted values for Model 1 (Fig. 3 [Step 2 (b)])	The residuals distributed around 0 and insignificance of the Levene's test
	(b) Levene's test using the con. STD residuals for Model 1	Homo. assumed ⇒ Model 1 is retained.
Level-2 Outliers	Normal Q-Q plot of the STD EB of the intercept for Model 1 (Fig. 3 [Step 2 (c)])	No level-2 outliers ⇒ Model 1 is retained.
Level-2 Influ.	Cook's distance vs. School IDs (Fig. 3 [Step 2 (d)])	Two level-2 outliers detected, but not deleted yet ⇒ Model 1 is retained.
Level-1 Outliers	Con. STD residuals for Model 1 vs. Fitted values for Model 1 (Fig. 3 [Step 2 (e)])	No level-1 outliers detected ⇒ Model 1 is retained.
Level-1 Influ.	Cook's distance vs. Student IDs (Fig. 3 [Step 2 (f)])	No level-1 influ. detected ⇒ Model 1 is retained.
Level-1 Normality	(a) Normal Q-Q plot of the con. STD residuals for Model 1 (Fig. 3 [Step 2 (g)])	Significance of the Shapiro–Wilk test, but non-normality was not extreme
	(b) Shapiro–Wilk test of the con. STD residuals	observed in the normal Q-Q plot and histogram
	(c) Histogram of the con. STD residuals (Fig. 3 [Step 2 (h)])	Level-1 normality assumed ⇒ Model 1 is retained.
Level-2 Normality	Normal Q-Q plot of the STD EB of the intercept for Model 1 (Fig. 3 [Step 2 (c)])	Level-2 normality assumed ⇒ Model 1 is retained.

Table 6 (continued)

Steps	Diagnostic Measures, Plots, and Tests	Results
Step 3	(a) SIQR and SIQR(SIQR) for Model 2 (Eq. 23) (Model 1 + random slope)	More level-1 variability captured by the random slope and
	(b) OLS regression lines by schools (Fig. 3 [Step 3 (a)])	Variability in intercepts and slopes across schools
	(c) The same analyses for Model 2 as Step 2	Level-1 homo., no-severe outliers and influ., and normality assumed
		⇒ Model 2 is selected.
Step 4		
Inclusion of Level-2 Cov.	A plot of STD EB of random slope vs. cluster means `parentHED`	Cluster means `parentHED` needed
	for Model 2 (Fig. 3 [Step 4 (a)]) and Model 3 (Eq. 24; Fig. 3 [Step 4 (b)])	⇒ Model 3 is selected.
Level-2 Linearity	(a) A plot of the STD EB of the random slope for Model 3 vs. cluster means `parentHED` (Fig. 3 [Step 4 (c)])	A smooth curve not needed,
	with a smooth function	
	(b) Testing the square and cubic effects using Model 3a (Eq. 25) (Model 3 + the square and cubic effects)	Significant square and cubic effects
	(c) A plot of the STD EB of the random slope for Model 3a vs. cluster means `parentHED` (Fig. 3 [Step 4 (d)])	Nonlinear patters likely from the small number of clusters
		⇒ Model 3 is retained.
Level-2 Hetero.	(a) A plot of the STD EB of the random intercept for Model 3 vs. cluster means `parentHED` (Fig. 3 [Step 4 (e)])	The residuals distributed around 0 and insignificance of the Levene's test
	(b) Levene's test using the STD EB of the random intercept for Model 3	Homo. assumed ⇒ Model 3 is retained.
Level-2 Outliers	Normal probability plot and box plots of the STD EBs of random effects for Model 3	A few level-2 outlying cluster detected
Level-2 Influ.	Cook's distance vs. School IDs	One influential cluster detected and deleted
Level-2 Normality	Normal Q-Q plot and histogram of the STD EB of random effects for Model 3	Level-1 and level-2 normality assumed
Level-1	The same analyses for Model 3 as Step 2	No level-1 outliers and influ. detected
Step 5		
Control Cov.	Testing level-1 and level-2 effects of control covariates for Model 4	Significant fixed effects of `homework` and `white` in Model 4c
	A plot of the con. STD residuals vs. fitted values for Model 4c (Fig. 3 [Step 5 (b)])	No noticeable systematic patterns
	with a Bartels ratio test	⇒ Model 4c is selected.

"Hetero." indicates heteroscedasticity; "Homo" indicates homoscedasticity; "Influ." indicates influential points; "Cov." indicates a covariate; con. STD indicates conditional standardized; "mar. STD" indicates marginal standardized; "STD EB" indicates the standardized empirical Bayes residuals

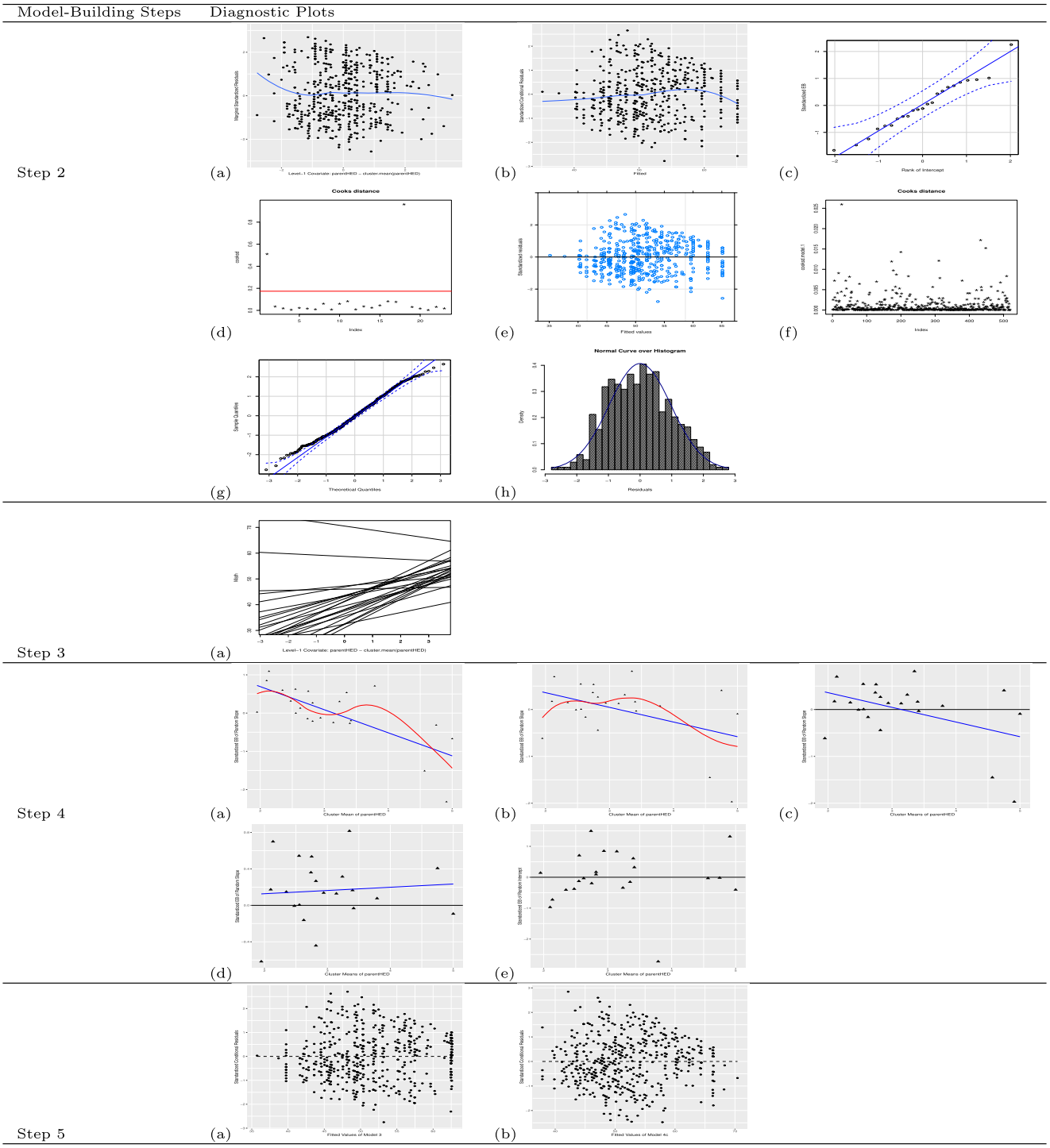

Fig. 3 Diagnostic plots for random effects selection in the two-level cross-sectional data (`Math` data)

some deviations from normality in the lower extreme (with conditional standardized residuals falling slightly outside the 95% confidence bands). A Shapiro–Wilk test indicated that conditional standardized residuals were significantly non-normal (p value = .0022), which as shown in the normal Q-Q plot above is due to deviances from normality in the extreme observations. However, a histogram of conditional standardized residuals of Model 1 overlaid a normal curve shows that this deviance from normality is not large (see Fig. 3 [Step 2 (h)]). To conclude, level-1 normality

was assumed because, although the p value was small, the deviance from normality was too small to give up on the normality assumption.

Level-2 normality A normal Q-Q plot was generated to examine whether the standardized EB of the intercept of Model 1 were normally distributed, as presented in Fig. 3 (Step 2 (c)).

Step 3. Random effects of the level-1 covariate of interest

In this step the random effect (i.e., random slope) of the level-1 covariate $(x_{ij}^{(1)} - x_{\cdot j}^{(2)})$, the cluster-mean-centered `parentHED`, is added to the Null Model Random (Equation 12) to create Model 2:

$$y_{ij} = \beta_0 + b_{0j} + \beta_1 (x_{ij}^{(1)} - x_{\cdot j}^{(2)}) + b_{1j} (x_{ij}^{(1)} - x_{\cdot j}^{(2)}) + \epsilon_{ij}, \quad (23)$$

where b_{1j} is the random slope of the cluster-mean-centered `parentHED`. The addition of the random effect of the level-1 covariate lowered the median SIQR from 0.679 in Model 1 to 0.627 in Model 2, but it increased the SIQR(SIQR) from 0.121 in Model 1 to 0.143 in Model 2. These results indicate that Model 2 better captured the level-1 variability in the data (by having a smaller median SIQR), but was slightly more heteroscedastic (by having a larger SIQR(SIQR)). The small difference in SIQR(SIQR) is likely influenced by the small number of level-2 units, as will be discussed in Step 5.

To show the variability in the effect of $x_{ij}^{(1)} - x_{\cdot j}^{(2)}$ across schools, the ordinary least squares (OLS) regression line predicting math scores with cluster-mean-centered `parentHED` was plotted for each school, as shown in Fig. 3 (Step 3 (a)). Variability in intercepts across schools in this plot is indicative of the need for a random intercept (b_{0j}), whereas variability in slopes across schools in this plot is indicative of the need for the random slope (b_{1j}).

