The Multilevel Crossed Random Effects Growth Model for Estimating Teacher and School Effects: Issues and Extensions

Gregory J. Palardy

Educational and Psychological Measurement 2010 70: 401 originally published online 28 December 2009
DOI: 10.1177/0013164409355693

Educational and Psychological Measurement
70(3) 401–419
© 2010 SAGE Publications
DOI: 10.1177/0013164409355693
http://epm.sagepub.com

The Multilevel Crossed Random Effects Growth Model for Estimating Teacher and School Effects: Issues and Extensions

Gregory J. Palardy[1]

Abstract

This article examines the multilevel linear crossed random effects growth model for estimating teacher and school effects from repeated measurements of student achievement. Results suggest that even a small degree of unmodeled nonlinearity can result in a substantial upward bias in the magnitude of the teacher effect, which raises concerns about its appropriateness for estimating teacher effects. To address this issue, a piecewise linear crossed random effect growth model is proposed. A comparison with the linear growth form shows that the piecewise specification provides more accurate estimates of teacher effects when achievement growth departs from linear growth across grade levels or over summer, which are prevalent conditions. Fitted examples using nationally representative data and Bayesian estimation methods are provided.

Keywords

crossed random effects, teacher effects, multilevel, growth model, piecewise, summer learning, value added, Bayesian

A typical assumption of multilevel models (MLMs) is that data are strictly hierarchical in structure. That is, each lower level unit is nested in a single unit or group at the next higher level. However, some important applications involve more than one type of grouping within a higher level where units from the immediate lower level are

[1]University of California–Riverside, Riverside, CA, USA

Corresponding Author:
Gregory J. Palardy, 2205 Sproul Hall, Graduate School of Education, University of California–Riverside, Riverside, CA 92521, USA
Email: gregory.palardy@ucr.edu

nested in a single unit of each. For example, students are generally nested in both schools and neighborhoods, but schools are often not nested in a single neighborhood or vice versa. In this case, students may be conceived of as cross-classified in schools and neighborhoods (Garner & Raudenbush, 1991). Cross-classified data are inappropriate for analysis using conventional MLMs. A special class of MLMs, known as crossed random effects models (CREM), has been developed for analyzing data with this structure (Goldstein, 1987; Raudenbush, 1993).

CREMs within the multilevel modeling framework can be considered flexible elaborations of the randomized factorial design that accommodate covariates and unbalanced data and can be extended to include additional levels of analysis. The essential characteristic is that two or more types of groupings or random factors are crossed within a given level of the model. The simplest CREM will have two factors crossed in a two-level model. In this case, each unit at Level 1 is nested in a cell at Level 2 that designates membership to a single unit of each Level 2 random factor. Extending the example above, each student is nested in a single school and a single neighborhood, whereas schools and neighborhoods are crossed random factors at Level 2 of the model. The variation in the respective random effects describes the part of the variance in the outcome attributable to the corresponding factor. Unlike randomized factorial design experiments, the random effects for interactions between factors are not commonly included in the model, because cell sample sizes are typically too small for reliable estimation (Raudenbush, 1993).

A special application of the CREM involves repeated measurements, which will be referred to as the crossed random effects growth model (CREGM). The linear specification of this model (L-CREGM) was proposed by Raudenbush (1993) for estimating teacher effects from repeated measurements of student achievement when students encounter multiple teachers over time. In this situation, students are not nested in a single teacher's classroom; hence, a conventional MLM is not appropriate. However, if the repeated measurements are treated as cross-classified in students and teachers, this can be considered a special case of CREM. This important contribution to the literature on teacher effects estimation was originally constrained to two levels. At Level 1 are repeated measurements of student achievement, which are cross-classified in students and teachers at Level 2. This model was extended by Rowan, Correnti, and Miller (2002) to include a third level—schools—whereas both students and teachers are nested in schools.

Figure 1 depicts a hypothetical database where repeated measurements of achievement test scores are cross-classified in students and teachers. The Xs represent achievement test scores collected on students as they progress through elementary school from kindergarten to the Jth grade, which is the highest grade level in the model. The achievement test scores are nested in N student's and M teacher's classrooms, and the student and teacher factors are crossed. Note that students not only change teachers annually as they progressed through grade levels, but the student composition of the classrooms also likely change.

The L-CREGM has received considerable attention in the literature on teacher effects because it has been suggested that it provides more accurate estimates of

Grade	Kindergarten				First Grade				. . .	J^{th} Grade
Teacher	1	2	3	K	K+1	K+2	K+3	K+L	. . .	M
Student										
1	x			.			x	.	. . .	.
2			x	.	x			.	. . .	.
3		x		.	x			.	. . .	.
4			x	.		x		.	. . .	.
5	x			.			x	.	. . .	.
6		x		.			x	.	. . .	.
7		x		.	x			.	. . .	.
.	.	.	.	.	.	.	.	.	. . .	.
.	.	.	.	.	.	.	.	.	. . .	.
.	.	.	.	.	.	.	.	.	. . .	.
N	.	.	.	.	.	.	.	.	. . .	.

