Multilevel models with random residual variances for joint modelling school value-added effects on the mean and variance of student achievement

George Leckie, Richard Parker, Harvey Goldstein, Kate Tilling

Centre for Multilevel Modelling

School of Education

University of Bristol, UK

Acknowledgements of support

This research was funded by UK Economic and Social Research Council (ESRC) grant ES/R010285/1 and UK Medical Research Council (MRC) grant MR/N027485/1. This work contains statistical data from Office for National Statistics (ONS), UK. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data.

Harvey Goldstein

Harvey Goldstein passed away on 9th April 2020 aged 80 while this article was in preparation. George, Richard, and Kate worked closely with Harvey over the last 15 years and remember him fondly as an outstanding scholar and mentor.

Address for correspondence

Professor George Leckie

Centre for Multilevel Modelling

School of Education

University of Bristol

35 Berkeley Square

Bristol

BS8 1JA

United Kingdom

g.leckie@bristol.ac.uk

Multilevel models with random residual variances for joint modelling school value-added effects on the mean and variance of student achievement

Abstract

School value-added models are widely applied to study the effects of schools on student achievement and to monitor and hold schools to account for their performances. The traditional model is a multilevel linear regression of student current achievement on student prior achievement, background characteristics, and a school random intercept effect. The predicted random effect aims to measure the mean academic progress students make in each school. In this article, we argue that much is to be gained by additionally studying the variance in student progress in each school. We therefore extend the traditional model to allow the residual variance to vary as a log-linear function of the student covariates and a new school random effect to predict the influence of schools on the variance in student progress. We illustrate this new model with an application to schools in London. Our results show the variance in student progress varies substantially across schools – even after adjusting for differences in the variance in student progress associated with different student groups – and that this variation is predicted by school characteristics. We discuss the implications of our work for research and school accountability.

Keywords: school value-added models, multilevel models, variance functions, mixed-effects location scale models, school effectiveness, school accountability

1. Introduction

School value-added models attempt to estimate the effects of individual schools on student achievement and are widely applied in educational (Goldstein, 1997; Reynolds et al., 2014; Teddlie and Reynolds, 2000; Townsend, 2007) and statistical research (American Statistical Association, 2014; Braun and Wainer, 2007; McCaffrey et al., 2004; Raudenbush and Willms, 1995; Wainer, 2004). They are also used in the US, UK and other school accountability systems where the predicted school effects, often referred to as school value-added scores, provide the basis of reward and sanction decisions on schools (Amrein-Beardsley, 2014; Castellano and Ho, 2013; Koretz, 2017; Leckie and Goldstein, 2017; OECD, 2008). In educational and statistical research, additional interest lies in identifying school policies and practices which predict the school effects and that might therefore prove effective at raising student achievement in schools in general.

The traditional school value-added model is a multilevel linear regression model (Goldstein, 2011; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012) of student current achievement on student prior achievement measured at the start of the value-added period (typically defined as one or more school years or a phase of schooling) and a school random intercept effect to predict the school effects (Aitkin and Longford, 1986; Goldstein et al., 1993; Raudenbush & Bryk, 1986). The adjustment for student prior achievement is fundamental as simpler comparisons of unadjusted school mean achievement would in large part reflect school differences in student achievement present at the start of the value-added period. Such differences are considered beyond the control of the school. Student sociodemographic characteristics are often added to more convincingly adjust for the non-random selection of students into schools (Ballou et al., 2004; Leckie and Goldstein, 2019). Schools with higher school effects are said to add more value: produce higher student achievement for any given set of students. The school effects are argued to reflect the net influences of differences in the quality of teaching, availability of resources, and other policies and practices across schools which are typically unobserved to the data analyst. While concerns remain around potential further omitted student characteristics and selection into schools (Castellano et al., 2014), the predicted school effects from school value-added models are widely viewed as fairer and more meaningful measures to compare schools for research and accountability purposes than comparing simple school mean achievement.

School value-added models are frequently motivated and discussed in terms of measuring the academic 'progress' (learning or improvement) made by students over the value-added period (Goldstein, 1997) and we shall adopt that language here. Specifically, student progress is measured by the difference between observed and predicted student current achievement, that is, the total residual. The total residual is in turn modelled as the summation of the school random intercept effect and the student residual. The school random effect measures the mean student progress in each school. In contrast, the constant residual variance assumes the variance in student progress is the same in every school. This inconsistent modelling of the mean and variance does not seem very realistic. Any given school policy or practice will have different effects on students as a function of their observed and unobserved characteristics and will therefore contribute to the variance in student progress operating in each school. The different sets of school policies and practices operating in each school will therefore lead the variance in student progress to vary across schools as well as the mean.

Studying the variance in student progress in each school would provide valuable new information as to the influence of schools on student learning. Consider two schools which show similar high levels of mean student progress. The traditional school value-added model would view these two schools as equally effective. Suppose, however, the two schools differ in their variance in student progress. Should the two schools continue to be viewed as equally effective? The school which shows higher variance in student progress might now be viewed as the less effective as their positive mean student progress does not appear to be spread as evenly across students as for the school which shows lower variance in student progress. More generally, schools which show more variable student progress might be regarded as showing less control over their student learning. They might also be viewed as struggling to mitigate inequalities in student achievement relative to schools which show less variable student progress. We might then wish to follow up schools which show very low or high variance in student progress to try to identify the specific school policies and practices which lead their students to progress at similar or dissimilar rates. These policies might then be promoted or deterred in schools in general in the same way as is done for school policies and practices which predict higher or lower mean student progress. While we interpret lower school variances in student progress positively, such an interpretation is open to debate. For example, if lower school variances reflect schools which are hindering the progress of their higher progress students, preventing them from reaching their full potential, then lower school variances would then be viewed negatively. We will return to this point in the Discussion.

The aim of this article is to therefore broaden the traditional school value-added model to study the effects of schools on not just mean student achievement, but the variance in student achievement. Specifically, we propose modelling the residual variance in the underlying multilevel linear regression as a log-linear function of an intercept and a new school random effect. The log-linear link function ensures the resulting school-specific residual variances and therefore school variances of student progress are positive. We estimate this variance function simultaneously with the usual mean function and we allow the mean and variance school effects to correlate. Where there is an overall relationship between the residual variance and student prior achievement or other student background characteristics (e.g., where higher prior achieving students in general show less variable progress), we enter these variables into the residual variance function as covariates so that the resulting school variance differences more credibly reflect only variation arising from differences in school policies and practices (and not school differences in student backgrounds). We illustrate this new approach with an application to schools in London.

In biostatistics, our extended multilevel model would be referred to as a 'mixed-effect location scale model' where 'location' and 'scale' in the context of our study refer to the joint modelling of the mean and residual variance in student achievement. Hedeker et al. (2008) introduced this model in the context of studying intensive longitudinal data on mood. Subsequently, Hedeker and others further developed this model and applied it to a range of other longitudinal psychological and health data (e.g., Goldstein et al., 2018; Hedeker et al., 2012; Nordgren et al., 2019; Parker et al., 2021; Rast et al., 2012). This new focus on joint modelling the mean and residual variance in multilevel models is also increasingly being explored in social science research, including in applications to clustered cross-sectional data (Brunton-Smith et al., 2017, 2018; Leckie et al., 2014; McNeish, 2020). However, the applicability of these extended multilevel models to school value-added studies has not yet been explored.

One existing extension to the traditional school value-added model, which partially recognizes that schools influence the variance in student progress, is to include a random slope on student prior achievement. This allows mean student progress in each school to vary systematically as a function of student prior achievement and is sometimes referred to as allowing for 'differential school effectiveness' since schools are now allowed to have different effects on different types of students (Nuttal et al., 1989; Thomas et al., 1997; Strand, 2010). Implicitly, the variance in student progress is then also modelled as a function of student prior achievement. In practice, however, this extension can only be used to account for a limited number (e.g., one, two, three) of observed student characteristics, not to students in general. Thus, random residual variances are still required.

This article proceeds as follows. In Section 2, we discuss the use of school value-added models in accountability in England and thereby motivate our application to London schools. In Section 3, we briefly review the traditional school value-added model and its implementation as a multilevel linear regression model. In Section 4, we propose our new school value-added model which measures the variance in student progress in each school and, where required, adjusts these variances for student covariates. In Section 5, we discuss data, models and software. In Section 6, we present the results. In Section 7, we provide a general discussion, including implications of our work for research and school accountability.

2. School value-added models and accountability in England

In England, since 2004, the Government has published school value-added scores for all secondary schools in the country in annual school performance tables (gov.uk/school-performance-tables). These scores aim to measure the mean student progress shown in each school between the end of primary schooling national Key Stage 2 (KS2) tests (age 11, academic year 6) and the end of compulsory secondary schooling General Certificate of Secondary Education (GCSE) examinations (age 16, academic year 11). The school value-added scores play a pivotal role in the national school accountability system, informing school inspections and judgements on schools. They are also promoted to parents as a source of information when choosing schools for their children. Their high stakes uses and very public presentation have drawn sustained criticism from the academic literature (Goldstein and Speigelhalter, 1996; Leckie and Goldstein, 2009, 2017, 2019). Nevertheless, these authors also argue that when used carefully and collaboratively with schools in a sensitive and less public manner there is still an important role for school value-added models to help identify and understand the different ways schools influence student learning and it is in this spirit that we have carried out the current research (Goldstein, 2020).

Leckie and Goldstein (2017) review the evolution of Government school value-added models and measures in England over time. From 2006-2015, multilevel linear regression models were used, regressing student age 16 GCSE examination score on student age 11 KS2 score and a school random intercept effect. From 2006-2010 these models additionally entered a range of student sociodemographic characteristics which vary across schools and are predictive of student age 16 score (even after adjusting for student age 11 score). In contrast, the current 'Progress 8' model introduced in 2016 (DfE, 2020), is a conventional linear regression of student age 16 score on only student age 11 score (Leckie and Goldstein, 2019; Prior et al., 2021b). Postestimation, the school effects are calculated as school averages of the predicted student residuals. Crucially, while the Government school value-added model has evolved over time, what has remained constant is that the Government have never reported any measure of the variance in student progress in each school and how this variance may potentially be larger in some schools than others.

3. Review of traditional school value-added models

3.1 Random-intercept model

Let y_{ij} denote the current achievement for student i ($i = 1, \dots, n_j$) in school j ($j = 1, \dots, J$). The traditional school value-added model can then be written as the following random-intercept linear regression

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_j + e_{ij} \quad (1)$$

where $\mathbf{x}_{ij}$ denotes the vector of student covariates including the intercept and student prior achievement, $\boldsymbol{\beta}$ the associated vector of regression coefficients, u_j the school random intercept effect, and e_{ij} the student residual. The school random intercept effect and student residual are assumed normally distributed with zero means and constant variances $u_j \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. The random effect and residual are assumed independent of one another and independent of the covariates.