Level-1 heteroscedasticity and level-specific outliers, influential points, and normality As in Step 2, level-1 heteroscedasticity, level-1 and level-2 outliers and influential points, and level-1 normality were checked by examining the conditional standardized residuals and standardized EB of the intercept of Model 2. In addition, level-2 normality and multivariate normality were checked by examining the standardized EB of the intercept and slope of Model 2.

The conditional standardized residuals of Model 2 were distributed around 0 along the continuum of fitted values, indicative of level-1 homoscedasticity. This was further supported by a Levene's test showing that the conditional standardized residuals were not significantly heteroscedastic (p value = .139). One level-1 outlier was detected with a conditional standardized residual of 2.797, though no level-1 units (including this outlier) were influential, with a maximum Cook's distance of 0.0229. One influential level-2 unit was detected with a Cook's distance of 0.237 (> 0.174), though no level-2 outliers were detected, with all standardized EB of the intercept ranging from -1.615 to 2.352. A normal Q-Q plot of the conditional standardized residuals of Model 2 (plotted to evaluate level-1 normality) resulted in a pattern similar to Model 1. As a result, level-1 normality is assumed for Model 2.

Normal Q-Q plots were generated to examine whether the standardized EB of the intercept and slope were normally distributed for Model 2 (the figure is not shown). The standardized EB of the intercept were normally distributed, with all standardized EB falling within the 95% confidence bands. The standardized EB of the slope were mostly normally distributed, with two level-2 units falling outside the 95% confidence bands. To further examine level-2 normality, histograms of the standardized EB of the intercept and slope were plotted (the figure is not shown). In both cases, level-2 normality was questionable to investigate, as any potential non-normality could be the result of the small number of level-2 units (23). Because there were no drastic violations of level-2 normality (and no exceptional outliers observed), level-2 normality was assumed for Model 2.

Step 4. Fixed effects of a level-2 covariate of interest

In this step the fixed effect of the level-2 covariate $x_{\cdot j}^{(2)}$, the cluster mean of `parentHED`, was added to create Model 3:

$$y_{ij} = \beta_0 + b_{0j} + \beta_1 (x_{ij}^{(1)} - x_{\cdot j}^{(2)}) + b_{1j} (x_{ij}^{(1)} - x_{\cdot j}^{(2)}) + \beta_2 x_{\cdot j}^{(2)} + \epsilon_{ij}, \quad (24)$$

where β_2 is the fixed effect of the cluster mean of `parentHED`.

Potential inclusion of level-2 covariate To explore whether the `parentHED` cluster means should be included in the model, the level-2 covariate (which was not previously included) was plotted vs. the standardized EB of the random slope for Model 2 (Equation 23, which does not include the level-2 covariate), as presented in Fig. 3 (Step 3 (a)). Standardized EB of the random slope had an identifiable pattern (a negative linear trend) across the range of `parentHED` cluster means, justifying the inclusion of the cluster mean of `parentHED` in the model. After including the cluster mean of `parentHED` in the model, the standardized EB of the random slope for Model 3 was plotted (see Fig. 3 [Step 4 (b)]). Although there was still a negative linear trend in the standardized EB, the slope of

this negative trend was reduced from −0.6035 (in Model 2) to −0.3132 (in Model 3).

Level-2 linearity To examine whether the relationship between the included `parentHED` cluster means and math scores is strictly linear, a scatter plot of the standardized EB of the random slope for Model 3 vs. the level-2 covariate was generated (see Fig. 3 [Step 4 (c)]). The standardized EB were not randomly dispersed around 0 along the full range of the level-2 covariate, as would be expected if the relationship between `parentHED` cluster means and math scores was nonlinear. Instead, there was a significantly negative linear trend (slope = −0.313, p value= .0337). In addition, a third-degree smooth curve fitted to predict standardized EB as a function of `parentHED` cluster means (using the `mgcv` package in `R`) was found to be significantly nonlinear ($F = 3.545, edf = 1.720, p$ value = .0325), which suggests that there is a nonlinear relationship that needs to be included in the model. As Fig. 3 (Step 4 (c)) shows, there is a potentially nonlinear relationship between `parentHED` cluster means and math scores in Model 3, indicating that `parentHED` cluster means may have a nonlinear (square and/or cubic) relationship with math scores. To test higher-degree effects of `parentHED` cluster means on math scores, an alternative version of Model 3 including the square and cubic effects of `parentHED` cluster means ($x_{\cdot j}^{(2)}$) on math scores, called Model 3a, was tested:

$$\begin{aligned} y_{ij} = {} & \beta_0 + b_{0j} + \beta_1 (x_{ij}^{(1)} - x_{\cdot j}^{(2)}) + b_{1j} (x_{ij}^{(1)} - x_{\cdot j}^{(2)}) \\ & + \beta_2 x_{\cdot j}^{(2)} + \beta_3 (x_{\cdot j}^{(2)})^2 + \beta_4 (x_{\cdot j}^{(2)})^3 + \epsilon_{ij}, \end{aligned} \tag{25}$$

where β_2, β_3, and β_4 are the linear, square, and cubic (respectively) fixed effects of `parentHED` cluster means on math scores. Both of the higher-order (square and cubic) terms in Model 3a were statistically significant (p value= .0255 and p value= .0282, respectively). A scatter plot of the level-2 covariate vs. the standardized EB of the random slope for Model 3a was generated to compare with Model 3 (the figure is not shown.) The pattern observed for Model 3a was almost identical to that observed for Model 3 without the inclusion of the higher-order terms of `parentHED` cluster means. These results indicate that the original pattern observed in the plot of the level-2 covariate vs. the standardized EB of the random slope for Model 3 was not caused by a strong nonlinear relationship between math scores and `parentHED` cluster means. Because the number of level-2 units is small (23), it is possible that the results stem from a few clusters. The negative slope and the nonlinearity may have been caused by two clusters with more negative standardized EB (−1.970 and −1.451) than the rest of the clusters. To examine this possibility, the level-2 covariate vs. the standardized EB of the random slope for Model 3 was plotted without these two clusters, as presented in Fig. 3 (Step 4 (d)). The standardized EB without these two clusters were much more consistently centered around 0 along the full range of the level-2 covariate, with a nonsignificantly positive intercept (intercept = 0.056, p value= .858) and nonsignificantly positive slope (slope = 0.036, p value= .727). Based on these results, it is likely that the pattern observed was caused by the two clusters, rather than being a systematic pattern in the data indicative of level-2 nonlinearity. In addition, Model 3 had a smaller RMSE (0.2028) than Model 3a (0.2033), and Model 3 fits better than Model 3a based on BIC and AIC (BIC= 3628.188 for Model 3 vs. BIC= 3772.394 for Model 3a; AIC= 3598.728 for Model 3 vs. AIC= 3739.928 for Model 3a). Taking all of these results together, we considered Model 3 to be preferable to Model 3a. Going forward, Model 3 is used (rather than Model 3a) throughout the model-building process.

Level-2 heteroscedasticity To explore the level-2 heteroscedasticity, a scatter plot of the standardized EB of the random intercept for Model 3 vs. the level-2 covariate was generated (see Fig. 3 [Step 4 (e)]). A Levene's test ($F(21, 1) = 249.4$, p value= .050) suggests that the variance of the standardized EB is constant along the full range of `parentHED` cluster means (indicative of level-2 homoscedasticity).

Level-2 outliers To detect level-2 outliers, a normal probability plot of standardized EB for the random intercept was plotted against a theoretical normal distribution (the figure is not shown). One cluster at the lower end deviated extremely from the line in the normal probability plot. This cluster can also be observed as an outlier of the box plot of standardized EB (the figure is not shown). Similar plots were created with standardized EB for the random slope. In the normal probability and box plots, there were two deviate clusters.

Level-2 influential points There were two level-2 influential clusters, having Cook's distances of 0.261 and 0.319, exceeding the cutoff of $0.174 = 4/23$ for a sample size of 23 schools (the figure is not shown). One of these influential clusters (having a Cook's distance of 0.261) was also the level-2 outlier observed. Because this cluster drastically differs from the rest of the data and is expected to influence parameter estimates, this cluster was removed from the data. Although a second cluster was found to have an influence on parameter estimates (having a Cook's distance of 0.319), it was not found to be an outlier. This means that this cluster is expected to influence parameter estimates in agreeance with the rest of the data, and as a result it is not necessary to remove this cluster from the data.

Outlier removal Because this is the final step of the model-building procedure regarding random effects, the single level-2 outlying cluster was removed from the data. The level-2 outlying cluster contained 19 level-1 units, meaning that the resulting data set after removing this outlier contained 500 (519 - 19) level-1 units and 22 (23 - 1) level-2 clusters.

A second iteration of Steps 1–4 was made with this reduced data set. There were a few differences in the results of Steps 1–4 in this second iteration. First, no level-1 outliers were detected in Step 3 (as opposed to the single level-1 outlier previously detected). Second, the higher-order (square and cubic) terms in Model 3a in Step 4 were no longer significant (p value = .1496 and p value = .1887, respectively). Third, two level-2 influential points were detected in Step 4, having Cook's distances of 0.208 and 0.387, exceeding the cutoff of 0.182 = 4/22 for a sample size of 22 schools. However, neither of these level-2 units were found to be outliers, meaning that these clusters are expected to influence parameter estimates in agreeance with the rest of the data, and as a result their removal from the data was not necessary. Fourth, several median SIQR and SIQR(SIQR) values, as well as the ranking of these values across models, differed between the two iterations. These differing median SIQR and SIQR(SIQR) values are presented and discussed below.

Level-2 normality Normal Q-Q plots were generated to examine whether the standardized EB of the intercept and slope for Model 3 were normally distributed (the figure is not shown). The standardized EB of the intercept were normally distributed, with all standardized EB falling within the 95% confidence bands. The standardized EB of the slope appeared non-normal, with four level-2 units falling outside the 95% confidence bands at the lower extreme. To further examine level-2 normality, histograms were plotted for the standardized EB of the intercept and slope (figures are not shown). The outlying standardized EB of the slope at the lower extreme likely appear to be outliers due to the small number of level-2 units (22, after the outlying level-2 unit was removed). The four smallest standardized EB of the slope ranged from -1.968 to -0.4438, which although not large in magnitude were considered outlying in the normal Q-Q plot because the other 19 standardized EB of the slope ranged from −0.1533 to 0.8252. Because these outlying standardized EB of the slope were not drastically large in magnitude, and the potential level-2 non-normality observed is explainable by the small number of level-2 units, level-2 normality was assumed for Model 3.