Figure 1. Example of a cross-classified data structure for estimating teacher effects
Note: Each x is an achievement test score. J = highest grade level in the sample; K = number of kindergarten teachers; L = number of first-grade teachers; M = total number of teachers from kindergarten through the Jth grade; N = number of students (Inspired by Raudenbush, 1993, p. 326). Notes: Each x is an achievement test score; J = highest grade level in the sample; K = number of kindergarten teachers; L = number of first grade teachers; M = total number

teacher effects on student learning compared with other models and because estimates using this model indicate teacher effects account for a substantially larger proportion of the variance in student learning compared with earlier modeling approaches (Raudenbush, 1993; Rowan et al., 2002). This model is believed to provide more accurate estimates because student learning trajectories are based on multiple years of data rather than constrained to a single year, as is typically the case when the CREGM is not used. Multiple years of repeated measurements improves the stability of the growth estimates. Yet, the model is based on the assumption that the underlying student achievement growth is linear over time and that deviations in the trajectories from linear are because of variation in the effectiveness of the teachers. However, achievement growth may be nonlinear for other reasons, particularly because of a summer drop in learning that depends in part on the social background of the child (Alexander, Entwisle, & Olsen, 2001; Burkam, Ready, Lee, & Logerfo, 2004). Violations of the linearity assumption can be problematic, because unmodeled nonlinearity can inflate the variance in the random effects—in this case, it can inflate teacher effects (Bauer & Cai, 2009).

To address this issue, this article extends the L-CREGM to the piecewise growth specification model, which is referred to as the P-CREGM. The piecewise growth form is designed for situations where change in the outcome differs during discrete periods, such as across grade levels or during summer, and can address concerns

with unmodeled nonlinearity. The P-CREGM estimates the growth rate for each period or piece of the trajectories and requires a minimum of two repeated measurements per randomly varying period, preferably collected near the beginning and end of the periods.

The remainder of this article is devoted to describing the details of the CREGM and its application for estimating teacher effects. I begin with a brief review of the literature on the estimation of teacher effects. This is followed by a detailed description of the L-CREGM and its extension to the piecewise growth specification. Next, is an empirical analysis comparing the results from the linear and piecewise growth specifications for estimating teacher effects. Last, I discuss the advantages of using the P-CREGM for estimating teacher effects and the implications of the results to building data systems for studying or evaluating teacher effectiveness.

Estimation of Teacher Effects

There is debate in the research community regarding the best modeling approach for estimating teacher effects on student learning. Not surprisingly, the expected magnitude of the teacher effect is linked to the modeling approach employed. A recent review of 17 studies that estimate the magnitude of teacher effects using multilevel achievement gain model or residual gain model (which use prior achievement as a covariate in the model) found that between 7% and 21% of the variance in achievement gain is because of teacher effects and depend partially on the subject area being tested and the grade level of the students (Nye, Konstantopoulos, & Hedges, 2004). However, gain score and residual gain score models may underestimate the magnitude of teacher effects, because these models do not have a residual for measurement error in the achievement tests. In these models, measurement error is subsumed largely by the student level random effect rather than the teacher or school level random effects, which reduces the proportion of the total variance that is attributed to teacher effects.

Growth models overcome the problem of measure error by having an additional lower level of analysis, sometimes referred to as the measurement model, where the parameters of the individual growth trajectories are generated. The measurement model residual term captures the measurement error (and sampling error) of the repeated measurements, which alleviates the concerns of underestimating the teacher effect that has been noted with the gain score and residual gain score models (Willett, 1988). However, a literature on teacher effects based on the standard growth model has not accumulated. This is partially because of data limitations; the model requires that student achievement is measured at least three times during the school year, which is rare. Indeed, no known national database meets this requirement. Data sets with three or more repeated measurements collected over multiple school years and teachers are available; however, in those data sets children are no longer nested in a single teacher's classroom. This dilemma necessitates a more sophisticated approach than the conventional growth model that takes into account the complex structure of the data.

Contrasting with the literature on achievement gains models, studies employing the L-CREGM suggest that teacher effects are larger. For example, using a two-level L-CREGM that controlled for whether teachers had a master's degree, Raudenbush (1993) found that teacher effects accounted for 46% of the variance in student learning in math during early elementary school grades. Although this estimate would be smaller if a school level of analysis and student background variables were added to the model, it would still likely be higher than the upper bounds of the range established by previous research using a gain score or residual gain score outcome. In another example, Rowan et al. (2002) used a large national database and a multilevel L-CREGM (including a school level) to estimate teacher effects, concluding that approximately 60% of the total variance in student reading achievement growth during elementary school is because of teacher effects. This estimate is 3 to 8 times the magnitude established in the literature using gain score outcomes.