The total residual $u_j + e_{ij}$ measures student progress over the value-added period relative to the overall average student who has a total residual and therefore progress of 0. The random effect u_j measures the mean student progress in each school while the residual e_{ij} measures the progress of each student relative to their school mean. The random effect variance σ_u^2 measures the variation in school mean progress across schools. The residual variance σ_e^2 measures the average variance in student progress within schools (averaged across all schools). Crucially, this parameter is assumed constant across schools (homoscedasticity). Thus, while the model allows mean student progress to vary from school to school u_j, it assumes the variance in student progress is the same in every school σ_e^2.

3.2 Random-slope model

The differential effects version of the traditional school value-added model can be written as the following random-slope linear regression

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_j + e_{ij} \tag{2}$$

where $\mathbf{z}_{ij}$ denotes the vector of covariates with random slopes (a subset of $\mathbf{x}_{ij}$ including the intercept), and $\mathbf{u}_j$ the associated vector of school random effects. The random effects are assumed multivariate normally distributed with zero mean vector and constant covariance matrix $\mathbf{u}_j \sim N(\mathbf{0}, \boldsymbol{\Omega}_\mathbf{u})$. All other terms are defined as before. When $\mathbf{z}_{ij} = 1$ the model simplifies to the random-intercept model (Equation 1).

In the differential effects version of the model, the total residual, now $\mathbf{z}'_{ij}\mathbf{u}_j + e_{ij}$, again measures the progress each student makes over the value-added period relative to the overall average student who again has a total residual and therefore progress of 0. School mean student progress $\mathbf{z}'_{ij}\mathbf{u}_j$ now varies not only across schools, but also across students as a function of the covariates $\mathbf{z}_{ij}$ with random slopes. Thus, this version of the model allows schools to be potentially more or less effective for different types of students.

School mean student progress averaging over all students in each school is given by $\bar{\mathbf{z}}'_{.j}\mathbf{u}_j$ where $\bar{\mathbf{z}}_{.j}$ denotes the average of $\mathbf{z}_{ij}$ in school j. For the purpose of identifying effective schools, it is necessary to evaluate $\bar{\mathbf{z}}'_{.j}\mathbf{u}_j$ at common values of $\bar{\mathbf{z}}_{.j}$ for all schools. The variance in student progress in each school (over all students) is given by $\boldsymbol{u}'_j \mathrm{Var}_j(\mathbf{z}_{ij})\boldsymbol{u}_j + \sigma_e^2$. The first component of this expression $\boldsymbol{u}'_j \mathrm{Var}_j(\mathbf{z}_{ij})\boldsymbol{u}_j$ captures the variance in student progress attributable to interactions between the school effect $\mathbf{u}_j$ and the covariates with random slopes $\mathbf{z}_{ij}$. The magnitude of this component varies across schools. For the purpose of identifying effective schools, it is necessary to evaluate $\boldsymbol{u}'_j \mathrm{Var}_j(\mathbf{z}_{ij})\boldsymbol{u}_j$ at common values of $\mathrm{Var}_j(\mathbf{z}_{ij})$ for all schools. The second component σ_e^2 is attributable to all other sources of variance in student progress. Crucially, this continues to be assumed constant across schools (homoskedasticity). Thus, adding random slopes only partially recognizes that schools may influence the variance in student progress.

4. Proposed new school value-added models

4.1 Random-intercept model with random residual variance function

We allow the variance in student progress to vary across schools by modelling the residual variance as a log-linear function of the covariates and a new random school effect. The variance function can be written as

$$\ln\left(\sigma_{e,ij}^2\right) = \mathbf{w}_{ij}'\boldsymbol{\alpha} + v_j \qquad (3)$$

where $\mathbf{w}_{ij}$ denotes the vector of covariates (typically a subset of $\mathbf{x}_{ij}$ including an intercept), $\boldsymbol{\alpha}$ the associated vector of regression coefficients, v_j the new school random effect assumed normally distributed with zero mean and constant variance, $v_j \sim N(0, \sigma_v^2)$. We additionally assume that u_j and v_j follow a bivariate normal distribution and so u_j and v_j are allowed to correlate.

The variance function allows the inclusion of covariates $\mathbf{w}_{ij}$ and this will typically be necessary. Recall the reason for entering student prior achievement (and potentially further student covariates) into the mean function is that schools should not be held accountable for pre-existing differences in student achievement across schools at the start of the value-added period. A similar argument applies when comparing the variance in student progress across schools since such differences will also reflect school mean differences in student prior achievement even though we have adjusted for this in the mean function. For example, suppose the residual variance decreases with increasing prior achievement. This would suggest that schools with higher mean student prior achievement would in general be expected to show less variable student progress than schools with lower mean student prior achievement. However, following the arguments underpinning the traditional value-added model, this would be viewed as a reflection of their school intake rather than the influence of their school policies and practices. By entering student prior achievement into the model for the variance, we adjust for this overall variance trend. Focus then shifts to how schools deviate from this overall trend. We can calculate school intake adjusted estimates of the school variance in student progress as $\sigma^2_{e,j} = \exp(\bar{\bar{\mathbf{w}}}'_{..}\boldsymbol{\alpha} + v_j)$ where $\bar{\bar{\mathbf{w}}}_{..}$ denotes the overall average value for $\mathbf{w}_{ij}$ across all students and schools (or any other desired value common to all schools).

4.2 Random-slope model with random residual variance function

We can also extend the differential effects version of the model (Equation 2) to include a random residual variance function (Equation 3), in which case $\mathbf{u}_j$ and v_j are assumed multivariate normal distributed. School mean student progress (averaging over all students) is then given by $\bar{\mathbf{z}}'_{.j}\mathbf{u}_j$ as it was in the constant residual variance case and so we will again need to evaluate this at a common value of $\bar{\mathbf{z}}_{.j}$ for all schools. The variance in student progress in each school (over all students) is now given by $\boldsymbol{u}_j \mathrm{Var}_j(\mathbf{z}_{ij})\boldsymbol{u}'_j + \mathrm{E}_j\left(\sigma^2_{e_{ij}}\right)$. Crucially, the second component of this expression is now also free to vary across schools (Equation 3).

5. Data, models and software

5.1 Data

We focus on schools in London whose Progress 8 scores (i.e., school mean progress scores) were published in the Government's 2019 secondary school performance tables. The data are drawn from the National Pupil Database (DfE, 2021a) and consist of 71,321 students in 465 schools (mean = 153 students per school, range = 14 to 330). To make our analyses accessible to a broad audience we standardise student age 16 and age 11 scores to have means of 0 and SDs of 1 so that the measures can be interpreted in SD units. Histograms show the age 16 and age 11 scores are approximately normally distributed (Figure S1 and S2 in the supplemental information) while a scatterplot shows the scores are approximately linearly related with a strong Pearson correlation of 0.72 (Figure S3 in the supplemental information). There are very slight floor and ceiling effects in age 16 scores.

5.2 Models

We fit a series of increasingly complex models. Model 1 is the traditional random-intercept model which only adjusts for student prior achievement and assumes the residual variance and therefore variance in student progress is the same in every school (Section 3.1). Model 2 is the new version of this model which allows this variance to vary randomly across schools (Section 4.1). Model 3 adjusts these school variances for any London-wide relationship between the variance and student prior achievement to better isolate the effects of schools' policies and practices. Model 4 adds a random slope on student prior achievement to the mean function to explore the role of school by prior achievement interactions in inducing variation in the variances of student progress across schools (Sections 3.2 and 4.2). Model 5 adds student sociodemographic characteristics to both the mean and residual variance functions to better measure student progress (remove factors beyond the control of schools) and therefore the school means and variances of student progress (we remove the random slope on prior achievement for simplicity and because it does not prove substantively important in our application). Model 6 adds in school characteristics to attempt to explain school differences in the school means and variances of student progress.

5.3 Software

The traditional school value-added models reviewed in Section 3 are typically fitted via maximum likelihood estimation using conventional multilevel modelling routines in standard software (R, SAS, SPSS, Stata). However, the extended versions of these models proposed in Section 4 cannot be fitted using these routines, nor can then be fitted in specialised multilevel modelling packages (HLM, MLwiN). Hedeker and colleagues have developed the MIXWILD software to fit these models by maximum likelihood estimation (Dzubur et al., 2020) but this proves computationally challenging on larger datasets with many random effects. In contrast, these models can be fitted relatively easily via Markov Chain Monte Carlo (MCMC) methods as implemented in Stata, R (Parker et al., 2021), and Mplus (McNeish, 2020), as well as dedicated Bayesian software such as Stan, WinBUGS, OpenBUGS, and JAGS.

We fit all models using the bayesmh command in Stata (StataCorp, 2021) which implements an adaptive Metropolis-Hastings MCMC algorithm. We use hierarchical centring reparametrisations to improve mixing. We specify vague (diffuse) normal priors for all regression coefficients and minimally informative inverse Wishart prior for the random effects variance-covariance matrices. We specify overdispersed initial values for all parameters. We fit all models with four chains each with 5,000 burnin iterations and 10,000 monitoring iterations. We judge convergence using Gelman-Rubin convergence diagnostics (Gelman and Rubin, 1992) and trace, autocorrelation, and scatter plots. All models converged and all parameters had effective sample sizes > 400. We compare model fit using the deviance information criterion (DIC) (Spiegelhalter et al., 2002). Smaller values are preferred. To support readers wishing to implement these models, we present annotated R and Stata syntax and simulated data in the supplemental information.

6. Results

6.1 Model 1 = Traditional random-intercept school value-added model

Model 1 regresses student age 16 score y_{ij} on student age 11 score x_{ij}. The model is written as $y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_j + e_{ij}$, where $u_j \sim N(0, \sigma_u^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. This model allows mean student progress to vary across schools u_j, but assumes the variance in student progress to be the same in every school σ_e^2. For the purpose of comparing to subsequent models, we parameterise σ_e^2 as $\exp(\alpha_0)$. Plots confirm that the normality assumptions for u_j and e_{ij} are reasonable (Figure S4 in the supplemental information).

Table 1 presents the results. The slope coefficient on student age 11 score β_1 is estimated as 0.678, and so a 1 SD difference in age 11 score is associated with a 0.678 SD difference in age 16 score. The total variance in student progress $\sigma_u^2 + \sigma_e^2$ is estimated to be 0.487 (and so student age 11 scores accounts for 51% of the variation in student age 16 scores $(= 100\{1 - (\sigma_u^2 + \sigma_e^2)\})$. The between-school variance in school mean progress σ_u^2 is estimated as 0.067 and so 14% of the total variation in student progress $(= 100\,\sigma_u^2/(\sigma_u^2 + \sigma_e^2))$ is variation in the schools means. The between-school variance implies an interdecile range (IDR) for the school means of student progress of $(-0.34, 0.32) = \beta_0 \pm \Phi^{-1}(0.90)\sqrt{\sigma_u^2}$, and so students in the most effective schools (operating at the 90th percentile of the distribution of all schools) are predicted to make 0.66 SD more progress (score 0.66 SD higher at age 16 than other students with the same age 11 score) than students in the least effective schools (operating at the 10th percentile). In contrast, the student residual variance $\sigma_e^2 = \exp(\alpha_0)$, estimated as 0.419, is assumed constant, naively implying the variance in student progress is the same in every school.