Level-1 outliers, level-1 influential points, and level-1 normality Similar plots were created to explore level-1 outliers, influential points, and normality as shown in Step 2. To detect any outliers, the fitted values from Model 3 were plotted against the conditional standardized residuals. No level-1 outliers were detected as having unusually high conditional standardized residuals. The largest observed conditional standardized residuals were −2.751 and 2.704, which although large in magnitude are not unexpected given the large number of level-1 units (519). In addition, no level-1 influential points were detected, as no point had a Cook's distance greater than the cut-off value of 1. The highest Cook's distance detected was 0.01655.

Diagnostic measures and model selection from Steps 1–4 In Table 7, the three diagnostic measures considered for comparing models are RMSE, median SIQR, and SIQR(SIQR), in addition to AIC and BIC. Based on the AIC and BIC presented in Table 7, Model 3 was selected as the best-fitting model regarding the level-1 and level-2 fixed and random effects of `parentHED`. These results agree with the analyses in Step 4 illustrating the importance of the level-2 covariate of `parentHED` cluster means, a parameter which was only included in Model 3. This added

Table 7 Model comparisons regarding diagnostic measures of `Math` data

Model	Fixed Effects	Random Effects	RMSE	AIC	BIC	LL	Median SIQR	SIQR(SIQR)
Null	Intercept	Intercept	0.2177 [4]	3671.709 [4]	3684.347 [4]	-1832.854	0.797 [4]	0.160 [4]
1	Intercept, L-1	Intercept	0.2077 [3]	3628.927 [3]	3645.770 [3]	-1810.464	0.673 [3]	0.115 [1]
2	Intercept, L-1	Intercept, L-1	0.2029 [2]	3618.717 [2]	3643.980 [2]	-1803.358	0.627 [1]	0.123 [2]
3	Intercept, L-1, L-2	Intercept, L-1	0.2028 [1]	3598.728 [1]	3628.188 [1]	-1792.364	0.650 [2]	0.133 [3]

L-1 and L-2 in the above table refer to the Level-1 and Level-2 covariates of `parentHED`, respectively; the values of median SIQR and SIQR(SIQR) in the above table (as well as their rankings from lowest to highest) differ from those results described in Steps 1–4. This is because the values presented in Steps 1–4 were obtained before the outlying level-2 unit was removed from the data, whereas the values presented in the table above were obtained in the second iteration of Steps 1–4; numbers in brackets rank models from worst [4] to best [1] regarding each evaluation measure

complexity of Model 3 was motivated by AIC and BIC, which still ranked Model 3 as the best model despite the penalization for a larger number of parameters. The only diagnostic measures for which Model 3 did not outperform the other models was in median SIQR and SIQR(SIQR). Model 3 had similar median SIQR to both Model 1 and Model 2, indicating that these three models all captured the level-1 variability in the data about equally well. Model 3 had the highest SIQR(SIQR) of the four models, indicating that Model 3 had the highest heteroscedasticity of the four models.[9]

The conditional standardized residuals of Model 3 had no noticeable systematic trend, with residuals scattered uniformly around zero. The lack of a systematic trend in residuals is indicative of Model 3 adequately estimating math scores without omitting a critical fixed or random effect. A Bartels ratio test conducted on the conditional standardized residuals of Model 3 showed that residuals were not significantly nonrandom ($T = 0.0823$, $n = 500$, p value= .5327). Histograms of the level-1 conditional standardized residuals, the level-2 standardized EB of the random intercept, and the level-2 standardized EB of the random slope for Model 3 were plotted to evaluate the normality of residuals (see Fig. 3 [Step 5 (a)]). The level-1 conditional standardized residuals are clearly normally distributed, and based on Shapiro's test the conditional standardized residuals are not significantly non-normal (p value = .0936). The small number of level-2 units makes it difficult to visually determine if the level-2 standardized EB of the random intercept and of the random slope are normally distributed. Shapiro tests concluded that the standardized EB of the random intercept are not significantly non-normal (p value= .337), however, the standardized EB of the random slope are significantly non-normal (p value= .00148). A multivariate normality test of the random intercept and the random slope, Mardia's test, suggested that there is evidence of non-multivariate skewness (Statistic=11.762, p value= .019), but there is evidence of multivariate kurtosis (Statistic=1.782, p value= .075). To conclude, a multivariate normality is assumed because the deviations are not large enough. After selecting Model 3 with the level-1 and level-2 fixed and random effects of `parentHED`, variants of Model 3 were tested with additional level-1 and level-2 fixed effects of the other variables in the data to determine which fixed effects were significant when added to the model in Step 5.

[9]The values of SIQR for each school did not change very much from Model 2 to Model 3, but they changed enough for the ranking of SIQRs for the schools to change between models. Specifically, the 6th smallest SIQR (the one that the first-quartile is largely dependent upon for 22 schools) changed from 0.5461 (for schid= 25456) in Model 2 to 0.4958 (for schid=68493) in Model 3, whereas the median SIQR and third-quartile SIQR stayed largely consistent between Model 2 and Model 3. Because the first-quartile of SIQR was smaller in Model 3, the SIQR(SIQR) increased.

Step 5. Model selection regarding fixed and random effects

Level-1 fixed and random effects Each of the fixed effects of the five additional level-1 variables (cluster-mean-centered `SES`, cluster-mean-centered `homework`, cluster-mean-centered `white`, cluster-mean-centered `sex`, and cluster-mean-centered `race`) was added to the model one at a time. If a fixed effect was significant (p value $< .05$), it would remain included in the model for the remainder of the model building procedure. For example, the fixed effect of cluster-mean-centered (level-1) `SES` was added to Model 3 (with the pre-existing fixed effects β_0, β_1, and β_2) to create Model 4a:

$$y_{ij} = \beta_0 + b_{0j} + \beta_1(x_{ij}^{(1)} - x_{.j}^{(2)}) + b_{1j}(x_{ij}^{(1)} - x_{.j}^{(2)}) + \beta_2 x_{.j}^{(2)} + \beta_3(SES_{ij}^{(1)} - SES_{.j}^{(2)}) + \epsilon_{ij}, \quad (26)$$

where β_3 is the fixed effect of cluster-mean-centered `SES`. If β_3 is significant (p value $< .05$), the fixed effect of cluster-mean-centered `SES` is kept in the model when testing the next fixed effect (cluster-mean-centered `homework`). However, if β_3 is nonsignificant, the fixed effect of cluster-mean-centered `homework` would be tested by adding it to Model 3 (because the fixed effect of cluster-mean-centered `SES` was not kept in the model). Of the five level-1 fixed effects tested, only the fixed effects of cluster-mean-centered `homework` (p value $< .001$) and cluster-mean-centered `white` (p value= .0172) were significant and added to the model. A summary of the models tested is presented in Table 8. Based on the results in Table 8, the final model regarding the additional level-1 fixed effects was Model 4c:

$$y_{ij} = \beta_0 + b_{0j} + \beta_1(x_{ij}^{(1)} - x_{.j}^{(2)}) + b_{1j}(x_{ij}^{(1)} - x_{.j}^{(2)}) + \beta_2 x_{.j}^{(2)} + \beta_3(homework_{ij}^{(1)} - homework_{.j}^{(2)}) + \beta_4(white_{ij}^{(1)} - white_{.j}^{(2)}) + \epsilon_{ij}, \quad (27)$$

where β_3 is the fixed effect of cluster-mean-centered `homework`, and β_4 is the fixed effect of cluster-mean-centered `white`. The addition of the level-1 fixed effects of cluster-mean-centered `homework` and `white` lowered the median SIQR from 0.7013 in Model 3 to 0.627 in Model 4c.

Level-2 fixed effects A similar model-building procedure was used to test the fixed effects of the thirteen additional level-2 variables (`public`, `ratio`, `percmin`, `sctype`, `cstr`, `scsize`, `urban`, `region`, `SES` cluster means, `homework` cluster means, `white` cluster means, `sex` cluster means, and `race` cluster means), with each fixed

Table 8 Model comparisons with additional level-1 fixed effects of `math` data

Model	L-1 Covariate Added	p value	RMSE	AIC	BIC	LL	Median SIQR	SIQR(SIQR)	Covariate Added
Model 3	–	–	0.203	3598.728	3628.188	−1792.364	0.650	0.133	–
Model 4a	ses	0.304	0.203	3597.795	3631.447	−1790.897	0.642	0.127	No
Model 4b	homework*	2.13E-13	0.192	3547.865	3581.517	−1765.932	0.651	0.138	Yes
Model 4c	white*	0.017	0.191	3542.501	3580.342	−1762.251	0.618	0.109	Yes
Model 4d	sex	0.724	0.191	3543.134	3585.160	−1761.567	0.620	0.106	No
Model 4e	race	0.800	0.191	3542.753	3584.779	−1761.377	0.618	0.113	No

*indicates that the effect was significant at the .05 level, and was added to the model

effect (if significant) being added to Model 4c one at a time. None of the thirteen level-2 fixed effects tested were significant when added to the model, with the smallest p value observed being .0914 for `region`. A summary of the models tested is presented in Table 9. None of the additional level-2 covariates was significant. Based on the results in Table 9, the final model with the additional level-1 and level-2 fixed effects was Model 4c. The parameter estimates of Model 3 (without the fixed effects of `homework` and `white`) are compared to those of Model 4c to examine the impact of these added parameters on parameter estimates, shown in Table 10. The estimates and standard errors of the fixed and random effects that were in both Model 3 and Model 4c were similar between the two models. In addition, the residual SD decreased (from 8.507 in Model 3 to 8.031 in Model 4c), indicative of the additional variability in math scores being accounted for in Model 4c with the inclusion of the level-1 fixed effects of cluster-mean-centered `homework` and `white`. Based on these results, Model 4c was selected as the final model regarding all level-1 and level-2 fixed and random effects for all variables.