Although the L-CREGM is arguably superior to gain score models for estimating teacher effects, it has assumptions that, when violated, can impact parameter estimates. An important assumption of the L-CREGM (and other linear growth models) is that achievement growth is a linear function of time. Violations of this assumption can affect the magnitude of the teacher effect estimate. As we will see, violations of the linearity assumption results in substantially inflated teacher effect estimates using the L-CREGM, whereas it is a far more minor problem with other growth models. This is of particular concern, because student achievement growth may be nonlinear for a variety of reasons. The P-CREGM is suitable for modeling nonlinearity across periods, such as grade levels and summer, to produce more accurate estimates of teacher effects. It also has the advantages of other growth models in that it contains measurement error in the repeated measurements and, like the L-CREGM, is applicable to the multiyear context where students encounter multiple teachers over time, which results in a cross-classified data structure.

Model Descriptions and Formulations

The L-CREGM

The L-CREGM assumes linear change over time, in which case intrastudent growth trajectories can be represented by an intercept and a single slope parameter. Level one of the three level model, which is sometimes referred to as the measurement model, can be described using Equation (1):

$$Y_{t(ij)k} = \pi_{0(ij)k} + \pi_{1(ij)k} a_{t(ij)k} + e_{t(ij)k}, \quad e_{t(ij)k} \sim N(0, \sigma^2). \tag{1}$$

This formulation uses a letter subscript to denote each classification, which also typically corresponds to a level of the model. However, subscripts gathered within parentheses occupy the same level in the hierarchical structure and units from the immediately lower level of analysis are cross-classified in them. $Y_{t(ij)k}$ is the observed value on the achievement outcome variable at time t, for individual i, in classroom j, and school k, where individual and classroom classifications are crossed, both

occupying Level 2. $\pi_{0(ij)k}$ is the expected value of Y for individual i when time is zero, $\pi_{1(ij)k}$ is the expected change in Y per unit change in time for individual i, $a_{t(ij)k}$ is a variable measuring the passage of time between achievement measurements for each individual, and $e_{t(ij)k}$ denotes the residual or random error associated with the repeated achievement measurements for each individual. This measurement model generates the intercepts ($\pi_{0(ij)k}$) and slopes ($\pi_{1(ij)k}$) for each child, which together represents their linear achievement growth trajectory. These random coefficients are the outcomes at Level 2.

Level 2, which includes both student and teacher classifications, is represented by Equation (2):

$$\begin{aligned}\pi_{0(ij)k} &= \beta_{00k} + \sum_{p=1}^{P} \beta_{0pk} X_{pik} + \sum_{q=1}^{Q} \eta_{0qk} Z_{qjk} + r_{0ik} + c_{0jk}, \quad \mathbf{r}_{ik} \sim N(0, \mathbf{T}_r), \\ \pi_{1(ij)k} &= \beta_{10k} + \sum_{p=1}^{P} \beta_{1pk} X_{pik} + \sum_{q=1}^{Q} \eta_{1qk} Z_{qjk} + r_{1ik} + c_{1jk}, \quad \mathbf{c}_{jk} \sim N(0, \mathbf{T}_c),\end{aligned} \tag{2}$$

X_{pik} is one of a set of P covariates measuring student characteristics on which the intercepts and slopes are regressed to explain variation across individuals, and Z_{qjk} is one of a set of Q covariates measuring aspects of the classroom (including aspects of the teacher). In the intercept equation, β_{00k} is the expected value on the outcome when time equals zero and the P student and Q classroom covariates take on the value zero. Similarly, in the slope equation β_{10k} represents the expected change in the outcome per unit change in time within each of k schools controlling for the P and Q covariates. β_{0qk} and β_{1qk} are a set of P slope coefficients that describe the relationship between the student covariates and the respective outcome. η_{0qk} and η_{1qk} are a set of Q slope coefficients for the classroom variables. Both the intercept and slope equations have random effects for students, notated by r_{0ik} and r_{1ik}, and for classrooms, notated by c_{0jk} and c_{1jk}. The random effects are assumed to be normally distributed, to have a mean of zero, and to covary within classification. However, student and classroom random effects are independent.

Note that the random classroom effect for the slope can only reasonably be interpreted as a teacher effect when P and Q covariates are in the model controlling for student and classroom factors that affect student learning but are not within the teachers' control. However, for the remainder of this article, the terms *classroom effects* and *teacher effects* are used interchangeably when describing this random effect.

In the somewhat common situation where achievement is measured on students once annually near the end of the school year, the classroom random effect for the slope is conceptually flawed for estimating teacher effects, because the slope is supposed to measure the change in achievement over time, and there will be little or no variation in the time measure within each classroom, making it infeasible to reliably estimate the classroom slopes. In this situation the classroom random effects (c_{1jk}) and fixed effects ($\sum_{q=1}^{Q} \eta_{1qk} Z_{qjk}$) in the slope equation would be dropped from the model,

and the classroom random effect in the intercept equation (c_{0jk}) would become the classroom effect (Raudenbush, 1993). In this specification, c_{0jk} is conceptualized as the offset or ''deflection'' in the child's growth trajectory resulting from being a member of classroom j, and the variation in c_{0jk} is conceptualized as the variation in achievement growth due to classroom effects. This conceptualization is only reasonable when achievement is measured once annually, near the end of the school year after children have been members of the classroom for the greater part of the school year. If achievement is measured early in the year, much of the observed ''deflection'' in student learning cannot be logically attributed to the present classroom.