6.2 Model 2 = Model 1 + random residual variance

Model 2 extends Model 1 by allowing the variance in student progress to vary across schools. Specifically, we model the student residual variance as $\ln(\sigma_{e,j}^2) = \alpha_0 + v_j$ where $v_j \sim N(0, \sigma_v^2)$ and where we allow u_j and v_j to correlate. Thus, this model allows both mean student progress u_j and the variance in student progress $\sigma_{e,j}^2$ to vary across schools.

Model 2 shows a reduction in the DIC of 972 points confirming that the variance in student progress varies significantly across schools. Plots confirm that the normality assumptions for u_j, v_j and e_{ij} are reasonable (Figure S5 in the supplemental information). The mean function parameter estimates are largely unchanged. The residual variance function intercept α_0 and variance of the new school random effect σ_v^2 are estimated as -0.881 and 0.037. The population-averaged school variance in student progress is estimated to be $0.422 = \exp(\alpha_0 + \frac{\sigma_v^2}{2})$, which, as expected, is close to the Model 1 estimate of 0.419. The population IDR of school variances of student progress is estimated to be $(0.32, 053) = \exp\left\{\alpha_0 \pm \Phi^{-1}(0.90)\sqrt{\sigma_v^2}\right\}$. This range is substantial. For example, the difference in age 16 scores between otherwise equal students performing at the 90th and 10th percentile of student progress within the most variable schools $\sigma_{e,j}^2 = 0.53$ is 1.87 SD while in the least variable schools $\sigma_{e,j}^2 = 0.32$ it is 1.46 SD.

Figure 1 plots the predicted school means of student progress u_j (y-axis) against the predicted school variances $\sigma_{e,j}^2 = \exp(\alpha_0 + v_j)$ (x-axis). The means and variances are posterior mean predictions and so have been shrunk towards their population average values as a function of their sample size. The London average values are illustrated by the horizontal and vertical reference lines. Marker size is drawn proportional to school size. The plot visualizes the substantial variation in both school means and variances of student progress described above. The figure also shows a negative association between the school means and variances $r = -0.54$. Thus, schools which are traditionally viewed as more effective by virtue of showing higher mean student progress would tend now to be viewed as doubly effective in that their students not only make high progress but do so consistently across their student intakes (i.e., schools in the top-left quadrant).

Figure 2 presents 'caterpillar plots' of the 465 predicted school means (left panel) and school variances (right panel). Such plots are routinely used by researchers and accountability systems to identify schools that are significant different from average. The distribution of the school variances is positively skewed, consistent with being modeled as log-normally distributed. Schools with fewer students have wider 95% credible intervals than schools with more students. Only 117 out of 465 schools (25%) can be statistically separated from the overall average in terms of their school variances compared to 320 out of 465 schools (69%) when we consider the school means.

6.3 Model 3 = Model 2 + student prior achievement adjustment to random residual variance

Model 3 extends Model 2 by adding student age 11 scores to the residual variance function to adjust for any London-wide relationship between prior attainment and the variance in student progress to better isolate the effects of school policies and practices on the variance in student achievement. The variance function becomes $\ln\left(\sigma_{e,ij}^2\right) = \alpha_0 + \alpha_1 x_{1ij} + v_j$.

Model 3 is preferred to Model 2 ($\Delta\text{DIC} = 34$) showing the residual variance significantly increases with student age 11 scores. However this relationship is very weak. The population IDR of school intake adjusted variances of student progress is effectively the same as in the previous model where we did not adjust for school intake, $(0.32, 0.53) = \exp\left\{\alpha_0 + \alpha_1 \bar{\bar{x}}_{1..} \pm \Phi^{-1}(0.90)\sqrt{\sigma_v^2}\right\}$ where $\bar{\bar{x}}_{1..} = 0$ denotes the London-wide average covariate value for x_{1ij}.

6.4 Model 4 = Model 3 + random slope on student prior achievement

Model 4 is a random-slope version of Model 3 where we add a random slope on age 11 score. The model is written as $y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + u_{1j} x_{1ij} + e_{ij}$, where $u_{0j} \sim N(0, \sigma_{u0}^2)$, $u_{1j} \sim N(0, \sigma_{u1}^2)$ and where u_{0j}, u_{1j} and the random residual variance effect v_j are allowed to correlate. Postestimation, plots confirm that the normality assumptions for u_{0j}, u_{1j}, v_j and e_{ij} are reasonable (Figure S6 in the supplemental information).

Model 4 is preferred to Model 3 ($\Delta\text{DIC} = 301$) confirming the age 11 slope varies significantly across schools. The mean and variance of the age 11 slope across schools β_1 and σ_{u1}^2 are estimated to be 0.672 and 0.004. The latter implies an IDR of school slopes of $(0.59, 0.75) = \beta_1 \pm \Phi^{-1}(0.90)\sqrt{\sigma_{u1}^2}$. Figure 3 visualises this variation for the sample schools by plotting the predicted school lines based on the previous random-intercept model (left panel) and the current random-slope model (right panel). The plots appear very similar suggesting the variance in student progress in each school attributable to the interactions between the school and student prior achievement $u_{1j}x_{1ij}$ while statistically significant is substantively modest.

We calculate the variance in student progress for each school in our sample for a common reference distribution of students with a common mean $\bar{\bar{x}}_{1..} = 0.00$ (the mean of the school means of student prior achievement) and variance $\bar{\sigma}_{x_{1..}}^2 = 0.83$ (the mean of the school variances of student prior achievement) of student age 11 scores. The resulting expression is $u_{1j}^2\bar{\sigma}_{x_{1..}}^2 + \sigma_{e,j}^2$ where $\sigma_{e,j}^2 = \exp(\alpha_0 + \alpha_1\bar{\bar{x}}_{1..} + v_j)$. The first component $u_{1j}^2\bar{\sigma}_{x_{1..}}^2$ gives the variance attributable to the interactions. The IDR in the sample ranges from just 0.00005 to 0.0045. The second component $\sigma_{e,j}^2$ captures all remaining variance. The IDR in the sample ranges from 0.33 to 0.51. In sum, the inclusion of the random slope on prior achievement has done very little to explain the variance in student progress in each school. For this reason and to illustrate the subsequent models as simply as possible, we remove the random slope.

6.5 Model 5 = Model 3 + student sociodemographic characteristics

Model 5 extends Model 3 by adding student age (summer born or not), gender, ethnicity (white, black, Asian, Chinese, mixed, other), first language (English or not), special educational needs (SEN) status, and free school meal (FSM) status into the mean and residual variance functions. Table S1 in the supplemental information presents definitions and summary statistics. Adding these characteristics to the mean function implies students are now compared to other students across London who not only share the same age 11 score, but who also share the same sociodemographic characteristics. The aim is to ensure that schools do not appear more or less effective simply as a result of recruiting more or less educationally advantaged students (Leckie and Goldstein, 2019). The resulting improvement in predicted age 16 scores will lead the student progress scores to in general reduce in absolute magnitude (and reorder) leading the overall variance in student progress to decrease. In turn, the school means and variances of student progress scores will also change, again in general reducing in magnitude and reordering. We then further adjust the school variances of student progress via including the student characteristics in the student residual variance function. This ensures that if there are any London-wide relationships between the variance in student progress and particular student characteristics this again will not benefit or count against schools with disproportionate numbers of these students.

Table 2 presents the results. Model 5 is preferred to Model 3 ($\Delta\text{DIC} = 7247$) confirming the statistical importance of the student characteristics. First consider the mean function. The results show that summer born students, girls, all ethnic minority groups except Mixed ethnicity students (relative to White), and students who speak English as a second language, are all predicted to score higher at age 16, than otherwise equivalent students. SEN and FSM students, in contrast, are predicted to score lower than otherwise equivalent students. These results are established and consistent with the literature (Leckie and Goldstein, 2019). What is not known is whether there are also sociodemographic differences in the variance in student progress. The results show that, all else equal, the residual variance and therefore variance in student progress now increases with age 11 scores and is also higher for SEN and FSM students than for otherwise equal students. Thus, it proves harder to predict reliably the age 16 scores of these student groups relative to other student groups. In contrast, summer born students, girls, Black and Asian students show lower variance in student progress and therefore appear to perform in a more consistent fashion than otherwise equal student groups within schools.

Figure 4 presents scatterplots of the school means and variances of student progress based on the current model which adjusts for student background against those based on Model 3 which ignores student background. We calculate the school variances in each model by plugging in the sample mean values for the covariates $\bar{\mathbf{w}}_{..}$ into $\sigma^2_{e,j} = \exp(\bar{\mathbf{w}}'_{..}\boldsymbol{\alpha} + v_j)$. The plots show both the school means and the school variances are correlated 0.94 across the two models. Thus, schools which show high mean progress when one ignores student background nearly always still show high mean progress after adjustment. The same applies for school variances of student progress. However, even with such high correlations, the rank ordering of those schools whose social mix differ most markedly from the London-wide average still change considerably as shown by schools located furthest away from the 45-degree line in the bottom plots.

6.6 Model 6 = Model 5 + school characteristics

We now shift from attempting to best define and measure student progress, and therefore the school means and variances of student progress, to attempting to explain why some schools show higher mean student progress and lower variance in student progress than others. Unfortunately, we do not observe school policies and practices in our data. However, we do observe school characteristics. We add school type (standard, converter academy, sponsored academy, other), school admissions (comprehensive, grammar, secondary modern), school gender (mixed, boys, girls), and school religion (none, religious) to the mean and residual variance functions. Table S2 in the supplemental information presents definitions and summary statistics.

The results for the existing mean and residual variance function regression coefficients are very similar to before and so we restrict our interpretation here to the new results. First, consider the mean function. Relative to standard school types, school mean progress is in somewhat higher in sponsored and converter academies having adjusted for the other covariates. Similarly, school mean progress is higher in girls schools and religious schools, all else equal. However, the most sizeable differential related to school admissions: school mean progress is considerably higher in grammar schools and lower in secondary modern schools relative to comprehensive schools. These results agree with the literature (Leckie and Goldstein, 2019). With respect to the residual variance function, we see new findings. School variances in student progress tend to be lower in converter academies compared to standard school types, lower in grammar schools versus comprehensive school types, and lower in religious schools versus non-religious schools, and this is after adjusting for London wide relationships between the variance in student progress and student characteristics. Thus, students in converter academies, grammar, and religious schools not only tend to show higher student progress on average, but also tend to show more consistent student progress.

7. Discussion

In this article, we have argued that the focus of school value-added models should broaden to measure not just school mean differences in student progress (student achievement beyond that predicted by student prior achievement and other student background characteristics), but school variance differences in student progress. We have suggested that schools which show lower variance in student progress might, all else equal, be viewed more positively as their lower variances might arguably signal greater control of their students' performances. Put differently, such schools appear to enable students to progress at similar rates and so limit the extent to which their learning might otherwise be overpowered by external idiosyncratic influences. To study school variance differences in student progress, we have proposed extending the traditional school value-added model – a random-intercept linear regression of current achievement on prior achievement and other student background characteristics – to model the residual variance as a log-linear function of the student covariates and a new random school effect. The school random intercept effect and random residual variance in this model measure the school mean and variance in student progress. This model can be viewed as an application of the mixed-effects location scale model popular in biostatics (Hedeker et al., 2008).