Evaluation of the selected model The residuals of Model 4c were examined to determine if Model 4c adequately predicted math scores with the included level-1 and level-2 fixed and random effects, and whether the residuals of Model 4c are randomly and normally distributed. A scatter plot of the conditional standardized residuals vs. fitted values based on Model 4c was generated (see Fig. 3 [Step 5 (b)]). The conditional standardized residuals of Model 4c had no noticeable systematic trend, as residuals were scattered uniformly around zero. The lack of a systematic trend in residuals is indicative of Model 4c adequately estimating math scores without omitting a critical fixed or random effect. A Bartels ratio test conducted on the conditional standardized residuals of Model 4c showed that residuals were not significantly nonrandom ($T = -1.0716$, $n = 500$, p value$= .1422$). Histograms of the level-1 conditional standardized residuals, the level-2 standardized EB of the random intercept, and the level-2 standardized EB of the random slope for Model 4c were plotted to evaluate the normality of residuals (these plots are not shown in the paper). The level-1 conditional standardized residuals are clearly normally distributed, which was not contradicted by a Shapiro's test with a p value of .707. The small number of level-2 units makes it difficult to visually determine if the level-2 standardized EB of the random intercept and of the random slope are normally distributed. Shapiro tests concluded that the standardized EB of the random intercept are not significantly non-normal (p value$= .839$), however, the standardized EB of the random slope are significantly non-normal (p value$= .010$). A multivariate normality test of the random intercept and the random slope, Mardia's test, indicated that multivariate normality assumption is rejected because of skewness (Statistic=11.973, p value$= .018$), but not because of kurtosis (Statistic=0.979, p value$= .332$).

Answers to the research question As mentioned earlier, the goals of analysis using the `math` data set was to predict math scores from parents' highest level of education (`parentHED`). Estimates of Model 4c reported in Table 10 were interpreted to answer this research question. Controlling for the level-1 `homework` and `white` covariates, the effect of the level-1 parentHED ($x_{ij}^{(1)} - x_{.j}^{(2)}$) was 2.520 (SE=0.464, p value $< 1e-04$) and the effect of the level-2 parentHED ($x_{.j}^{(2)}$) was 4.549 (SE=0.647, p value $< 1e-04$).

Example 2: Two-level longitudinal data (`HD` data)

Table 11 presents a summary of analyses and results.

Step 0. A preliminary descriptive analysis

The primary research interest is the relationship between depression (measured with the HD rating scale) and the effect of a drug over time (using the `Week` variable for time). To begin, the HD rating was plotted over time (with 6 measurements taken over 5 weeks) for each of the two

Table 9 Model comparisons with additional level-2 fixed effects of `math` data

Model	L-2 Covariate Added	*p* value	RMSE	AIC	BIC	LL	Median SIQR	SIQR(SIQR)	Covariate Added
Model 4c	–	–	0.191	3542.501	3580.342	− 1762.251	0.618	0.109	–
Model 4f	public	0.538	0.191	3540.870	3582.895	− 1760.435	0.623	0.112	No
Model 4g	ratio	0.848	0.191	3546.732	3588.758	− 1763.366	0.620	0.110	No
Model 4h	percmin	0.099	0.192	3543.036	3585.062	− 1761.518	0.656	0.115	No
Model 4i	sctype	0.620	0.191	3542.086	3584.111	− 1761.043	0.623	0.108	No
Model 4j	cstr	0.452	0.191	3542.401	3584.426	− 1761.200	0.625	0.112	No
Model 4k	scsize	0.153	0.191	3542.621	3584.646	− 1761.310	0.614	0.100	No
Model 4l	urban	0.623	0.191	3543.161	3585.187	− 1761.581	0.607	0.106	No
Model 4m	region	0.091	0.192	3540.190	3582.216	− 1760.095	0.623	0.106	No
Model 4n	ses	0.765	0.191	3539.234	3581.260	− 1759.617	0.617	0.109	No
Model 4o	homework	0.104	0.191	3539.831	3581.857	− 1759.916	0.644	0.121	No
Model 4p	white	0.182	0.191	3539.469	3581.495	− 1759.735	0.646	0.116	No
Model 4q	sex	0.891	0.191	3539.869	3581.895	− 1759.935	0.617	0.109	No
Model 4r	race	0.969	0.191	3541.750	3583.775	− 1760.875	0.613	0.109	No

Table 10 Parameter estimate comparisons between Model 3 and Model 4c of `math` data

Covariate	Model 3		Model 4c	
	EST	SE	EST	SE
Fixed				
Intercept	**36.700**	2.126	**37.007**	2.106
$x_{ij}^{(1)} - x_{.j}^{(2)}$	**2.886**	0.510	**2.520**	0.464
$x_{.j}^{(2)}$	**4.653**	0.653	**4.549**	0.647
`homework`	–		**2.098**	0.275
`white`	–		1.907	1.170
Random				
	SD	Correlation $x_{ij}^{(1)} - x_{.j}^{(2)}$	SD	Correlation $x_{ij}^{(1)} - x_{.j}^{(2)}$
Intercept	2.089	−0.241	2.166	−0.367
$x_{ij}^{(1)} - x_{.j}^{(2)}$	1.535		1.310	
Residuals	8.507		8.031	

- indicates not-modeled; EST in bold indicates significance at alpha=.05 level

groups (`Endog` $= 0$, left, and `Endog` $= 1$, right, based on whether or not the depression was endogenous). Figure 4 (Step 0 (a)) shows a clear negative trend in HD rating over time for both groups. The overlapping red (linear trend) and blue (smooth curve) lines in the plots indicate that the negative trend in HD rating was linear.

Step 1. Random intercepts for the clusters

In this step, HD rating (y_{ij}) was modeled without any covariates. The first null model (Null Model Fixed) includes only a fixed intercept:

$$y_{ij} = \beta_0 + \epsilon_{ij}, \tag{28}$$

where y_{ij} is the HD rating for person j at time i, β_0 is the fixed intercept parameter, and ϵ_{ij} is the random error[10]. The second null model (Null Model Random) includes only a random intercept:

$$y_{ij} = \beta_0 + b_{0j} + \epsilon_{ij}, \tag{29}$$

where b_{0j} is the random intercept parameter. The random errors for Null Model Fixed and Null Model Random (ϵ_{ij}) are assumed to be distributed as $N(0, \sigma^2 R_j)$, with $R_j = \Lambda_j C_j \Lambda_j$ for $\Lambda_j = I_{n_j}$ (with homoscedasticity) and $C_j = I_{n_j}$ (with uncorrelated errors), where n_j is the number of observations for person j ($1 \leq n_j \leq 6$).

To examine the multilevel nature of the data (with 6 measurements nested within persons), the standardized errors for Null Model Fixed and the conditional standardized errors for Null Model Random were plotted. For Null Model Fixed, standardized errors varied across persons, as shown in Fig. 4 (Step 1 (a)). The variability in standardized errors across persons in Fig. 4 (Step 1 (a)) is indicative of the multilevel nature of the data. Allowing the intercept to vary across persons in Null Model Random resulted in the conditional standardized errors, presented in Fig. 4 (Step 1 (b)). As shown in Fig. 4 (Step 1 (b)), conditional standardized errors for each person are distributed more consistently around 0 when the intercept is allowed to vary across persons. An ICC = .268 (meaning that 26.8% of the variance in HD scores is accounted for by the variability across persons) supports the conclusion that the data are multilevel, and the inclusion of a random intercept in the model is necessary.

Step 2. Fixed effect of level-1 covariate of interest

In this step the fixed effect of the level-1 covariate `Week` (the linear effect of the drug treatment on HD ratings over time) is added to Null Model Random to create Model 1:

$$y_{ij} = \beta_0 + \beta_1 \texttt{Week}_{ij}^{(1)} + b_{0j} + \epsilon_{ij}, \tag{30}$$

where β_1 is the fixed effect of $\texttt{Week}_{ij}^{(1)}$, with $\texttt{Week}_{ij}^{(1)} = i$ being the number of weeks ($0 \leq i \leq 5$) since person j began the study.

Adding the fixed effect of the level-1 covariate `Week` decreased median SIQR from 0.487 in Null Model Random to 0.467 in Model 1 (indicating that Model 1 captured level-1 variability better than Null Model Random), and slightly decreased the SIQR(SIQR) from 0.178 in Null Model Random to 0.176 in Model 1 (indicating that Model

[10] ϵ_{ij} is referred to as "random error" for the null models (without covariates), and is referred to as "random residual" after covariates are modeled.

Table 11 Summary of analyses and results for HD data

Steps	Diagnostic Measures, Plots, and Tests	Results
Step 0	A plot of the HD rating vs. Week by Endog with liner and smooth lines (Fig. 4 [Step 0 (a)])	The negative trend in HD observed
Step 1	(a) Box plots of con. STD errors for Null Model Fixed (Eq. 28) for each school (Fig. 4 [Step 1 (a)])	The residuals distributed around 0 when adding a random intercept
	(b) Box plots of con. STD errors for Null Model Random (Eq. 29) for each school (Fig. 4 [Step 1 (b)])	⇒Null Model Random is selected
Step 2	SIQR and SIQR(SIQR) comparisons between Null Model Random and Model 1 (Eq. 30)	Reduced SIQR and SIQR(SIQR) in Model 1
Level-1 Linearity	A plot of the mar. STD for Model 1 vs. Week (Fig. 4 [Step 2 (a)])	A smooth curve not needed and with a test of a smooth function
Level-1 Hetero.	(a) Con. STD residuals for Model 1 vs. Fitted values for Model 1 (Fig. 4 [Step 2 (b)])	The residuals not distributed around 0 and significance of the Levene's test
	(b) Levene's test using the con. STD residuals for Model 1	Hetero. assumed ⇒ Model 1a is selected.
Corr. Res.	A plot of ARs of the mar. STD residuals for Model 1a (Fig. 4 [Step 2 (c)])	ARMA(1,0) assumed ⇒ Model 1b is selected.
Level-2 Outliers	Normal Q-Q plot of the STD EB of the intercept for Model 1b (Fig. 4 [Step 2 (d)])	No level-2 outliers ⇒ Model 1b is retained.
Level-2 Influ.	Cook's distance vs. Person IDs (Fig. 4 [Step 2 (e)])	Detected influ., but none were outliers ⇒ Model 1b is retained.
Level-1 Outliers	Con. ind. residuals for Model 1b vs. Fitted values for Model 1b (Fig. 4 [Step 2 (f)])	Six level-1 outliers detected, but not deleted yet ⇒ Model 1b is retained.
Level-1 Influ.	Cook's distance vs. Observation IDs (Fig. 4 [Step 2 (g)])	No level-1 influ. detected ⇒ Model 1b is retained.
Level-1 Normality	(a) Normal Q-Q plot of the con. ind. residuals for Model 1b (Fig. 4 [Step 2 (h)])	Insignificance of the Shapiro–Wilk test and
	(b) Shapiro–Wilk test of the con. ind. residuals	ignorable non-normality in the plots
	(c) Histogram of the con. ind. residuals	Level-1 normality assumed ⇒ Model 1b is retained.
Level-2 Normality	Normal Q-Q plot of the STD EB of the intercept for Model 1b (Fig. 4 [Step 2 (i)])	Level-2 normality assumed ⇒ Model 1b is retained.
Step 3	(a) SIQR(SIQR) for Model 2 (Eq. 31) (Model 1b + random slope)	Increased SIQR(SIQR) due to outliers of SIQR (shown in Fig. 4 [Step 3 (a)])
	(b) OLS regression lines by persons (Fig. 4 [Step 3 (b)])	Variability in intercepts and slopes across persons
	(c) The same analyses for Model 2 as Step 2	Level-1 homo., no-severe outliers and or influ., and normality assumed
		⇒ Model 2 is selected.
Step 4		
Inclusion of Level-2 Cov.	(a) A plot of STD EB of random slope of Model 2 vs. Endog (Fig. 4 [Step 4 (a)])	Highly similar STD EB of random slope and SIQR(SIQR)
	(b) A plot of STD EB of random slope of Model 3 (Eq. 32) vs. Endog (Fig. 4 [Step 4 (b)])	between Model 2 and Model 3
	(c) SIQR(SIQR) comparisons between Model 2 and Model 3 (Fig. 4 [Step 4 (c)])	⇒ Model 2 is retained.
Outliers and Influ.	The same analyses for Model 2 as Step 2	No outliers and influ. detected for Model 2
Step 5	A plot of the con. ind. residuals vs. fitted values for Model 2 (Fig. 4 [Step 5 (a)])	No noticeable systematic patterns
	with Bartels ratio and Durbin-Watson tests	⇒ Model 2 is retained.