Level 3 contains the school classification, for which the equations are shown below:

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + \sum_{m=1}^{M} \gamma_{00m} W_{mk} + u_{0k}, \\ \beta_{10k} &= \gamma_{100} + \sum_{m=1}^{M} \gamma_{10m} W_{mk} + u_{1k}, \end{aligned} \qquad \mathbf{u}_k \sim N(0, \mathbf{T}_u), \tag{3}$$

γ_{000} is the grand mean of the school intercepts, and u_{0k} is the random intercept effect for school k. In the school slope equation, γ_{100} is the grand mean of the learning rates in the sample of schools, and u_{1k} is the unique effect associated with school k. The Ws represent a set of M school-level covariates, each of which has a coefficient associated with it that describes the linear relationship between the respective W and the school intercepts or slopes.

The P-CREGM

The P-CREGM extends the linear specification so that the rate of growth may differ across discrete periods, such as grade levels or during summer. This model is more suitable than the L-CREGM for estimating teacher effects when student achievement growth is not linear across the periods. The equations for the P-CREGM provided below are for a period of 2 school years, including the summer period between them, which is consistent with the application that follows. However, these formulations can easily be generalized. Level 1, the measurement model, can be represented by Equation (4):

$$Y_{t(ij)k} = \pi_{0(ij)k} + \pi_{1(ij)k} a_{1t(ij)k} + \pi_2 a_{2t(ij)k} + \pi_{3(ij)k} a_{3t(ij)k} + e_{t(ij)k}, \; e_{t(ik)j} \sim N(0, \sigma^2). \tag{4}$$

The interpretation of the coefficients will depend on how the time variables are coded. The most readily interpretable coding scheme yields coefficients that estimate the rate of achievement growth during corresponding periods, which is used here and in the application below where details on coding are provided. $\pi_{0(ij)k}$ is the expected achievement for individual i when all three time variables are zero. $\pi_{1(ij)k}$ is the learning rate

during the first school year for individual i, π_2 is the summer learning rate, and $\pi_{3(ij)k}$ is the learning rate during the second school year.

Like the linear specification, the repeated measurements at Level 1 are cross-classified in students and classrooms at Level 2, and the Level 1 coefficients are the outcomes at Level 2. Equation (5) represents Level 2 of the piecewise growth model:

$$\begin{aligned}
\pi_{0(ij)k} &= \beta_{00k} + \sum_{p=1}^{P} \beta_{0pk} X_{pik} + \sum_{q=1}^{Q} \eta_{0qk} Z_{qjk} + r_{0ik} + c_{0jk}, \qquad \mathbf{r}_{ik} \sim N(0, \mathbf{T}_r), \\
\pi_{1(ij)k} &= \beta_{10k} + \sum_{p=1}^{P} \beta_{1pk} X_{pik} + \sum_{q=1}^{Q} \eta_{1qk} Z_{qjk} + r_{1ik} + c_{1jk}, \\
\pi_2 &= \beta_2, \\
\pi_{3(ij)k} &= \beta_{30k} + \sum_{p=1}^{P} \beta_{3pk} X_{pik} + \sum_{q=1}^{Q} \eta_{3qk} Z_{qjk} + r_{3ik} + c_{3jk}. \qquad \mathbf{c}_{jk} \sim N(0, \mathbf{T}_c),
\end{aligned} \tag{5}$$

The intercept and slope outcomes ($\pi_{0(ij)k}$, $\pi_{1(ij)k}$, and $\pi_{3(ij)k}$) have both student (r_{nik}) and classroom (c_{njk}) random effects that are assumed to be normally distributed and to covary within classification. β_{00k} represents the mean student achievement in school k when time equals zero, whereas β_{10k} and β_{30k} represent the mean student learning rates in school k during two different grade levels (e.g., kindergarten and first grade). As we saw with the L-CREGM, the equations include P student predictors and Q classroom predictors that account for variation in the student and classroom random effects.

The individual growth trajectories have $t - 1$ degrees of freedom, where t is the number of repeated measurements taken on the individual. These degrees of freedom correspond to the maximum number of parameters that can be estimated by the intra-individual trajectories, which may limit the complexity of their form. For example, in the application presented below there are four repeated measurements collected over a 2-year period, which limits the individual trajectories to three parameters. Although the linear specification has only two parameters—intercept and slope—the piecewise model typically necessitates additional slope parameters. Four time points provide sufficient data for a random intercept and two random slopes. Additional slopes must be fixed across students or the intercept can be fixed to accommodate an additional random slope. For this reason, the summer slope (β_2) was fixed in Equation (5) and in the example below; however, the summer can be specified as random if additional repeated measurements are available. When there are insufficient degrees of freedom to estimate all parameters of the individual growth trajectories, fixing the summer slope is a reasonable strategy when the objective is to estimate teacher effects, because it is largely a statistical control in the model. Note also that in the event school year learning rates do not differ across grade levels, they can be constrained to be equal. In this case, the P-CREGM simplifies to have one random school year slope,

a specification that will be referred to as the school year linear CREGM (SL-CREGM) in the application below.