We have illustrated our new school value-added model with an application to schools in London. Our results suggest meaningful differences in the variance in student progress across schools. We also find a moderate to large negative association between the school mean and variance in student progress. Thus, schools which show the highest mean student progress also tend to be the schools which show the lowest variance in student progress. These schools might therefore be viewed as doubly effective. One process by which school variance differences may arise is if there is a London-wide negative relationship between the variance in student progress and student prior achievement. We adjusted for this by entering student prior achievement into the residual variance function. A second process by which school variance differences may arise is via interaction effects between the different school policies and practices envisaged to be represented by the school random intercept effect and observed and unobserved student characteristics. Previous research has studied this via entering a school random slope on student prior achievement and this showed schools to be differentially effective for students with low, middle, and high prior achievement. In our application, however, these school-by-student prior achievement interactions are small and explain little of the variation in school variances between schools. We then turned our attention to entering student characteristics into the model, both in the mean and residual variance functions, to better measure student progress. In terms of new results, we find that FSM and SEN students show greater variance in student progress and therefore less predictable age 16 scores than otherwise equal students. The resulting predicted school means and variances of student progress, however, are similar to those based on the model which only adjusts for student prior achievement. Nevertheless, schools whose sociodemographic student mix differ most from the average school still move up and down the London-wide rankings considerably, demonstrating the importance of adjusting for student background at least for some schools (Leckie and Goldstein, 2019). Finally, we shifted our emphasis from measuring school means and variances of student progress to seeking to explain them. We find converter academies and grammar schools tend to show lower variances in student progress than other school types. That is, students in these schools progress at more similar rates versus students in other schools. Importantly, here too we adjusted for any overall relationship between the variance in student progress and student prior achievement and background characteristics and so these differences in school variances lie beyond this simple explanation. Future work might seek to identify whether school variance differences in general as well as those relating to these two school types can be predicted by specific school policies and practices.

Expanding the focus of school value-added models to consider schools effects on the variance in student achievement raises interpretational challenges that future work will need to deal with. In particular, while we have interpreted lower school variances in student progress positively, we acknowledge that this is not necessarily so clear cut. For example, where two schools show the same mean student progress, the school with the smaller variance will not only have fewer students making unacceptably low progress (a positive), but also fewer students making exceptionally high progress (a negative). It is not immediately clear which school would therefore be viewed more positively. Similarly, where one school has a higher variance as well as a higher mean student progress versus a second school, would the first school still be viewed as more effective, especially if its higher variance is such that the first school substantially increases the number of students actually scoring lower than they would have had they attended the second school? More generally, faced now with two summaries of school effects on student learning (mean and variance effects), researchers and school accountability systems must make value judgements as to how to best combine them into any overall summary of school effectiveness for the purpose of making overall inferences, judgements and decisions about schools (Prior et al., 2021a). Crucially, it is only by extending the school value-added model to allow for school effects on the variance in student achievement that such debates are made possible. The extension we have presented paves the way for new substantive research into the reasons behind differences in variability and therefore how best differences should be interpreted.

The new school value-added model presented here can also be extended in various ways beyond simply adding further covariates and random slopes suggesting avenues for new methodological research. First, in the school effectiveness literature, there is interest in studying the consistency of school effects across academic subjects (Goldstein, 1997; Reynolds et al., 2014; Teddlie and Reynolds, 2000; Townsend, 2007). We can extend our new school value-added model to study this phenomenon with respect to the school variance in student progress. Essentially, we would fit a multivariate response version of our model for multiple student achievement scores (Leckie, 2018). The model would have multiple residual variance functions, one for each academic subject. We can then study the correlations of the school means and variances of student progress across subjects. Second, the same multivariate response version of the model can be used to study the stability of school effects over time. Here we would fit a multivariate response model to a single achievement score, but for multiple student cohorts (Leckie and Goldstein, 2009). Third, we could include a random slope in the residual variance function (Goldstein et al., 2018) to study whether schools exacerbate or mitigate any overall relationship between the variance in student progress and student prior achievement. Fourth, while we have flexibly modelled the residual variance, we have not modelled the random intercept variance (the random slope model relaxed this, but in a rather specific way). It is also possible to model the random intercept variance as a log-linear function of school covariates (Hedeker et al, 2008). For example, the variability of school mean progress scores across schools may appear greater for some school groups than others and this could then be tested by introducing the school group variable as a covariate in this second variance function. Fifth, we can expand the model to three levels to incorporate an additional random effect into the mean and residual variance functions relating to, for example, school district and thereby study school district differences in the mean and variance in student progress. This then raises the possibility of entering school district random effects into the school random intercept variance function since school mean progress might vary more in some school districts than in others and so with this extension we can potentially study differential school level inequalities in the education system by school district (Leckie and Goldstein, 2015). Alternatively, teacher random effects could be introduced as a new level between the student and school level. Finally, our focus has been on shifting attention from studying school mean of student progress to additionally focussing on the variance in student progress. In future work it would be interesting to explore further ways the distribution of student progress might vary across schools, for example, with respect to skewness.

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. Journal of the Royal Statistical Society. Series A (General), 149, 1-43.

Amrein-Beardsley, A. (2014). Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability (Routledge).

American Statistical Association (2014). ASA Statement on Using Value-Added Models for Educational Assessment. URL: http://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. Journal of Educational and Behavioral Statistics, 29, 37-65.

Braun, H. I., & Wainer, H. (2007). Value-added assessment. Handbook of statistics, 27, 867-892.

Brunton-Smith, I., Sturgis, P., & Leckie, G. (2017). Detecting and understanding interviewer effects on survey data by using a cross-classified mixed effects location–scale model. Journal of the Royal Statistical Society: Series A (Statistics in Society), 180(2), 551-568.

Brunton-Smith, I., Sturgis, P., & Leckie, G. (2018). How collective is collective efficacy? The importance of consensus in judgments about community cohesion and willingness to intervene. Criminology, 56(3), 608-637.

Castellano, K. E., & Ho, A. D. (2013). A Practitioner's Guide to Growth Models. Council of Chief State School Officers.

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. Journal of Educational and Behavioral Statistics, 39, 333-367.

DfE (2021a). National pupil database. Department for Education, London. URL: https://www.gov.uk/government/collections/national-pupil-database.

DfE (2020). Secondary accountability measures: Guide for maintained secondary schools; academies, and free schools (London, Department for Education).

Dzubur, E., Ponnada, A., Nordgren, R., Yang, C. H., Intille, S., Dunton, G., & Hedeker, D. (2020). MixWILD: A program for examining the effects of variance and slope of time-varying variables in intensive longitudinal data. Behavior Research Methods, 1-25.

Gelman, A., & Rubin., D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical Science, 7: 457–472.

Goldstein, H. (1997). Methods in school effectiveness research. School Effectiveness and School Improvement, 8, 369-395.

Goldstein, H. (2011). Multilevel Statistical Models (4th ed.). Chichester, UK: Wiley.

Goldstein, H. (2020). Living by the evidence. Significance, 17, 38-40.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences, 1, 223-231.

Goldstein, H., Leckie, G., Charlton, C., Tilling, K., & Browne, W. J. (2018). Multilevel growth curve models that incorporate a random coefficient model for the level 1 variance function. Statistical Methods in Medical Research, 27(11), 3478-3491.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D., & Thomas, S. (1993). A Multilevel Analysis of School Examination Results. Oxford Review of Education, 19, 425-433.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society: Series A (Statistics in Society), 159(3), 385-409.

Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of school and college performance. Journal of the Royal Statistical Society. Series A (Statistics in Society), 159, 149-163.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data. Biometrics, 64(2), 627-634.

Hedeker, D., Mermelstein, R.J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. Statistics in Medicine, 31(27), 3328-3336

Koretz, D. (2017). The testing charade: Pretending to make schools better (University of Chicago Press).

Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling heterogeneous variance-covariance components in two-level models. Journal of Educational and Behavioral Statistics, 39, 307-332.

Leckie, G., & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. Journal of the Royal Statistical Society: Series A (Statistics in Society), 172, 835-851.

Leckie, G., & Goldstein, H. (2015). A multilevel modelling approach to measuring changing patterns of ethnic composition and segregation among London secondary schools, 2001-2010. Journal of the Royal Statistical Society: Series A (Statistics in Society), 178, 405-422. DOI: 10.1111/rssa.12066.

Leckie, G., & Goldstein, H. (2017). The evolution of school league tables in England 1992-2016: 'Contextual value-added', 'expected progress' and 'progress 8'. British Educational Research Journal, 43, 193-212.

Leckie, G. and Goldstein, H. (2019). The importance of adjusting for student background in school value-added models: A study of Progress 8 and school accountability in England. British Educational Research Journal, 45, 518-537.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. Journal of Educational and Behavioral Statistics, 29, 67-101.

McNeish, D. (2020). Specifying location scale models for heterogeneous variances as multilevel SEMs. Organizational Research Methods, 24, 630-653.

Nordgren, R., Hedeker, D., Dunton, G., & Yang, C. H. (2019). Extending the mixed-effects model to consider within-subject variance for Ecological Momentary Assessment data. Statistics in Medicine.

Nuttall, D. L., Goldstein, H., Prosser, R., & Rasbash, J. (1989). Differential school effectiveness. International Journal of Educational Research, 13(7), 769-776.

OECD (2008). Measuring improvements in learning outcomes: Best practices to assess the value-added of schools (Organisation for Economic Co-operation and Development Publishing & Centre for Educational Research and Innovation).

Parker, R. M., Leckie, G., Goldstein, H., Howe, L. D., Heron, J., Hughes, A. D., ... & Tilling, K. (2021). Joint modeling of individual trajectories, within-individual variability, and a later outcome: systolic blood pressure through childhood and left ventricular mass in early adulthood. American journal of epidemiology, 190(4), 652-662.

Prior, L., Goldstein, H., & Leckie, G. (2021a). School value-added models for multivariate academic and non-academic outcomes: exploring implications for performance monitoring and accountability. School Effectiveness and School Improvement, 1-22.

Prior, L., Jerrim, J., Thomson, D., Leckie, G. (2021b). A review and evaluation of secondary school accountability in England: Statistical strengths, weaknesses, and challenges for 'Progress 8'. Bristol Working Paper in Education Series. Working Paper, 11/18. Article.

Rast, P., Hofer, S. M., & Sparks, C. (2012). Modeling individual differences in within-person variation of negative and positive affect in a mixed effects location scale model using BUGS/JAGS. Multivariate Behavioral Research, 47, 177-200.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. Sociology of Education, 59, 1–17.

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

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. Journal of Educational and Behavioral Statistics, 20, 307-335.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. School Effectiveness and School Improvement, 25, 197-230.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde., A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B 64: 583–639.