"Hetero." indicates heteroscedasticity; "Homo" indicates homoscedasticity; "Corr. Res. " indicates correlated residuals; "Influ." indicates influential points; "Cov." indicates a covariate; Con. ind. indicates conditional independent; "mar. STD" indicates marginal standardized; "STD EB" indicates the standardized empirical Bayes residuals.

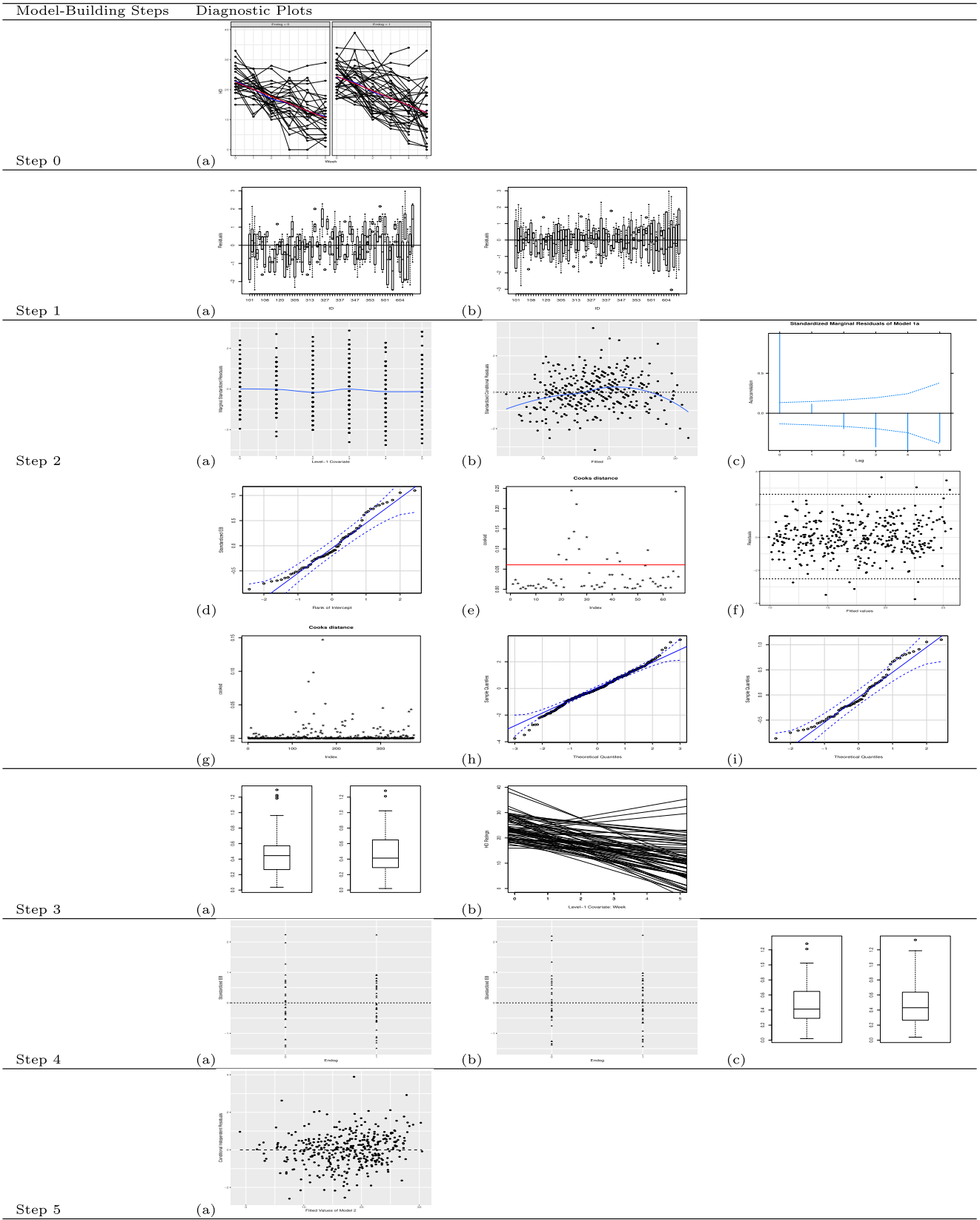

Fig. 4 Diagnostic plots for random effects selection in the two-level longitudinal data (HD data)

1 had highly comparable level-1 heteroscedasticity to Null Model Random).

Level-1 linearity `Week` was plotted against the marginal standardized residuals of Model 1 to examine whether the relationship between `Week` and HD ratings was strictly linear. Figure 4 (Step 2 (a)) shows that there is a linear relationship (the mean of the marginal standardized residuals is approximately equal to zero) between `Week` and the marginal standardized residuals, indicating that there is no higher-order (e.g., square and/or cubic) relationship between `Week` and HD ratings. In addition, a smooth curve fitted to the marginal standardized residuals was not found to be significantly nonlinear ($F = 0.027$, $edf = 1$, p value= .941).

Level-1 heteroscedasticity Fitted values of Model 1 were plotted against the conditional standardized residuals of Model 1 to explore the level-1 heteroscedasticity, as shown in Fig. 4 (Step 2 (b)). The conditional standardized residuals were not evenly distributed around 0 across the full range of fitted values, which is indicative of potential heteroscedasticity. This was further supported by a Levene's test indicating that the conditional standardized residuals were significantly heteroscedastic ($F = 2.666$, $df = 5$, p value= .022).

Level-1 heteroscedasticity was included in the model to create Model 1a. The inclusion of level-1 heteroscedasticity changed the variance of the random residuals (ϵ_{ij}) from being fixed as $\Lambda_j = I_{n_j}$ (constant across time) in Model 1 to being estimable parameters (allowing variance to differ across time) in Model 1a. The fitted values of Model 1a were plotted against the conditional standardized residuals of Model 1a to investigate whether including level-1 heteroscedasticity had an impact (this plot is not shown in the paper). The conditional standardized residuals for Model 1a appear more evenly distributed around 0 than those of Model 1 (particularly for extreme fitted values). A Levene's test indicated that the conditional standardized residuals were no longer significantly heteroscedastic ($F = 0.262$, $df = 5$, p value= .934). In addition, the SIQR(SIQR) decreased from 0.176 for Model 1 to 0.160 for Model 1a, indicating that Model 1a had less level-1 heteroscedasticity than Model 1. Based on these results, level-1 heteroscedasticity was assumed, and Model 1a was used instead of Model 1 for the remainder of the model-building process.

Correlated residuals ARs of the marginal standardized residuals of Model 1a were plotted at each time lag to explore whether the residuals of Model 1a are correlated, as presented in Fig. 4 (Step 2 (c)). Solid lines in Fig. 4 (Step 2 (c)) represent the AR effects at each time lag, with dotted lines indicating the 99% confidence intervals centered at zero. There were significant AR effects at time lags 2-4 for Model 1a. Variations of Model 1a with different residual correlation structures (unstructured, compound, ARMA(1,0), ARMA(2,1), and ARMA(2,2)) were modeled in an attempt to reduce AR. However, results were unobtainable for Model 1a with the ARMA(2,1) and ARMA(2,2) correlation structures, due to the coefficient matrix being uninvertible (possibly due to overfitting). Autocorrelations of the conditional independent residuals for Model 1a with the unstructured, compound, and ARMA(1,0) correlation structures were plotted to examine the effectiveness of these correlation structures at reducing AR (these plots are not shown in the paper).[11] All three correlation structures resulted in decreased AR at each time lag, with all ARs falling within the 99% confidence intervals centered at zero. Although the compound correlation structure resulted in the smallest (or highly similar) AR at each time lag among the correlation structures examined, further analyses showed that the compound correlation structure resulted in a large number of level-1 outliers, with conditional independent residuals ranging from −13.026 to 12.576. Although none of these level-1 outliers were influential enough to merit removal from the model, they resulted in significant violations of level-1 normality. For these reasons, the ARMA(1,0) correlation structure (which had generally lower AR in the residuals than the unstructured correlation structure) was selected instead.[12] Note that all AR with the ARMA(1,0) correlation structure were non-significant at the .01 confidence level, and the level-1 outliers and level-1 normality are less problematic with the ARMA(1,0) correlation structure than with the compound correlation structure, as discussed below. The version of Model 1a with the ARMA(1,0) correlation structure, referred to as Model 1b, was used for the remainder of the model-building process.

For the rest of the model-building process, conditional independent residuals of the fitted models are used for analyses instead of marginal standardized residuals, because errors are now allowed to correlate with the inclusion of the ARMA(1,0) structure in Model 1b.

Level-2 outliers To detect level-2 outliers, a normal Q-Q plot of the standardized EB of the intercept for Model

[11] Conditional independent residuals were plotted instead of marginal standardized residuals (which were plotted for Model 1a) because errors are now allowed to correlate.

[12] Model 1a with ARMA(1,0) was also selected by AIC and BIC of the three candidate models: Model 1a with unstructured, compound, and ARMA(1,0): Model 1a with unstructured (AIC=2244.172, BIC=2338.290), Model 1a with compound (AIC=2286.203, BIC=2325.419), and Model 1a with ARMA(1,0) (AIC=2242.170, BIC=2281.386).

1b was plotted against a theoretical normal distribution, presented in Fig. 4 (Step 2 (d)). The standardized EB of the intercept were largely normal, with several standardized EB falling slightly outside the 95% confidence bands. The standardized EB of the intercept for all level-2 units ranged from -0.864 to 1.098. Based on these results, no level-2 units were considered to be outliers.