Level 3 of the piecewise model, shown in equation (6), contains only the school classification:

$$\begin{aligned}
\beta_{00k} &= \gamma_{000} + \sum_{m=1}^{M} \gamma_{00m} W_{mk} + u_{0k}, \\
\beta_{10k} &= \gamma_{100} + \sum_{m=1}^{M} \gamma_{10m} W_{mk} + u_{1k}, \qquad \mathbf{u}_k \sim N(0, \mathbf{T}_u), \\
\beta_{2} &= \gamma_{2}, \\
\beta_{30k} &= \gamma_{300} + \sum_{m=1}^{M} \gamma_{30m} W_{mk} + u_{3k},
\end{aligned} \tag{6}$$

γ_{000} represents the grand mean of achievement when time is zero, γ_{100} and γ_{300} are the grand mean learning rates for two different grade levels, and γ_2 is the fixed effect for the summer learning rate. u_{0k}, u_{1k}, and u_{3k} are the random effects for schools—each of which may be adjusted by a set of M school-level predictors.

Application

In this section, a national database is used to demonstrate the application of the P-CREGM for estimating fixed and random teacher effects on student achievement growth. To assess the performance of the P-CREGM for estimating teacher effects, the results are compared with two other CREGM specifications including the L-CREGM and the intermediate SL-CREGM, which assumes linear achievement growth across grade levels during the school year and a differential summer growth rate. The data source and the Bayesian analysis methods used are described first, followed by the results.

Data Source

This analysis used a sample of children from the Early Childhood Longitudinal Study of the Kindergarten Class of 1998-1999 (ECLS-K) on whom reading achievement test scores were collected at the beginning and end of kindergarten and first grade (National Center for Educational Statistics [NCES], 2002). Students with missing teacher or school IDs, who transferred schools, or who repeated kindergarten were omitted from the analysis. In all, approximately 35% of the children were omitted from the analysis, resulting in a sample of 3,250 children. A weight developed by NCES (C1C4cw0) was applied to the student data in all analyses. A comparison of the weighted sample used in this analysis with the weighted full sample on key demographic variables and achievement test scores indicate that the means and variances of those variables are highly similar (Palardy & Rumberger, 2008). This suggests that the

Table 1. Descriptive Statistics of Test Scores and Certification Variable

Variable	*N*	Minimum	Maximum	Mean	*SD*
Reading, Fall-K IRT score	3,180	10.50	81.26	23.83	9.14
Reading, Spring-K IRT score	3,180	12.01	85.01	34.50	11.18
Reading, Fall-1 IRT score	3,250	12.82	86.63	39.60	12.69
Reading, Spring-1 IRT score	3,250	16.02	88.95	57.29	13.50
Full certification	1,553	0.00	1.00	0.89	0.32

Note: IRT = item response theory.

weighted sample used in this analysis is approximately nationally representative. The sample includes a total of 1,553 teachers consisting of 710 kindergarten teachers during the first year and 924 first-grade teachers during the second year. Note that 81 of the kindergarten teachers also served as first-grade teachers the following year.

Variables

Table 1 provides a list of the variables used in this study and their descriptive statistics. Repeated reading achievement test scores is the outcome. Test items were scaled using item response theory methods, and tests of varying difficulty were constructed and vertically equated, making them appropriate for growth modeling (NCES, 2002). A dummy-coded variable indicating whether the teacher has attained full certification is used in the analyses to compare fixed teacher effects across the models.

Several time variables are used in the measurement model to estimate learning rates. The optimal coding scheme for time variables in a growth model depends on the preferred interpretation of the intercept and slope estimates. In this application, the time variables are coded to produce coefficients that are estimates of the linear rate of change in achievement during the respective segment. Table 2 shows the values on each time variable for two students attending different schools. Note that the repeated measurements were not collected on the same schedule for each child. To achieve the above interpretation, each time variable is coded zero up to the start of the respective segment, then coded in units of month during the segment, and then it remains constant from the last measure of the segment. The exception is the summer time variable for the SL-CREGM.