Strand, S. (2010). Do some schools narrow the gap? Differential school effectiveness by ethnicity, gender, poverty, and prior achievement. School Effectiveness and School Improvement, 21(3), 289-314.

StataCorp. (2021). Stata 17 Bayesian Analysis Reference Manual. College Station, TX: Stata Press.

Teddlie, C., & Reynolds, D. (2000). The international handbook of school effectiveness research. Psychology Press.

Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997). Differential secondary school effectiveness: Comparing the performance of different pupil groups. British Educational Research Journal, 23(4), 451-469.

Townsend, T. (Ed.). (2007). International handbook of school effectiveness and improvement: Review, reflection and reframing (Vol. 17). Springer Science & Business Media.

Wainer, H. (2004). Introduction to a special issue of the Journal of Educational and Behavioral Statistics on Value-Added Assessment. Journal of Educational and Behavioral Statistics, 29, 1–2.

Tables

Table 1.

Results for traditional and new school value-added models adjusting only for student prior achievement.

		Model 1 Traditional school value-added model		Model 2 Model 1 + random residual variance		Model 3 Model 2 + student prior achievement adjustment to random residual variance		Model 4 Model 3 + random slope on student prior achievement	
		Est.	SE	Est.	SE	Est.	SE	Est.	SE
		Mean function							
β_0	Intercept	-0.011	0.012	-0.011	0.013	-0.011	0.012	-0.015	0.013
β_1	Age 11 score	0.678	0.003	0.679	0.003	0.679	0.003	0.672	0.004
σ_{u0}^2	School intercept effect variance	0.067	0.005	0.067	0.005	0.067	0.005	0.069	0.005
σ_{u1}^2	School slope effect variance							0.004	0.000
ρ_{u0u1}	Intercept slope effects correlation							0.229	0.067
		Residual variance function							

α_0	Intercept	-0.870	0.005	-0.881	0.010	-0.881	0.011	-0.889	0.011
α_1	Age 11 score					0.029	0.006	0.036	0.006
σ_v^2	School intercept effect variance			0.037	0.003	0.040	0.004	0.040	0.004
		Association between mean and variance function random effects							
ρ_{u0v}	Intercept residual effects correlation			-0.472	0.048	-0.484	0.047	-0.494	0.047
ρ_{u1v}	Slope residual effects correlation							-0.111	0.076
		Fit statistics							
	Deviance information criterion (DIC)	140803		139831		139796		139495	

Note.

Est. and SE denote the posterior means and SDs of the parameter chains.

Table 2.

Results for new school value-added models adjusting for student prior achievement, sociodemographic characteristics, and school characteristics.

		Model 5 Model 3 + student sociodemographic characteristics		Model 6 Model 5 + school characteristics	
		Est.	SE	Est.	SE
		Mean function			
β_0	Intercept	-0.129	0.012	-0.235	0.017
β_1	Age 11 score	0.634	0.003	0.632	0.003
β_2	Summer born	0.045	0.005	0.044	0.005
β_3	Girl	0.219	0.005	0.218	0.005
β_4	Ethnicity: Black	0.015	0.006	0.014	0.007
β_5	Ethnicity: Asian	0.152	0.008	0.150	0.008
β_6	Ethnicity: Chinese	0.296	0.028	0.290	0.028
β_7	Ethnicity: Mixed	0.001	0.009	0.000	0.009
β_8	Ethnicity: Other	0.089	0.010	0.088	0.009
β_9	First language not English	0.162	0.006	0.162	0.006
β_{10}	Special educational needs (SEN)	-0.276	0.008	-0.276	0.008
β_{11}	Free school meal (FSM)	-0.193	0.005	-0.192	0.005
β_{12}	School type: Sponsored academy			0.055	0.025
β_{13}	School type: Converter academy			0.082	0.020
β_{14}	School type: Other			0.023	0.038

β_{15}	School admissions: Grammar			0.396	0.049
β_{16}	School admissions: Secondary modern			-0.118	0.045
β_{17}	School gender: Boys			0.053	0.032
β_{18}	School gender: Girls			0.064	0.027
β_{19}	School religious			0.139	0.022
σ_{u0}^2	School intercept effect variance	0.050	0.004	0.037	0.003
		Residual variance function			
α_0	Intercept	-0.948	0.015	-0.889	0.024
α_1	Age 11 score	0.077	0.006	0.081	0.006
α_2	Summer born	-0.044	0.012	-0.045	0.012
α_3	Girl	-0.059	0.012	-0.061	0.012
α_4	Ethnicity: Black	-0.154	0.016	-0.156	0.016
α_5	Ethnicity: Asian	-0.105	0.018	-0.106	0.018
α_6	Ethnicity: Chinese	-0.088	0.072	-0.080	0.069
α_7	Ethnicity: Mixed	-0.028	0.022	-0.035	0.021
α_8	Ethnicity: Other	-0.014	0.020	-0.015	0.021
α_9	First language not English	-0.002	0.013	-0.005	0.013
α_{10}	Special educational needs (SEN)	0.204	0.016	0.203	0.016
α_{11}	Free school meal (FSM)	0.103	0.012	0.099	0.012
α_{12}	School type: Sponsored academy			0.011	0.028
α_{13}	School type: Converter academy			-0.048	0.023
α_{14}	School type: Other			0.053	0.042
α_{15}	School admissions: Grammar			-0.280	0.052
α_{16}	School admissions: Secondary modern			-0.068	0.044

α_{17}	School gender: Boys			0.002	0.034
α_{18}	School gender: Girls			0.015	0.029
α_{19}	School religious			-0.110	0.023
σ_v^2	School intercept effect variance	0.032	0.003	0.026	0.003
		Association between mean and variance function random effects			
ρ_{u0v}	Intercept residual effects correlation	-0.409	0.050	-0.282	0.057
		Fit statistics			
	Deviance information criterion (DIC)	132549		132539	

Note.

Est. and SE denote the posterior means and SDs of the parameter chains.

Student ethnicity reference group is White.

School type reference group is standard.

School admissions reference group is comprehensive.

School gender reference group is mixed-sex school.

Tables S1 and S2 present definitions and summary statistics of all student and school characteristics.

Figures

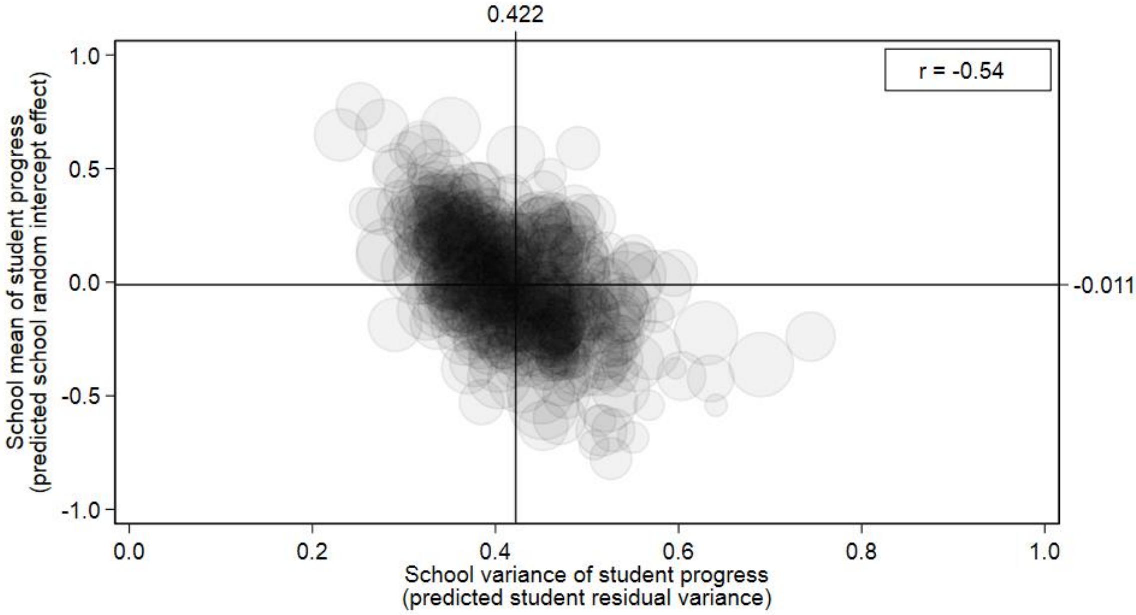

Figure 1.

Model 2 scatterplot of school means of student progress scores against school variances of student progress scores. The London average values are shown by horizontal and vertical reference lines. Marker size is proportional to school size.

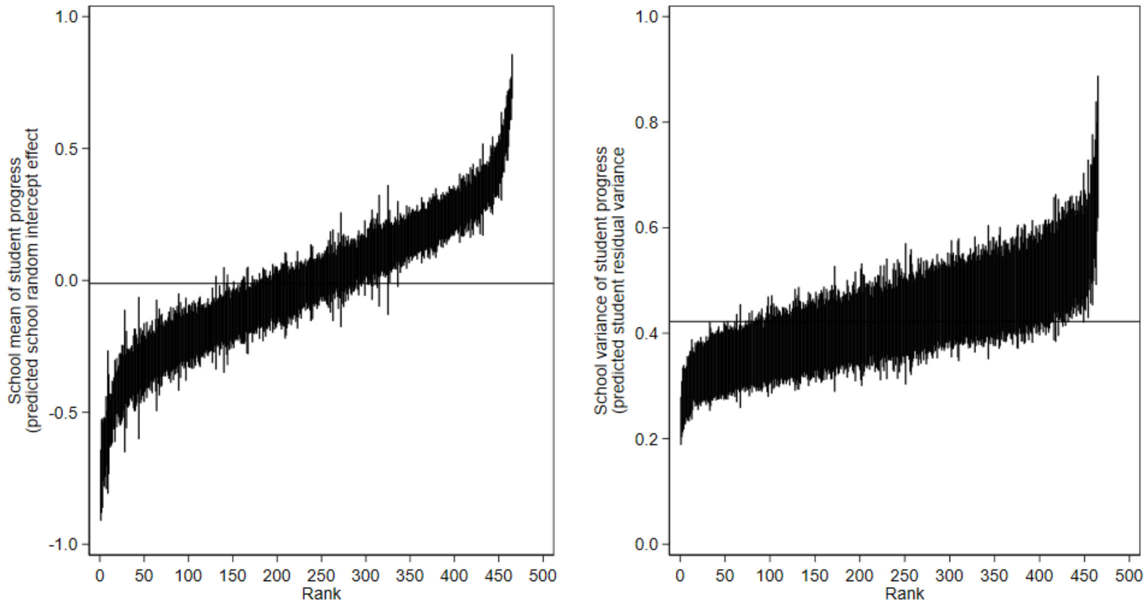

Figure 2.

Model 2 caterpillar plots for school means (left) and school variances right) of student progress presented in rank order. Posterior means with 95% credible intervals.