Level-2 influential points There were 13 level-2 influential points, having Cook's distances exceeding the cutoff of $0.0606 = 4/66$ for a sample size of 66 persons (see Fig. 4 [Step 2 (e)]). The Cook's distances of these influential points ranged from 0.0857 to 0.242. However, none of these influential level-2 units were considered to be outliers (as these level-2 units had standardized EB of the intercept ranging from -0.709 to 0.0246). Because these influential points are not expected to influence parameters in a way that disagreed with the rest of the data, removing these level-2 units from the data is not necessary.

Level-1 outliers The fitted values from Model 1b were plotted against conditional independent residuals to detect level-1 outliers. As shown in Fig. 4 (Step 2(f)), there were 6 level-1 units detected with high conditional independent residuals, ranging in magnitude from 2.362 to 3.649.

Level-1 influential points No level-1 influential points were detected as having a Cook's distance greater than the cut-off of 1. The highest Cook's distance observed was 0.147, as presented in Fig. 4 (Step 2 (g)). Because no level-1 unit (including those with large conditional independent residuals) was expected to influence parameter estimates, all level-1 units were considered acceptable to remain in the data. If any of the outlying level-1 units had been found to be influential as well, they would be marked for removal from the data in Step 4 (if they were found to be influential outliers in Step 4 as well).

Level-1 normality A normal Q-Q plot was generated to examine whether the conditional independent residuals of Model 1b were normally distributed (see Fig. 4 [Step 2 (h)]). Conditional independent residuals appeared somewhat non-normal in the extremes. In addition, a Shapiro–Wilk test indicated that the conditional independent residuals were significantly non-normal ($W = 0.989$, p value$= .005$). A histogram of the conditional independent residuals was overlaid with a normal curve to further examine normality (this plot is not shown in the paper). The histogram showed that the deviance from normality is not large. Therefore, level-1 normality was assumed for Model 1b based on this analysis.

Level-2 normality A normal Q-Q plot was generated to examine whether the standardized EB of the intercept of Model 1b were normally distributed (see Fig. 4 [Step 2 (i)]). The resulting Q-Q plot shows that the standardized EB of the intercept are mostly normal for Model 1b, with a few standardized EB falling slightly outside the 95% confidence bands. A histogram of the standardized EB of the intercept was plotted to further examine level-2 normality (this plot is not shown in the paper). The standardized EB of the intercept showed no drastic deviations from normality (such as outlying clusters with large standardized EB). Based on these results, level-2 normality was assumed for Model 1b.

Step 3. Random effects of the level-1 covariate

In this step the random effect of the level-1 covariate `Week` was added to Model 1b, creating Model 2:

$$y_{ij} = \beta_0 + b_{0j} + \beta_1 \texttt{Week}_{ij}^{(1)} + b_{1j}\texttt{Week}_{ij}^{(1)} + \epsilon_{ij}, \qquad (31)$$

where b_{1j} is the random slope of $\texttt{Week}_{ij}^{(1)}$. Note that Model 2 still includes the random effect for level-1 heteroscedasticity and the ARMA(1,0) correlation structure.

Adding the random effect of the level-1 covariate `Week` increased the SIQR(SIQR) from 0.117 in Model 1b to 0.174 in Model 2 (indicating that Model 2 had more level-1 heteroscedasticity than Model 1b). To explore this increase in SIQR(SIQR), boxplots of the SIQR for persons were plotted for Model 1b and Model 2, as shown in Fig. 4 (Step 3 (a)). The inclusion of the random effect of the level-1 covariate in Model 2 resulted in a few extreme outlying SIQR, causing the interquartile range of SIQR to "expand" at the upper end to include previously outlying SIQR. This "expansion" resulted in an increase in the SIQR(SIQR) of Model 2.

The ordinary least squares (OLS) regression lines predicting HD rating with `Week` for each person were plotted to show the variability in the effect of `Week` across persons, as presented in Fig. 4 (Step 3 (b)). Variability in the intercepts across persons is indicative of the need for the random effect of the intercept (b_{0j}), whereas variability in the slopes across persons is indicative of the need for the random effect of the slope (b_{1j}).

Plots for level-1 and level-2 outliers, influential points, and normality are not presented to save space. The plots are similar to the plots shown in Fig. 4 (Step 2 (d) - (i)).

Level-2 outliers To detect level-2 outliers, a normal Q-Q plot of the standardized EB of the intercept for Model 2 was plotted against a theoretical normal distribution. No level-2 units (persons) were detected as outliers, with all standardized EB of the intercept falling within the 95% confidence bands. In addition, none of these level-2 units were found to be outliers in a box plot of the standardized

EB of the intercept. The standardized EB of the intercept for all level-2 units ranged from -1.330 to 2.145. Based on these results, no level-2 units were considered to be outliers.

Level-2 influential points There were three level-2 influential points, having Cook's distances of 0.0635, 0.0779, and 0.114, exceeding the cutoff of $0.0606 = 4/66$ for a sample size of 66 persons. None of these level-2 influential points were considered to be outliers (having standardized EB of the intercept of 0.574, 1.373, and -0.770, respectively). Because none of these influential points were outliers, their removal from the data was not necessary.

Level-1 outliers The fitted values from Model 2 were plotted against the conditional independent residuals to detect level-1 outliers. There were 10 level-1 outliers detected, with conditional independent residuals ranging in magnitude from 2.116 to 3.893. They will be investigated on how influential they are.

Level-1 influential points No level-1 influential points were detected, as no point had a Cook's distance larger than the cut-off of 1. The largest Cook's distance observed was 0.027. Because no level-1 unit (including those with large conditional independent residuals) is expected to influence parameter estimates, all level-1 units were considered acceptable to remain in the data.

Level-1 normality A normal Q-Q plot was generated to examine whether conditional independent residuals of Model 2 were normally distributed. Conditional independent residuals appeared somewhat non-normal in the extremes, with several residuals falling outside the 95% confidence bands. To further examine level-1 normality, a histogram of the conditional independent residuals of Model 2 was overlaid with a normal curve. The Q-Q plot and histogram of the conditional independent residuals of Model 2 were highly similar to those for Model 1b, with no extreme violations of level-1 normality detected. As a result, level-1 normality was assumed for Model 2.

Level-2 normality Normal Q-Q plots were generated to examine whether the standardized EB of the intercept and of the slope of Model 2 were normally distributed. The resulting Q-Q plots show that the standardized EB of the intercept and of the slope are mostly normal for Model 2, with only a few standardized EB of the slope falling slightly outside the 95% confidence bands. Histograms of the standardized EB of the intercept and of the slope were plotted to further examine level-2 normality. The histograms of standardized EB of the intercept and of the slope showed no drastic deviations from normality (such as outlying clusters with large standardized EB). Based on these results, level-2 normality was assumed for Model 2.

A Mardia's test was conducted to evaluate the multivariate normality of the standardized EB of the intercept and of the slope of Model 2. The assumptions of multivariate non-skewness ($Statistic = 7.815$, $p = .0986$) and multivariate non-kurtosis ($Statistic = -0.545$, p value$= .586$) were not significantly violated at the .05 significance level. Based on these results, level-2 multivariate normality was assumed for Model 2.

Step 4. Fixed Effects of the Level-2 Covariate In this step the fixed effect of the level-2 covariate `Endog` was added to Model 2, creating Model 3:

$$y_{ij} = \beta_0 + b_{0j} + \beta_1 \texttt{Week}_{ij}^{(1)} + \beta_2 \texttt{Endog}_j^{(2)} + b_{1j} \texttt{Week}_{ij}^{(1)} + \epsilon_{ij}, \tag{32}$$

where β_2 is the fixed effect of the $\texttt{Endog}_j^{(2)}$ level-2 covariate. $\texttt{Endog}_j^{(2)} = 1$ if person j's depression is endogenous, and $\texttt{Endog}_j^{(2)} = 0$ otherwise. Model 3 still includes the random effect for level-1 heteroscedasticity and the ARMA(1,0) correlation structure.

Potential inclusion of the level-2 covariate To explore whether `Endog` should be included in the model, the standardized EB of the random slope for Model 2 (Equation 31, which does not include the level-2 covariate) was plotted for each value of `Endog`, as presented in Fig. 4 (Step 4 (a)). The histogram (overlaid with scatter plots of the standardized EB for group) shows that the mean standardized EB of the random slope for Model 2 was -0.077 when $\texttt{Endog} = 0$, and 0.060 when $\texttt{Endog} = 1$, illustrating the variability in HD ratings unaccounted for by omitting `Endog` in Model 2. The difference between these two groups was not very large (with a mean difference of 0.137). The standardized EB of the random slope for Model 3 (with `Endog` included in the model) was plotted for comparison, as shown in Fig. 4 (Step 4 (b)). With the inclusion of `Endog` in Model 3, the mean standardized EB of the random slope was highly similar between the two values of `Endog` (-0.016 when $\texttt{Endog} = 0$ and 0.012 when $\texttt{Endog} = 1$). The above histograms show that the standardized EB of the random slope were highly similar between Model 2 and Model 3.

The addition of the fixed effect of the level-2 covariate slightly increased the SIQR(SIQR) from 0.1741 in Model 2 to 0.1744 in Model 3. These highly similar SIQR(SIQR) (with the difference between the two models being $<$ 0.0004) indicate that Model 2 and Model 3 have similar levels of level-1 heteroscedasticity. To further illustrate the similarity in SIQR(SIQR) between these two models, boxplots of the SIQR for persons were plotted for Model

Table 12 Model comparisons regarding diagnostic measures of HD data

Model	Fixed Effects	Random Effects	RMSE	AIC	BIC	LL	Median SIQR	SIQR(SIQR)
Null Random	Intercept	Intercept	0.148 [4]	2506.428 [4]	2518.201 [4]	−1250.214	0.487 [4]	0.178 [4]
1b	Intercept, L-1	Intercept	0.135 [3]	2242.170 [3]	2281.386 [3]	−1111.085	0.347 [1]	0.117 [1]
2	Intercept, L-1	Intercept, L-1	0.0819 [2]	2231.860 [2]	2278.919 [1]	−1103.930	0.408 [2]	0.1741 [2]
3	Intercept, L-1, L-2	Intercept, L-1	0.0818 [1]	2228.600 [1]	2279.546 [2]	−1101.300	0.427 [3]	0.1744 [3]

L-1 and L-2 in the above table refer to the level-1 and level-2 covariates of `Week` and `Endog`, respectively; Numbers in brackets rank models from worst [4] to best [1] regarding each evaluation measure

2 and Model 3. Figure 4 (Step 4 (c)) shows that the interquartile range of the SIQR (and by extension the SIQR(SIQR)) are highly similar between Model 2 and Model 3.