Bayesian Estimation

Multilevel CREMs have seen only limited application in the research literature. One reason is software to estimate this class of models has only recently been developed. Another factor is that the typically complex error structure and often very small variance components of these models can result in estimation challenges, particularly for maximum likelihood–based estimators when applied to large data sets. Raudenbush (2008) notes that likelihood-based estimators are poorly suited for variance

Table 2. Examples of the Time Variables for Two Students

		P-CREGM			L-CREGM	SL-CREGM	
Student	Measurement Occasion	Kindergarten ($a_{1t(ij)k}$)	Summer (a_{2ti})	First Grade ($a_{1t(ij)k}$)	Linear ($a_{1t(ij)k}$)	Linear ($a_{1t(ij)k}$)	Summer (a_{2ti})
1	1	2.10	0.00	0.00	2.10	2.10	0.00
1	2	9.53	0.00	0.00	9.53	9.53	0.00
1	3	9.53	2.01	0.00	11.54	11.54	11.54
1	4	9.53	2.01	8.58	20.12	20.12	0.00
2	1	1.84	0.00	0.00	1.84	1.84	0.00
2	2	9.43	0.00	0.00	9.43	9.43	0.00
2	3	9.43	3.02	0.00	12.45	12.45	12.45
2	4	9.43	3.02	7.59	20.04	20.04	0.00

Note: P-CREGM = piecewise crossed random effects growth model; L-CREGM = linear crossed random effects growth model; SL-CREGM = school year linear crossed random effects growth model.

component estimation when the within group sample size is small and unbalanced, as is the case with the ECLS data used in this application and as is fairly common with educational data applicable for these models. CREMs have a complex error structures typically involving several variance components some of which tend to be very small in the context of teacher effects modeling. When these characteristics are combined with small and unbalanced within-group samples, maximum likelihood estimators tend to slow drastically and encounter estimation difficulties (Raudenbush, 2008).

These problems with maximum likelihood estimators for CREMs have led to recommendations to use Bayesian estimation via Monte Carlo Markov chain (MCMC) methods (Browne & Draper, 2006; Clayton & Rasbash, 1999; Goldstein, 2003). MCMC estimation is based on the principle that the probability distribution of a parameter can be determined to an arbitrary level of accuracy by repeatedly sampling from it (Draper, 2008). There are several estimation strategies and many diagnostic procedures for evaluating the quality of estimates. Because a detailed overview of Bayesian methods is beyond the scope of this article, interested readers are referred to Gelman, Carlin, Stern, and Rubin (2004) for a comprehensive presentation and to Browne (2003), Draper (2008), and Seltzer, Wong, and Bryk (1996) for multilevel applications.

The MCMC suite of MLwiN software (programming by Browne and Rasbash) was used to estimate the models in this application. Prior distributions were obtained from preliminary likelihood-based estimates. Raftery–Lewis and Brooks–Draper indices were used to determine the chain length necessary to reach a stable point in the determination of the posterior parameter distributions (Browne, 2003). Comparative model fit was assessed using the deviance information criterion (DIC), which was developed specifically for use on complex hierarchical models and Bayesian estimators (Spiegelhalter, Best, Carlin, & van der Linde, 2002).

Table 3. Piecewise, Linear, and School Year Linear Crossed Random Effects Growth Model Results

Parameter	Piecewise			Linear		School Year Linear	
Fixed effects (standard deviations of the posterior distributions)							
Intercept	19.21 (0.57)			13.86 (0.98)		19.03 (0.64)	
Full certification	0.52 (0.54)			0.08 (0.98)		0.24 (0.59)	
Linear slope	—			2.05 (0.07)		1.85 (0.05)	
Full certification	—			0.01 (0.07)		0.00 (0.05)	
Kindergarten slope	1.82 (0.07)			—		—	
Full certification	0.11 (0.07)			—		—	
Summer slope	0.11 (0.07)			—		0.45 (0.01)	
First slope	2.40 (0.09)			—		—	
Full certification	0.16 (0.08)			—		—	
Random effects (percentage of total variance)							
	Intercept	Kindergarten	First	Intercept	Slope	Intercept	Slope
School	15.91 (28)	0.16 (27)	0.21 (23)	21.00 (14)	0.07 (14)	19.00 (21)	0.06 (22)
Classroom	1.58 (03)	0.07 (11)	0.15 (17)	73.92 (48)	0.35 (70)	12.52 (14)	0.13 (48)
Student	39.73 (69)	0.37 (61)	0.54 (60)	57.70 (38)	0.08 (16)	57.48 (65)	0.08 (30)
Residual (Level 1)		16.76		26.66		26.38	
Model fit							
Deviance		72,738		78,714		78,576	
DIC		79,209		83,600		83,069	

Note: DIC = Deviance Information Criterion.