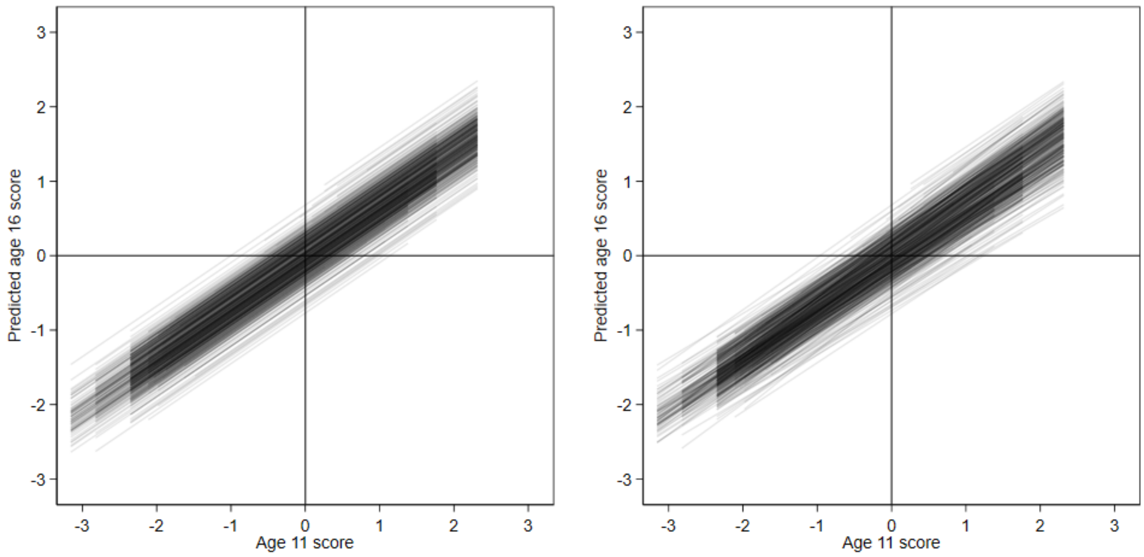

Figure 3.

Model 3 and 4 school regression lines of predicted age 16 scores against age 11 scores for random-intercept model (left) and random-slope model (right).

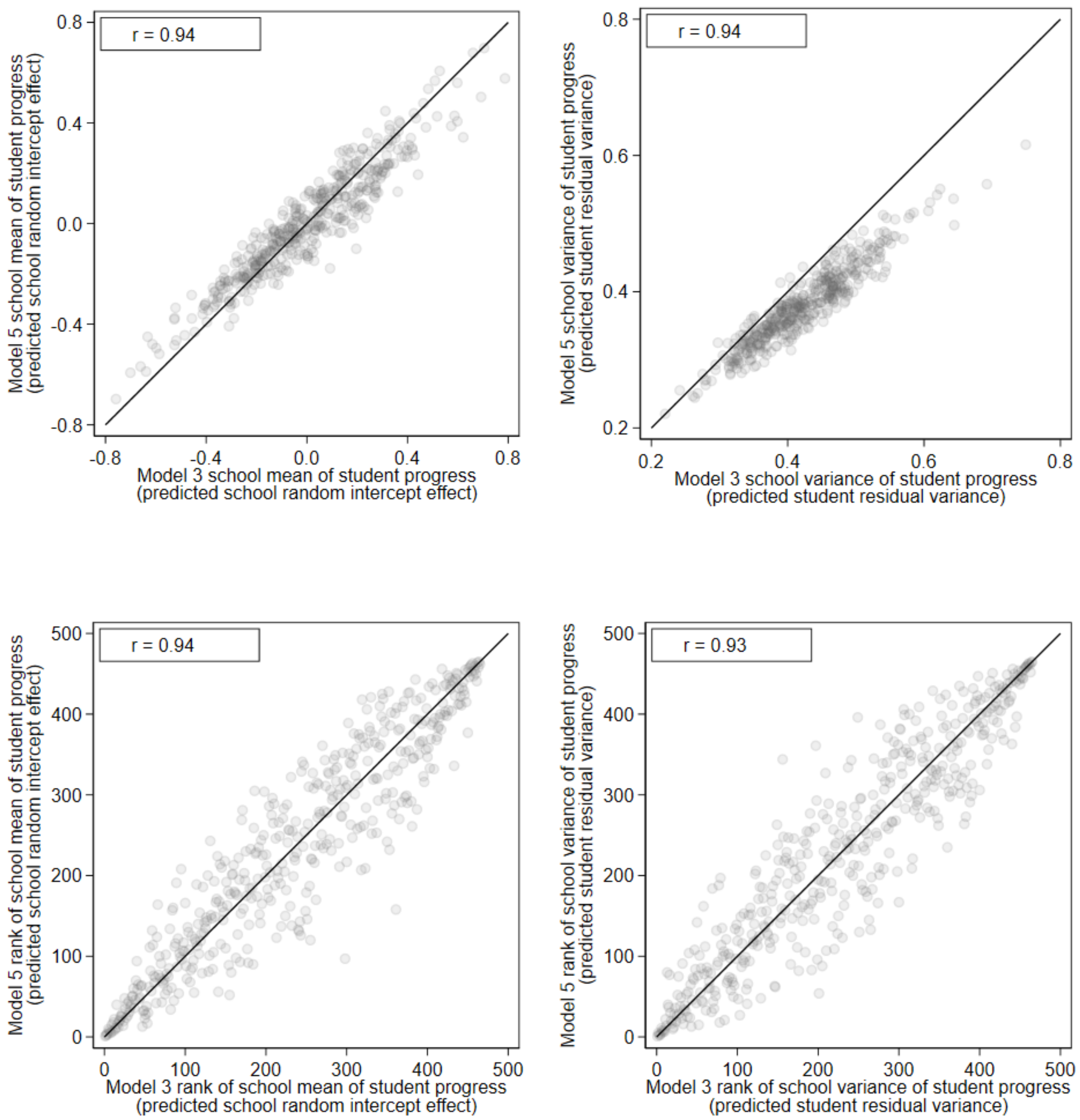

Figure 4.

Model 5 against Model 3 scatterplots of school means of student progress (top left), school variances of student progress (top right), ranks of school means of student progress 9bottom left), and ranks of school variances of student progress (bottom right).

Supplemental information

S1. Supplemental tables

Table S1.

Summary statistics for the student characteristics ($n = 71{,}321$).

	n	%
Age		
Not summer born	52,957	74.3
Summer born	18,364	25.8
Gender		
Boy	35,338	49.6
Girl	35,983	40.5
Ethnicity		
White	28,070	39.4
Black	15,633	21.9
Asian	14,987	21.0
Chinese	447	0.6
Mixed	5,795	8.1
Other	6,389	9.0
Language		
English	42,789	60.0
Not English	28,532	40.0
Special educational needs (SEN)		
Not SEN	61,189	85.8
SEN	10,132	14.2
Free school meal (FSM		

Not FSM	46,500	65.2
FSM	24,821	34.8

Note.

Summer born is defined as those born in June, July, or August.

FSM is defined as eligibility for Free School Meals (FSM) in any of the previous six years.

Table S2.

Summary statistics for the school characteristics ($n = 465$).

	n	%
Type		
Standard	151	32.5
Sponsored academy	93	20.0
Converter academy	184	39.6
Other	37	8.0
Admissions		
Comprehensive	425	91.4
Grammar	19	4.1
Secondary modern	21	4.5
School gender		
Mixed	340	73.1
Boys	50	10.8
Girls	75	16.1
Religious		
No	349	75.1
Yes	116	25.0

Note.

A range of school types operate in London and we have categorised these into four groups. Standard school type encompasses community, foundation, voluntary aided, voluntary controlled, and city technology colleges. In contrast to standard school types, academies receive their funding directly from the government rather than through local authorities (school districts). There are two types of academies. Sponsored academies are mostly underperforming schools which have been changed to academy status and run by sponsors. Converter academies are schools deemed to be performing well that have converted to academy status. Other school type encompasses free, studio, university technology colleges (UTCS), and further education colleges. These are more technically or vocationally oriented schools.

A minority of local authorities in London operates selective admissions. In these local authorities grammar school select students based on high performance in entrance examinations and so by definition have high mean age 11 scores and tend also to be educationally advantaged and homogenous in terms of student sociodemographic characteristics. Secondary modern schools take those students not admitted to grammar schools.

S2. Supplemental figures

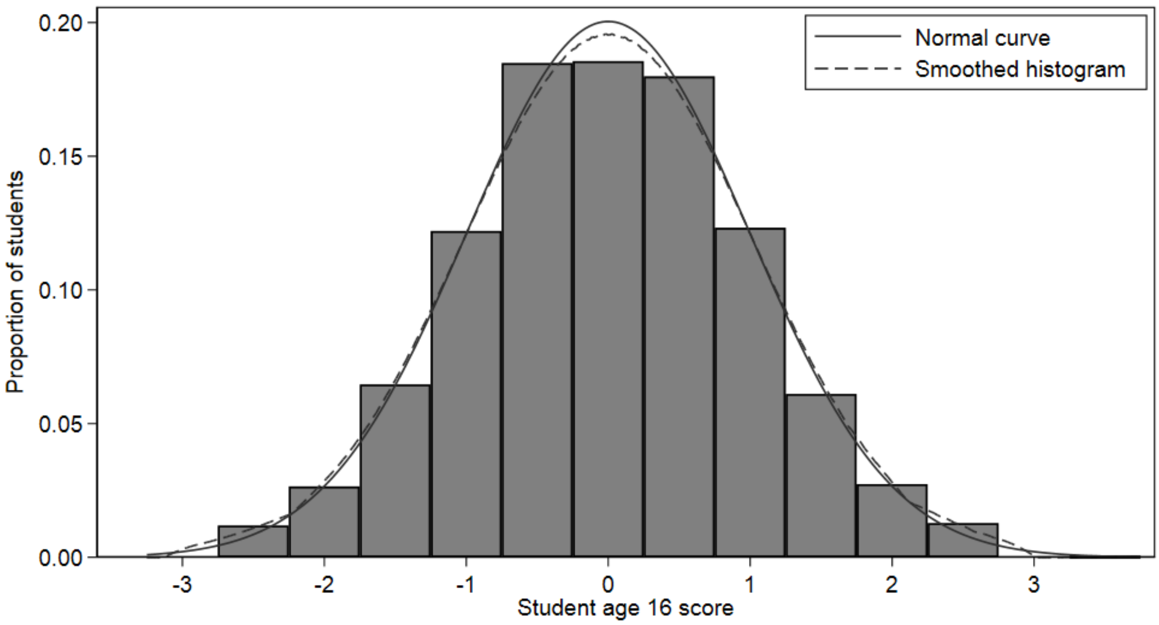

Figure S1.

Histogram of student age 16 scores.