Based on these analyses, the level-2 covariate `Endog` was not considered necessary to include in the model. Model 2 was used instead of Model 3 for the remainder of the model building process. Because Model 2 was selected, the analyses for outliers, influential points, and non-normality in this step are identical to the analyses presented in Step 3.

Outlier removal If any level-1 and/or level-2 units were found to be both outlying and influential in this step, they would be removed from the data and Steps 1–4 would be repeated. However, because no outlying and influential level-1 and/or level-2 units were detected for Model 2 in Step 3, such outlier removal was not necessary for this illustration.

Step 5. Model selection regarding fixed and random effects

In this step, the models analyzed in Steps 1–4 are compared regarding differences between their predicted values and the observed data. In Table 12, the diagnostic measures (RMSE, Median SIQR, and SIQR(SIQR)) for the summary of results and model selection methods (AIC and BIC) are reported.[13]

As discussed in Step 4, Models 2 and 3 were highly similar, with the inclusion of the `Endog` variable not found to be necessary. The added model complexity of Model 3 was evaluated by AIC and BIC, with AIC indicating the fixed effect of `Endog` worth including in the model (despite the added complexity), and BIC (which punishes model complexity more harshly than AIC) indicating that this parameter was not worth including in the model (with Model 2 having a lower BIC than Model 3).

[13] Although Model 2 was selected instead of Model 3 in Step 4, Model 3 is included in this table for comparison.

As investigated in Step 3, Models 2 and 3 (which both include the random effect of the level-1 `Week` covariate) had several outlying SIQR for persons, which caused the interquartile range of SIQR across persons (and thus the SIQR(SIQR)) to "expand." As a result, Model 1b (which does not include the random effect of the level-1 covariate, and therefore does not have these outlying SIQR) had the smallest median SIQR and SIQR(SIQR) of the four models. Model 2 had a slightly lower median SIQR and SIQR(SIQR) than Model 3, however, the boxplots of SIQR for persons presented in Step 4 were highly similar between Models 2 and 3. This result indicates that the degrees of variability and heteroscedasticity accounted by Models 2 and 3 are similar.

Taking all results together, Model 2 was selected as the best-fitting model, with level-1 fixed and random effects of `Week`, level-1 heteroscedasticity, and an ARMA(1,0) correlation structure. The added value of the `Endog` variable was not considered significant important to select Model 3. The parameter estimates of the selected model (Model 2) are presented in Table 13.

Evaluation of the selected model The residuals of Model 2 were examined to determine if Model 2 adequately explained HD rating, with `Week`, level-1 heteroscedasticity, and the ARMA(1,0) correlation structure, and whether the conditional independent residuals of Model 2 are randomly and normally distributed. A scatter plot of the conditional independent residuals vs. fitted values of Model 2 was generated, as shown in Fig. 4 (Step 5 (a)). The conditional independent residuals of Model 2 had no noticeable systematic pattern, with residuals being scattered uniformly around zero. The lack of a systematic pattern in the residuals is indicative of Model 2 adequately explaining HD rating without omitting a critical fixed or random effect (such as the level-2 fixed effect of `Endog`). A Bartels ratio test conducted on the conditional standardized residuals of Model 2 showed that residuals were not significantly nonrandom ($T = 4.253, n = 375, p$ value ≈ 1). In addition,

based on a Durbin-Watson test conducted on the conditional standardized residuals of Model 2, it was concluded that the first-order AR was not statistically significantly ($DW = 2.453$, p value ≈ 1).

Histograms of the level-1 conditional independent residuals, the level-2 standardized EB of the random intercept, and the level-2 standardized EB of the random slope for Model 2 were plotted to evaluate the normality of residuals (these plots were not shown in the paper). As discussed in Step 3, level-1 normality in the conditional independent residuals and level-2 normality in the standardized EB of the intercept and of the slope were assumed for Model 2. In addition, the standardized EB of the intercept and of the slope of Model 2 were shown in Step 4 to be multivariate normally distributed.

Answers to the research question Results of the selected model (Model 2) are presented in Table 13. The `weeks` covariate was coded as 0, 1, 2, 3, 4, and 5. Given this coding, the intercept estimate (23.509, SE=0.533) indicates that patients start with an HD score of 23.509 on average. There was nonignorable variability around the average scores across patients ($Var(b_{0j}) = 3.248^2$). The average weekly linear change in HD scores for patients with average drug levels was -2.384 (SE=0.210), indicative of a decrease in the degree of depression over time per week. There was variability in the linear change in HD scores across patients ($Var(b_{1j}) = 1.364^2$) and there was no clear support for an effect of endogeneity of the depression.

Summary and discussion

Residual-based diagnostic plots and measures have been extensively used in single-level linear regression models. However, such plots and measures are rather unusual in model selection and model checking in MLM applications. In this paper, we listed types of random effects presented in MLMs for two-level cross-sectional and longitudinal data, and provided a generic description of these random effects to guide researchers towards selecting the necessary random effects. In addition, we reviewed level-specific diagnostic plots using various kinds of level-specific residuals to select a random effect and to check model assumptions. Furthermore, we presented statistical tests and diagnostic measures to interpret patterns in the diagnostic plots. Using two empirical data sets, the existing and proposed methods were illustrated to demonstrate how to select necessary

Table 13 Parameter estimates of Model 2 of HD data

	Model 2	
Covariate	EST	SE
Fixed		
Intercept	**23.509**	0.533
$\texttt{Week}_{ij}^{(1)}$	**-2.384**	0.210
Random		
	SD	Correlation $\texttt{Week}_{ij}^{(1)}$
Intercept	3.248	-0.143
$\texttt{Week}_{ij}^{(1)}$	1.364	
Residuals	3.186	
Level-1 heteroscedasticity		
	EST	
σ_0	1*	
σ_1	1.232	
σ_2	1.101	
σ_3	1.044	
σ_4	1.052	
σ_5	1.646	
Level-1 ARMA(1,0) correlation structure		
	EST	
ϕ	0.171	

*indicates a model identification constraint; Bold parameter estimates indicate significance at the .05 level based on a t test

fixed and random effects in model-building steps. R code is provided for all analyses conducted in these illustrations.

Guidelines for the use of diagnostic measures, plots, and tests in model-building steps

For the longitudinal and cross-sectional illustrations, only one or two iterations (respectively) of the analyses described were required to select a model to answer research questions. However, longer iterative processes may be necessary in practice. For example, in Step 5 (model selection regarding fixed and random effects), large discrepancies may be found between data and fitted values for a selected model. If the discrepancies stem from data characteristics missed in the earlier steps (e.g., some individuals have different slopes), one can return to Step 3 (random effects of the level-1 covariates) and/or Step 4 (fixed and random effects of the level-2 covariate). Not only do we recommend going through the model-building steps to obtain the best-fitting model to the data, but it may be necessary to use multiple iterations because earlier decisions may look sub-optimal at later steps. An optimal set of fixed and random effects is crucial for 'correct' statistical inferences regarding an effect of interest.

In the illustrative data sets, there is one covariate of interest for which Steps 2–4 were conducted (confirmatory hypothesis testing) and additional covariate(s) (functioning as control covariates) for which Step 5 was conducted (an exploratory approach). When there are multiple covariates of interest from research questions, we suggest conducting Step 2 and Step 3 for *each* of the level-1 covariates of interest and Step 4 for *each* of the level-2 covariates of interest. When there are multiple covariates of interest, the model complexity regarding random effects can dramatically increase. In the present study, a model is built by starting with a null model and then slowly adding fixed and random effects based on the diagnostic measures, plots, and statistical tests as described. As illustrated, the use of diagnostic measures, plots, and tests can be useful to have a parsimonious model that provides an adequate description of the data. The model-building steps starting with a null model tend to keep the models simple (Hox et al., 2018, p. 43).

The model-building steps in the two applications are exploratory in nature, so that in Step 5 hypotheses can be tested regarding covariate(s) of interest. It is possible that decisions leading to the selected model are based on sample variation. When the sample size is large enough, we recommend cross-validation of the selected model (see Camstra & Boomsma, 1992, for review). As an example, Hox et al. (2018) suggested using one half of the data to build up models and using the other half for cross-validation of the selected model.

We use diagnostic measures, plots, and tests for residuals as a supplement to common model selection methods (e.g., AIC and BIC) or significance tests (e.g., Wald test). As presented in the illustrations, we recommend using diagnostic measures, plots, and tests for residuals even when the common model selection methods and significance testing of effects suggest a certain model. For example, in the illustration of the cross-sectional data set, the nonlinearity of the level-2 covariate was observed in a diagnostic plot (a plot of the standardized EB of the random slope vs. the level-2 covariate) in the first iteration (prior to deleting the single level-2 outlier), and based on AIC, BIC, and significance tests of higher-order terms of the covariate. However, we found that the nonlinearity is caused by a single level-2 outlier, based on (a) further analyses of diagnostic plots and measures (plot comparisons of the standardized EB of the random slope vs. the level-2 covariate and RMSE comparisons for models with and without the higher-order terms of the covariates), and (b) analyses of level-specific outliers and influential points in the suggested model-building steps. Based on these results, a model with the linearity of the level-2 covariate (without the single level-2 outlier) was selected.

In the illustrative cross-sectional data set, the single level-2 outlier was detected and removed. When a large number of of outliers are detected, researchers can use robust estimation methods such as the rank based and heavy tailed methods (e.g., Finch, 2017 for comparisons) and robust S-estimation (Copt & Victoria-Feser, 2006) to avoid removing large quantities of the data. We also suggest looking into Demidenko (2004, Section 4.4) for alternative approaches to robust modeling.

As far as normality is concerned, extreme non-normality was not encountered in either of the illustrations, neither of level-1 residuals, nor of EB estimates of random effects. Maas and Hox (2004) found via simulation studies that the non-normality of the level-1 residuals in MLM does not affect the estimates and standard errors of fixed effects, but non-normality does result in biased standard errors of variances of random effects. In addition, Maas and Hox (2004) reported that robust standard errors do not solve the non-normality of the level-1 residuals when the residuals are largely skewed. When there is an extreme deviation from normality in the level-1 residuals, a nonparametric estimate of the bivariate density of the random intercept and slope can be considered using a penalized Gaussian mixture linear mixed model (e.g., Ghidey, Lesaffre, & Eilers, 2004).