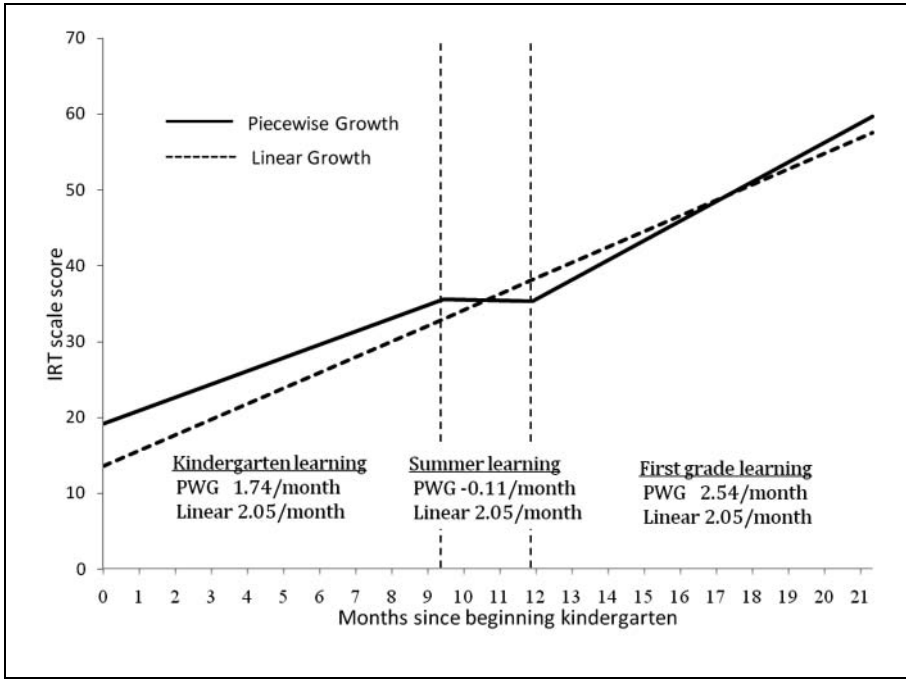

Figure 2. Piecewise and linear growth trajectories for reading in kindergarten through first grade

Results

The results of the fixed teacher effects are presented first followed by the random effects and model fit, all of which are summarized in Table 3. At each stage, the P-CREGM results are compared with the L- and SL-CREGM results to assess the strengths of the P-CREGM for estimating teacher effects. Note that this analysis began with the unconditional specification of each model, which is the typical starting point for multilevel analyses. However, in the subsequent model, which differs from the unconditional model by a single predictor—Teacher Certification—there were only miniscule changes in the variance components. Therefore, to conserve space, only the teacher certification model results are presented.

A visual comparison of the mean reading achievement growth trajectories for the piecewise and linear growth forms are shown in Figure 2. The figure illustrates that the rate of achievement growth differs across segments and that the linear trajectory seems to be a poor fit to the data. This suggests the assumption of linearity for the L-CREGM has not been met. An examination of the residual plot (not shown) verified this observation, although the violation appears minor.

Fixed Effects

The P-CREGM fixed effects results show that mean reading achievement growth fluctuates substantially over the three segments. During kindergarten children learn, on average, at a rate of 1.82 units per month, which then drops to −0.11 units per month during the summer, followed by 2.40 during first grade. Note that the first grade rate is 32% higher than the kindergarten rate. That the rate is markedly lower during summer compared with when school is in session is expected, given that most students are not receiving formal reading instruction during summer. The higher rate during first grade compared with kindergarten is also expected, because children typically attend school half of the day during kindergarten compared with the full day during first grade when reading instruction tends to receive a greater emphasis. Finally, Full Teacher Certification has a significant association with reading achievement growth in first grade but not kindergarten. This differential effect may also be because of the longer school day and greater emphasis on reading instruction in first grade.

The mean learning rate for the L-CREGM is 2.05 units per month compared with an average monthly rate of 1.85 for the P-CREGM. The higher average learning for the L-CREGM is compensated for with a lower intercept, which at 13.82 is several standard deviations of the posterior distribution below the P-CREGM intercept of 19.21. Moreover, the L-CREGM learning rate is dissimilar to either the kindergarten or the first-grade slope. Finally, Full Teacher Certification is not associated with linear growth in reading achievement. The coefficient is the approximate average of the kindergarten and first grade coefficients from the P-CREGM, which in this case cancel each other out. These results emphasize the insensitivity of the L-CREGM to differences in fixed effects across the segments of the trajectory, which can result in biased estimates for any given segment.

The SL-CREGM can be considered an intermediate model between the P- and L-CREGM. It assumes that children learn at approximately equal rates during school but at different rates during summer. To promote a direct comparison with the P- and L-CREGM, the school year slope is specified as random, and the summer slope is fixed. The results suggest that the SL-CREGM is only a minor improvement over the L-CREGM for these data for accurately estimating kindergarten and first-grade achievement growth. Although the SL-CREGM intercept is very similar to the P-CREGM intercept (19.21 vs. 19.03), the SL-CREGM school year slope (1.85) is far below the average school year growth rate for P-CREGM (2.11) and provides a particularly poor estimate of the first-grade learning rate (2.40). Moreover, the Full Certification coefficient is nonsignificant.

Random Effects and Model Fit

Substantial differences in the P- and L-CREGM random teacher effects were also noted. Comparisons are based on the percentage of the total variance in the trajectories between teachers within schools, because it can be compared directly with previous research on teacher effects. The P-CREGM results show that 3% of the

variance in the intercept is between teachers within schools, whereas 70% is between students within schools, and 27% is between schools. Recall that the intercept is the estimated mean reading achievement when each of the time variables equals zero, which corresponds to the start of kindergarten. This sparse variation in initial achievement among teachers is expected because ability tracking is rarely practiced in kindergarten. Contrasting these findings, 48% of the total variance in the intercept is between teachers within schools for the L-CREGM, whereas 38% is between students within schools, and 14% is between schools. These results are implausible, unless there is a high degree of ability tracking at the start of kindergarten and are the consequence of a poor fit of the L-CREGM to the repeated measurements that Figure 2 indicates is not linear. The SL-CREGM results suggest that adjusting the school year trajectory for the summer decline helps correct this problem as only 14% of the total variance in the SL-CREGM intercept is between classrooms.