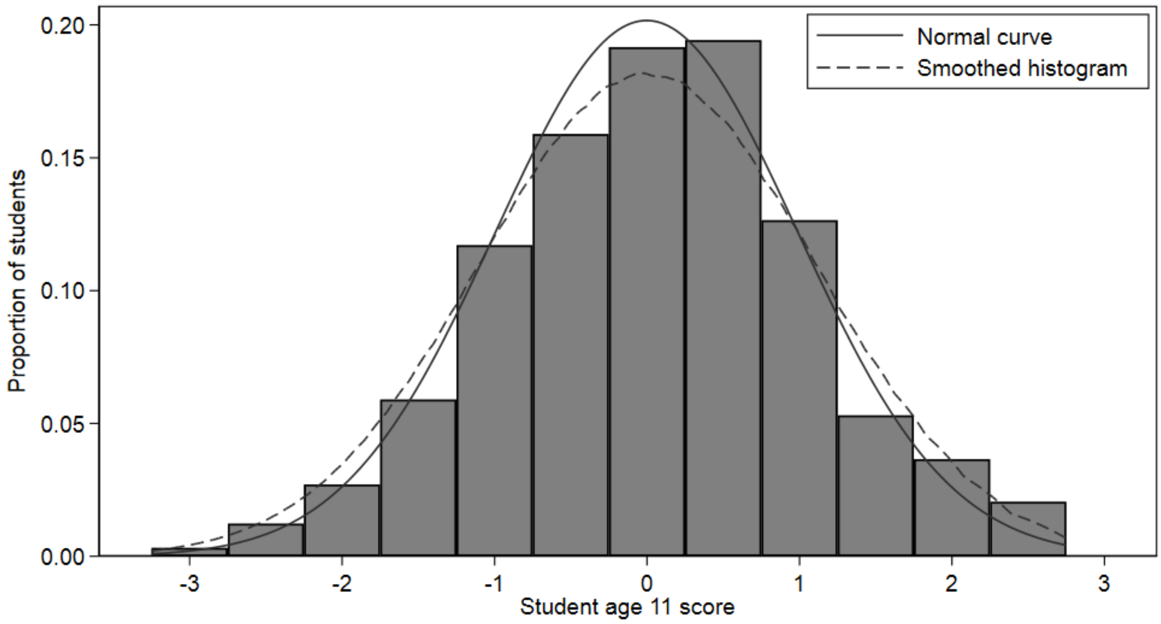

Figure S2.

Histogram of student age 11 scores.

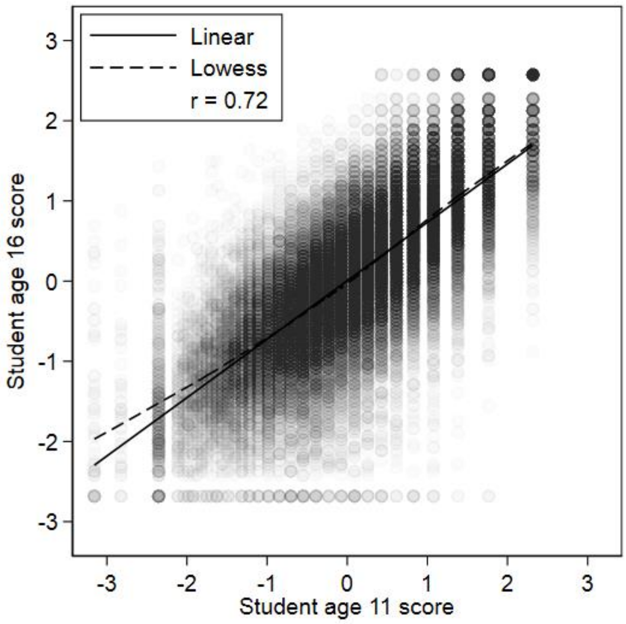

Figure S3.

Scatterplot of student age 16 scores against age 11 scores.

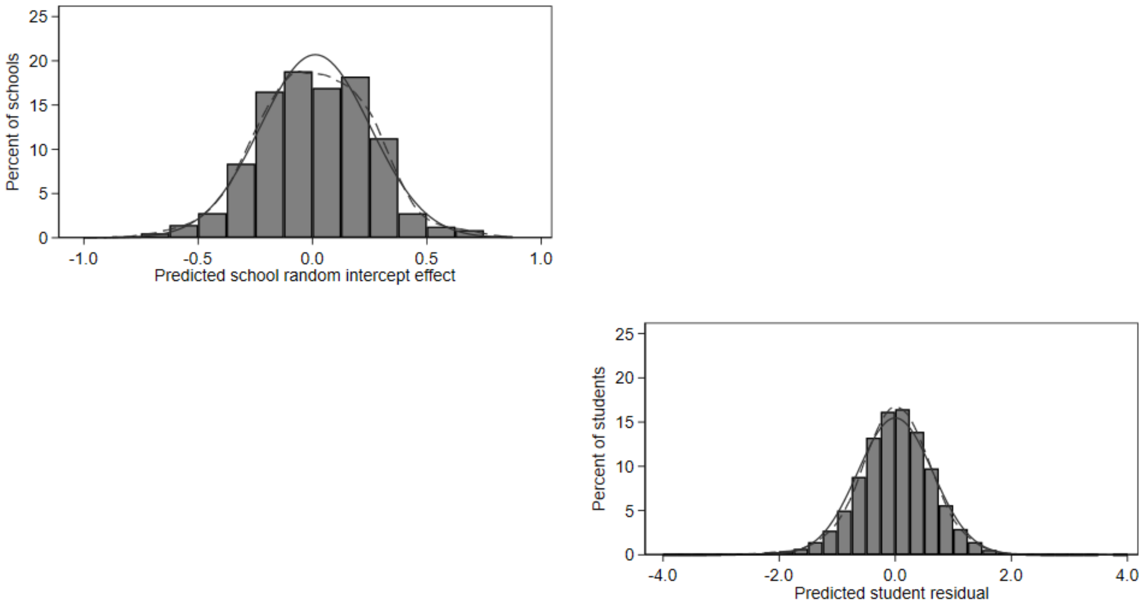

Figure S4.

Model 1 histograms of predicted school random intercept effects (top left) and student residuals (bottom right), each with superimposed normal curves (solid) and kernel density curves (dashed).

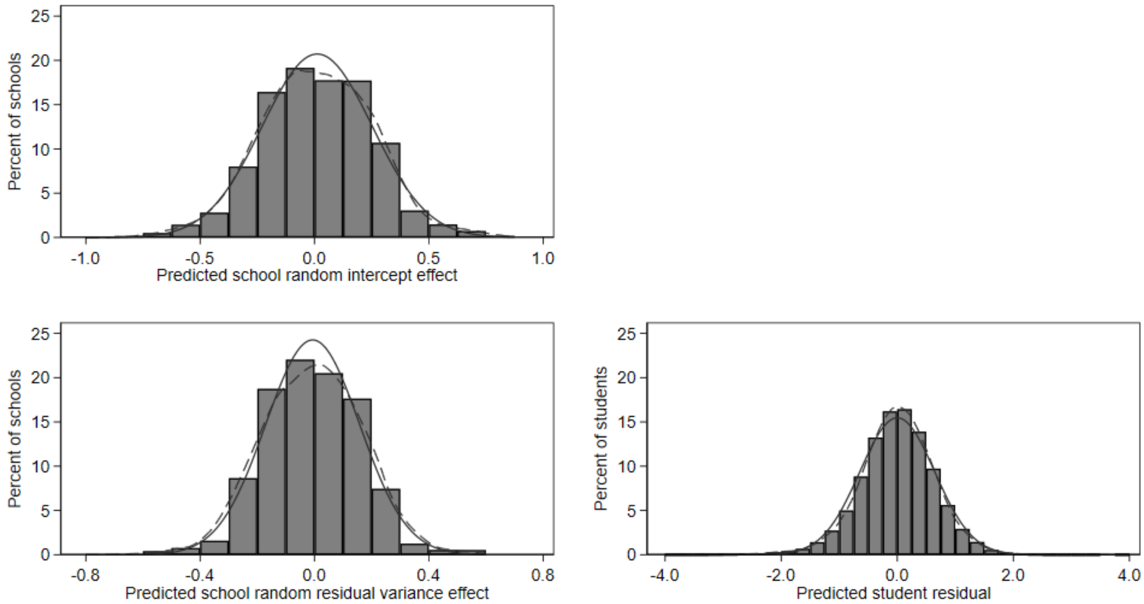

Figure S5.

Model 2 histograms of predicted school random intercept effects (top left), school random residual variance effects (bottom left), and student residuals (bottom right), each with superimposed normal curves (solid) and kernel density curves (dashed).

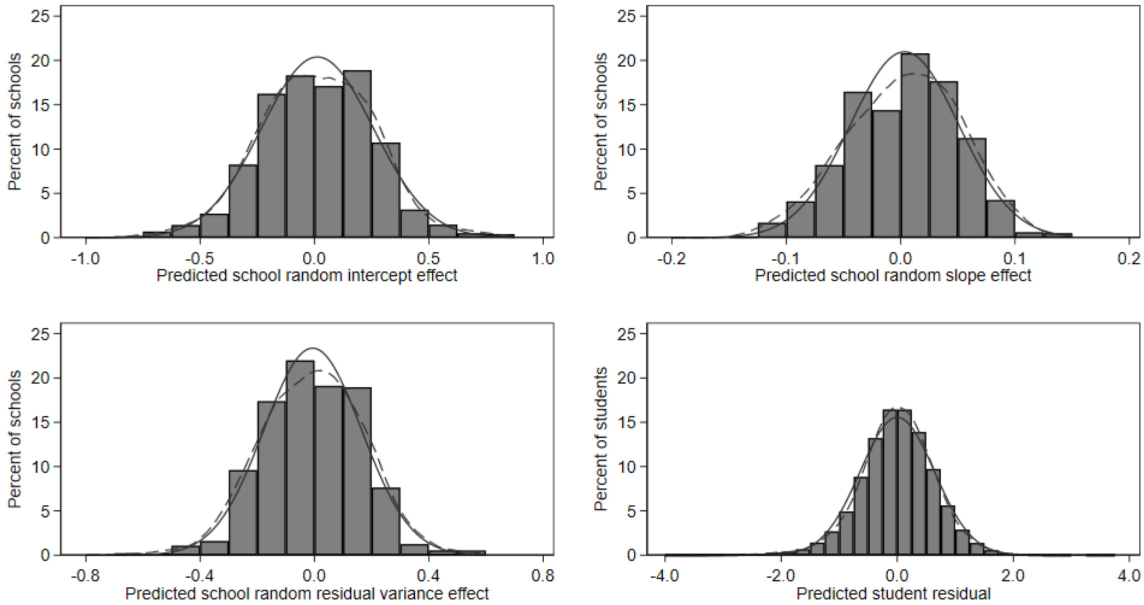

Figure S6.

Model 4 histograms of predicted school random intercept effects (top left), school random slope effects (top right), school residual variance effects (bottom left), and student residuals (bottom right), each with superimposed normal curves (solid) and kernel density curves (dashed).

S3. Stata and R software syntax and simulated data for fitting the models

In this section, we describe Stata and R syntax to fit the models explored in this article. To support readers, we provide script files and data to replicate the presented analysis.