Limitations of the present study

This study provides initial guidance to researchers to build up MLMs using diagnostic measures, plots, and tests for

2-level nested data. In applying MLMs for multilevel data having more than 2 nested levels, additional EB of random effects at the 3rd level or higher can be obtained, along with the level-1 residuals and level-2 random effects we described in the current study. Similar diagnostic measures, plots, and tests to those presented for the level-2 data are applicable to multilevel data with more than 2 levels. However, we expect that model-building strategies can be more complex for such data, especially when multiple iterations are desirable (i.e., returning to earlier steps). Future research applying these methods to higher-level data could be useful.

Illustrations of the present study are restricted to a case when there is a single covariate of interest based on a research question and in the presence of control covariates. Although the guidelines of model building for multiple covariates of interest are briefly discussed, step-by-step illustrations are needed in future research. In addition, additional diagnostic plots and tests are needed to explore additional complexities we did not illustrate in the present study. For example, when there are multiple level-1 covariates of interest, a plot of OLS regression coefficients per cluster for the level-1 covariates can be further considered in Step 3.

For the detection of the level-specific influential points, specific detection methods and their cut-off values were used in this study, as used in the MLM literature. To the best of our knowledge, there is no consensus regarding the "correct" detection method of the level-specific influential points and their specific cutoffs to use in MLM. Systematic comparisons of various detection methods are required in future studies. Furthermore, for the detection of the level-specific outliers, the univariate detection method was used for computational efficiency and the use of robust estimation methods was recommended when there are many outliers. However, without further studies on the level-specific outliers, it would be difficult to create an absolute guideline on when to use the univariate method instead of the multivariate detection method and on when to use the robust estimation methods that would be applied in the same way to all MLM applications.

This study uses a single software package (the nlme package in R), to fit MLMs and to calculate level-specific residuals. There are other software packages which provide different kinds of residuals and diagnostic measures, as reviewed by O'Connell et al. (2016) (see Table 4.1) and Loy and Hofmann (2014) (see Table 1). Currently, there are no other software packages which provide the functions required to perform all of the procedures we presented in this paper. Future research is required to provide guidelines on how to replicate the model-building and analyses conducted in this study using other software packages than nlme.

Despite these limitations, our work clearly underscores the benefits of using diagnostic measures, plots, and tests in the applications of MLMs. We hope to encourage researchers to explore and visualize data in model selection and model checking in their applications of MLMs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01709-z.

Acknowledgements Funding was provided in part by the National Science Foundation (SES:1851690) to Sun-Joo Cho, Sarah Brown-Schmidt, and Paul De Boeck. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are grateful to Sonya Sterba (Vanderbilt University) for helpful comments on earlier versions of this article.

References

Bartels, R. (1982). The rank version of von Neumann's ratio test for randomness. *Journal of the American Statistical Association*, *77*, 40–46. https://doi.org/10.1080/01621459.1982.10477764.

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

Bock, R. D. (1983). Within-subject experimentation in psychiatric research. In Gibbons, R. D., & Dysken, M. W. (Eds.) *Statistical and methodological advances in psychiatric research* (pp. 59–90). New York: Spectrum.

Bollen, K. A., & Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. In Fox, J., & Long, J. S. (Eds.) *Modern methods of data analysis* (pp. 11–35). Newbury Park: Sage.

Bryk, A. S., & Raudenbush, S. W. (1992). *Advanced qualitative techniques in the social sciences Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage.

Camstra, A., & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: an overview. *Sociological Methods and Research*, *21*, 89–115. https://doi.org/10.1177/0049124192021001004.

Chatfield, C. (2004). *The analysis of time series: an introduction*, (6th ed.). Boca Raton: Chapman & Hall/CRC.

Claeskens, G. (2013). Lack of fit, graphics, and multilevel model diagnostics. In Scott, M. A., Simonoff, J., & Marx, B. D. (Eds.) *SAGE handbook of multilevel modeling*, (pp. 425–443). Thousand Oaks: Sage, https://doi.org/10.4135/9781446247600.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*, 15–18. https://doi.org/10.2307/1268249.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall. http://conservancy.umn.edu/handle/11299/37076.

Coombs, C. H. (1964). *Theory of data*. New York: Wiley. https://doi.org/10.1177/001316446502500236.

Copt, S., & Victoria-Feser, M. (2006). High-breakdown inference for mixed linear models. *Journal of the American Statistical Association*, *101*, 292–300. https://doi.org/10.1198/016214505000000772.

de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, *11*(1), 57–85. https://doi.org/10.3102/10769986011001057.

Demidenko, E. (2004). *Mixed models: Theory and applications*. Hoboken: Wiley. https://doi.org/10.1002/0471728438.

Demidenko, E., & Stukel, T. A. (2005). Influence analysis for linear mixed-effect models. *Statistics in Medicine*, *24*, 893–909. https://doi.org/10.1002/sim.1974.

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression, I. *Biometrika*, *37*, 409–428. https://doi.org/10.2307/2332391.

Faraway, J. J. (2016). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models* (2 Ed.) Boca Raton: CRC Press. https://doi.org/10.1201/9781315382722.

Farrell, P. J., Salibian-Barrera, M., & Naczk, K. (2007). On tests for multivariate normality and associated simulation studies. *Journal of Statistical Computation and Simulation*, *77*, 1065–1080. https://doi.org/10.1080/10629360600878449.

Finch, W. H. (2017). Multilevel modeling in the presence of outliers: a comparison of robust estimation methods. *Psicológica*, *38*, 57–92.

Finch, W. H., Bolin, J. E., & Kelley, K. (2014). *Multilevel modeling using R*. Boca Raton: CRC Press. https://doi.org/10.1201/9781351062268.

Galecki, A., & Burzykowski, T. (2013). *Linear mixed-effects models using R: A step-by-step approach*. New York: Springer. https://doi.org/10.1007/978-1-4614-3900-4.

Ghidey, W., Lesaffre, E., & Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, *60*, 1412–1425. https://doi.org/10.1111/j.0006-341X.2004.00250.x.

Goldstein, H. (2003). *Multilevel statistical models,* (3 Ed.). New York: Oxford University Press [Distributor]. https://doi.org/10.1002/9780470973394.

Hamilton, M. (1960). A rating scale for depression. Journal of Neurology. *Neurosurgery, and Psychiatry*, *23*, 56–62. https://doi.org/10.1136/jnnp.23.1.56.

Hedeker, D. (2004). An introduction to growth modeling. In Kaplan, D. (Ed.) *Quantitative methodology for the social sciences*. https://doi.org/10.1016/S0005-7894(04)80042-X. Thousand Oaks: Sage.

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

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications*. New York: Routledge. https://doi.org/10.1177/0049124194022003001.

Kreft, I., & de Leeuw, J. (1998). *Introducing statistical methods: Introducing multilevel modeling*. Thousand Oaks: Sage. https://doi.org/10.4135/9781849209366.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963–974. https://doi.org/10.2307/2529876.

Langford, I. H., & Lewis, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *161*, 121–160. https://doi.org/10.1111/1467-985X.00094.

Lesaffre, E., & Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, *54*, 570–582.

Levene, H. et al. (1960). Robust tests for equality of variances. In Olkin, I., & Hotelling, H. (Eds.) *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278–292). Palo Alto: Stanford University Press.

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

Loy, A., & Hofmann, H. (2014). Are you normal? the problem of confounded residual structures in hierarchical linear models. *Journal of Computational and Graphical Statistics*, *24*, 1191–1209. https://doi.org/10.1080/10618600.2014.960084.

Lüdecke, D. (2020). Performance: Assessment of regression models performance. Retrieved from https://easystats.github.io/performance/.

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

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, *2*, 49–55.

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

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*, (2nd ed.). London: Chapman and Hall. https://doi.org/10.1007/978-1-4899-3242-6.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear and mixed models* (2nd Ed.) New York: Wiley.

O'Connell, A. A., Yeomans-Maldonado, G., & McCoach, D. B. (2016). Residual diagnostics and model assessment in a multilevel framework: Recommendations toward best practice. In Harring, J. R., Stapleton, L. M., & Beretvas, S. N. (Eds.) *Advances in multilevel modeling for educational research* (pp. 97–135). Information Age: Charlotte, NC.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer. https://doi.org/10.1007/b98882.

Pinheiro, J. C., Bates, D. M., DebRoy, S., Sarkar, D., & R Core Team (2020). nlme: Linear and nonlinear mixed effects models. R package version 3.1-148. https://CRAN.R-project.org/package=nlme.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Ed.). Thousand Oaks: Sage.

Reisby, N., Gram, L. F., Bech, P., Nagy, A., Petersen, G. O., Ortmann, J., et al. (1977). Imipramine: Clinical effects and pharmacokinetic variability. *Psychopharmacology*, *54*, 263–272. https://doi.org/10.1007/BF00426574.

Rights, J. D. (2019). *On the common but problematic specification of conflated random slopes in multilevel models [Unpublished doctoral dissertation].* Vanderbilt University.

Santos Nobre, J., & da Motta Singer, J. (2007). Residual analysis for linear mixed models. *Journal of Mathematical Methods in Biosciences*, *49*, 863–875. https://doi.org/10.1002/bimj.200610341.

Schabenberger, O. (2004). Mixed model influence diagnostics. *Proceedings of the twenty-ninth annual SAS users group international conference*, *189*, 29.

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

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

Snijders, T. . A. B., & Berkhof, J. (2007). Diagnostic checks for multilevel models. In Meijer, E., & de Leeuw, J. (Eds.) *Handbook of multilevel analysis*, (pp. 141–175). New York: Springer, https://doi.org/10.1007/978-0-387-73186-5.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks: Sage.

Verbeke, G., & Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, *23*(4), 541–556. https://doi.org/10.1016/S0167-9473(96)00047-3.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer. https://doi.org/10.1007/b98969.

von Eye, A., & Bogat, G. A. (2004). Testing the assumption of multivariate normality. *Psychology Science*, *46*, 243–258.

von Neumann, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *The Annals of Mathematical Statistics*, *12*, 367–395. https://doi.org/10.1214/aoms/1177731677.

Wang, Y., de Gil, P. R., Chen, Y.-H., Kromrey, J. D., Kim, E. S., Pham, T., . . . , Romano, J. L. (2016). Comparing the performance of approaches for testing the homogeneity of variance assumption in one-factor ANOVA models. *Educational and Psychological Measurement*, *77*, 305–329. https://doi.org/10.1177/0013164416645162.

Wood, S. N. (2012). On p values for smooth components of an extended generalized additive model. *Biometrika*, *100*, 221–228. https://doi.org/10.1093/biomet/ass048.

Wood, S. N. (2017). *Generalized additive models*. New York: Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279.

Wood, S. N. (2019). mgcv: Mixed gam computation vehicle with automatic smoothness estimation (published on the Comprehensive R Archive Network, CRAN). Retrieved from https://cran.r-project.org/web/packages/mgcv.