Example model

For simplicity, we focus on the two-level random-intercept model with a random residual variance function presented in Sections 3.1 and 4.1. To illustrate the syntax as simply as possible, we consider a version of this model with only one student characteristic (student prior achievement). This model can be written as

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \qquad \text{(S1)}$$

$$\ln\left(\sigma^2_{e,ij}\right) = \alpha_0 + \alpha_1 x_{ij} + v_j$$

$$\begin{pmatrix} u_j \\ v_j \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_u & \\ \sigma_{uv} & \sigma^2_v \end{pmatrix}\right\}$$

This model is the same as Model 3 presented in the article. Model 1 presented in the article can be viewed as a constrained version of this model (where the residual variance is assumed constant across all students and schools). Models 2, 5, and 6 vary in the covariates included in both the mean and residual variance functions. Model 4 is a random-slope version of this model (where a random slope is added to prior attainment).

Simulated data

As we cannot share the data analysed in the article, we analyse here simulated data where we use the above model as the data generating model. We simulate a single dataset with 100 schools and 25 students per school We simulate x_{ij} as standard normal variate with intraclass correlation of 0.2. We specify the true parameter values as $\beta_0 = 0$, $\beta_1 = 0.7$, $\sigma_u^2 = 0.05$, $\alpha_0 = -0.8$, $\alpha_1 = 0.05$, $\sigma_v^2 = 0.05$, $\sigma_{uv} = 0.025$. The resulting data can be found in data.dta.

Stata: The bayesmh command

We focus on the bayesmh Stata command (StataCorp, 2021). The bayesmh command implements an adaptive Metropolis-Hastings MCMC algorithm. We present the simplest possible syntax noting that mixing can be improved via model reparameterization (e.g., hierarchical centring) and by specifying various estimation options (initial values, blocking) and we encourage readers to consult the documentation for further details. The syntax is as follows.

```
. bayesmh y x U[school], ///
    likelihood(normal(exp({lnsigma2e:x,xb} + {V[school]}))) ///
    prior({y:}, normal(0, 10000)) ///
    prior({lnsigma2e:}, normal(0, 10000)) ///
    prior({U} {V}, mvnormal(2, 0, 0, {SIGMAUV, matrix})) ///
    prior({SIGMAUV, matrix}, iwishart(2, 3, S))
```

Line 1 of the syntax specifies the mean function. Line 2 specifies the normal response distribution and the residual variance function. The intercept is included in both functions by default. Lines 3 and 4 specify diffuse normal priors for the regression coefficients in each function with means of 0 and variances of 10000. Line 5 specifies the random effects to be bivariate normally distributed with zero means and a constant covariance matrix. Line 6 specifies a minimally informative inverse Wishart distribution for this covariance matrix (where S is pre-specified matrix such as an identity matrix).

The associated model output is as follows

```
Burn-in 2500 aaaaaaaaa1000aaaaaaaaa2000aaaaa done
Simulation 10000
.........1000.........2000.........3000.........4000.........5000.........6000.....
.
> ...7000.........8000.........9000.........10000 done

Model summary
------------------------------------------------------------------------------
Likelihood:
  y ~ normal(xb_y,exp({lnsigma2e:x,xb} + {V[school]}))

Prior:
  {y:x _cons} ~ normal(0,10000)                                            (1)

Hyperpriors:
    {lnsigma2e:x _cons} ~ normal(0,10000)
  {U[school] V[school]} ~ mvnormal(2,0,0,{SIGMAUV,m})
            {SIGMAUV,m} ~ iwishart(2,3,S)
------------------------------------------------------------------------------
(1) Parameters are elements of the linear form xb_y.

Bayesian normal regression                       MCMC iterations  =     12,500
Random-walk Metropolis-Hastings sampling         Burn-in          =      2,500
                                                 MCMC sample size =     10,000
                                                 Number of obs    =      2,500
                                                 Acceptance rate  =       .196
                                                 Efficiency:  min =    .002048
                                                              avg =     .01894
Log marginal-likelihood                                       max =     .04904

------------------------------------------------------------------------------
             |                                                Equal-tailed
             |      Mean   Std. dev.     MCSE     Median  [95% cred. interval]
-------------+----------------------------------------------------------------
y            |
           x |  .7005516   .0142504   .000858   .7005228   .6731723   .7301972
       _cons | -.0282751   .0220327   .001753  -.0289124  -.0706135   .0183987
-------------+----------------------------------------------------------------
lnsigma2e    |
           x |  .0520455   .0299735   .001694   .0520977  -.0059208   .1090707
       _cons | -.7796784   .0347738    .00157  -.7794015  -.8489381  -.7097778
-------------+----------------------------------------------------------------
 SIGMAUV_1_1 |  .0340151   .0068176   .001046   .0338658   .0200429   .0490111
 SIGMAUV_2_1 |  .0137237   .0077867   .001553   .0138888  -.0011668   .0284148
 SIGMAUV_2_2 |  .0431452   .0209756   .004636   .0395217   .0092352   .0902076
------------------------------------------------------------------------------
Note: There is a high autocorrelation after 500 lags.
Note: Adaptation tolerance is not met in at least one of the blocks.
```

The command ran one chain with 2500 burn-in iterations and 10000 monitoring iterations. The reader should note the warning messages at the end of the output. As noted above, mixing can be improved via model reparameterization and by specifying various estimation options. The results presented in tabular form are as follows

	Est.	SE
β_0	-0.028	0.022
β_1	0.701	0.014
σ_u^2	0.034	0.007
α_0	-0.78	0.035
α_1	0.052	0.030
σ_v^2	0.043	0.021
σ_{uv}	0.014	0.008

R: The brms package

We focus on the brm function of the brms R package (Bürkner, 2017, 2018). The brms package calls the Stan software (Stan Development Team, 2021) which implements Hamiltonian Monte Carlo (HMC) and no-U-turn samplers (NUTS). We present the simplest possible syntax noting that mixing can be improved via model reparameterization and by specifying various estimation options and we encourage readers to consult the documentation for further details. The syntax is as follows

```
brm(bf(y ~ 1 + x + (1 |s| school),
       sigma ~ 1 + x + (1 |s| school)),
    data = mydata,
    family = gaussian()
)
```

where for further simplicity we use the default priors for all model parameters and random effects. These are normal priors for the regression coefficients, half-Cauchy priors for the random effect standard deviations, and the LKJcorr prior for random effect correlation matrix.

Line 1 of the brm function syntax specifies the mean function. Line 2 specifies the residual variance function, but parameterizes this in terms of the residual SD rather than the residual variance. The desired residual variance function regression coefficients and random effect values can be recovered by multiplying the estimated quantities by 2. Line 3 specifies the dataframe. Line 4 specifies the normal response distribution.

The associated model output is as follows.

```
Compiling Stan program...
Start sampling

SAMPLING FOR MODEL '93a90408567ae7343eea598de7d7e540' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0.005 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 50 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 41.769 seconds (Warm-up)
Chain 1:                19.914 seconds (Sampling)
Chain 1:                61.683 seconds (Total)
Chain 1:

SAMPLING FOR MODEL '93a90408567ae7343eea598de7d7e540' NOW (CHAIN 2).
Chain 2:
Chain 2: Gradient evaluation took 0.001 seconds
Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 10 seconds.
Chain 2: Adjust your expectations accordingly!
Chain 2:
Chain 2:
Chain 2: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 2: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 2: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 2: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 2: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 2: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 2: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 2: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 2: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 2: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 2: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 2: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 2:
Chain 2:  Elapsed Time: 49.423 seconds (Warm-up)
```

```
Chain 2:                21.946 seconds (Sampling)
Chain 2:                71.369 seconds (Total)
Chain 2:

SAMPLING FOR MODEL '93a90408567ae7343eea598de7d7e540' NOW (CHAIN 3).
Chain 3:
Chain 3: Gradient evaluation took 0 seconds
Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 0 seconds.
Chain 3: Adjust your expectations accordingly!
Chain 3:
Chain 3:
Chain 3: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 3: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 3: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 3: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 3: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 3: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 3: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 3: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 3: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 3: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 3: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 3: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 3:
Chain 3:  Elapsed Time: 44.674 seconds (Warm-up)
Chain 3:                18.936 seconds (Sampling)
Chain 3:                63.61 seconds (Total)
Chain 3:

SAMPLING FOR MODEL '93a90408567ae7343eea598de7d7e540' NOW (CHAIN 4).
Chain 4:
Chain 4: Gradient evaluation took 0.001 seconds
Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 10 seconds.
Chain 4: Adjust your expectations accordingly!
Chain 4:
Chain 4:
Chain 4: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 4: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 4: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 4: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 4: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 4: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 4: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 4: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 4: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 4: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 4: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 4: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 4:
Chain 4:  Elapsed Time: 54.744 seconds (Warm-up)
Chain 4:                20.939 seconds (Sampling)
Chain 4:                75.683 seconds (Total)
Chain 4:
 Family: gaussian
  Links: mu = identity; sigma = log
Formula: y ~ 1 + x + (1 | s | school)
         sigma ~ 1 + x + (1 | s | school)
   Data: mydata (Number of observations: 2500)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Group-Level Effects:
~school (Number of levels: 100)
                              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                     0.19      0.02     0.15     0.24 1.00     1552     2474
sd(sigma_Intercept)               0.11      0.02     0.06     0.16 1.01     1333     1396
cor(Intercept,sigma_Intercept)    0.32      0.18    -0.04     0.67 1.00     2180     2381

Population-Level Effects:
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept          -0.03      0.02    -0.07     0.02 1.00     2188     2671
sigma_Intercept    -0.39      0.02    -0.43    -0.36 1.00     3600     3190
x                   0.70      0.01     0.67     0.73 1.00     6686     3234
sigma_x             0.03      0.02    -0.00     0.05 1.00     5863     3170

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
```

```
scale reduction factor on split chains (at convergence, Rhat = 1).
```

The model ran four chains each with 1000 warmup (burn-in) iterations and 1000 monitoring iterations. Recall that the residual variance function is specified in terms of the log of the residual SD, but that the desired regression coefficients can be easily recovered by multiplying the estimated quantities by 2. A second issue is that the elements of the random effect covariance matrix are presented as SDs and correlations rather than as variances and covariances. We can recover the random effect variances by squaring the random effect SDs. The random effect covariance can be recovered by multiplying the random effect correlation by the two random effect SDs. These calculations are best applied to the underlying chains rather than the means which are displayed in the output. Having carried out these steps, the results are as follows.

	Est.	SE
β_0	-0.028	0.024
β_1	0.701	0.014
σ_u^2	0.038	0.008
α_0	-0.784	0.037
α_1	0.050	0.030
σ_v^2	0.054	0.002
σ_{uv}	0.014	0.009

References

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software 80 (1): 1–28.

Bürkner P (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395–411.

Stan Development Team (2021). Stan Modeling Language User's Guide and Reference Manual, Version 2.27.0. URL: http://mc-stan.org.

StataCorp. (2021). Stata 17 Bayesian Analysis Reference Manual. College Station, TX: Stata Press.