The percentage of reading achievement growth between classrooms within schools for the P-CREGM was 11% and 17% for kindergarten and first grade respectively, compared with 70% for the L-CREGM and 48% for the SL-CREGM. These results suggest that nonlinearity across the segments not only biases the fixed effects for L-CREGM and SL-CREGM but also inflates the random teacher effects. The degree of the inflation is enormous in this case—more than 6 times for kindergarten and more than 4 times for first grade—which may strike some as surprising, given that the violation in the linearity assumption is only moderate for the school year slopes. This suggests that when left unmodeled, even relatively minor nonlinearity in the growth trajectories can have substantial inflationary consequences on the magnitude of the random teacher effects.

The P-CREGM also fits the data better than the L-CREGM or the SL-CREGM. The DIC for the P-CREGM is 79205 compared with 83602 and 83069 for the L-CREGM and the SL-CREGM. According to Spiegelhalter et al. (2002), DIC differences greater than 7 are substantial, whereas the differences here are in the magnitude of a 3,000. The fit of the Level 1 model can be assessed using the variance in the Level 1 residuals, which measures variation of the observed repeated achievement measurements about their respective intrachild trajectories. Misspecification of the form of the growth trajectory will result in greater residual variation. The Level 1 residual variance of the P-CREGM (16.76) is approximately 27% less than the L-CREGM (26.66) or the SL-CREGM (26.38).

Discussion

Unmodeled Nonlinearity Inflates Teacher Effects

The results suggest that ignoring even relatively minor departures from linearity in the repeated measurements across segments can substantially inflate the L-CREGM teacher effect estimates. The explanation for this phenomenon is a little complicated. The variance in the teacher random effect (in Equation (5)) measures the

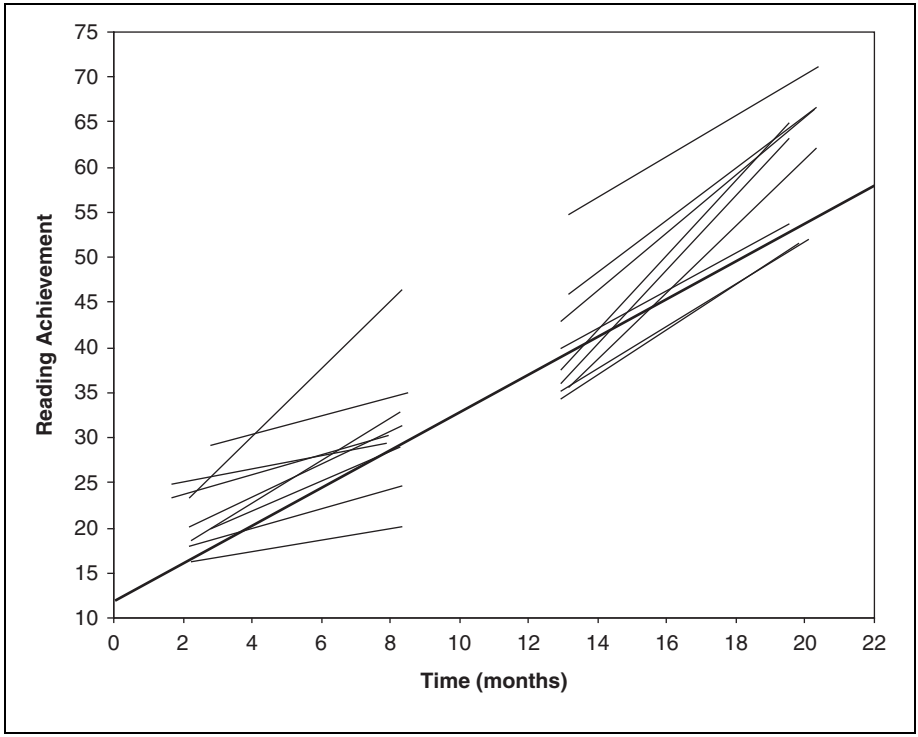

Figure 3. Kindergarten and first-grade classroom trajectories about the mean school trajectory based on the L-CREGM

deviation in the classroom achievement growth rates from the respective school rate. In this application, the classroom slopes for each school year are estimated from achievement data collected at two points only—the beginning and end of each school year—whereas the school trajectories use all four repeated measurements. Therefore, the L-CREGM estimates the linear trajectory for classrooms within each year as would the piecewise model, whereas the school trajectories are linear across the 2 years. Figure 3, which shows the kindergarten and first-grade classroom trajectories in a single school and the mean school trajectory, illustrates this phenomenon. The classroom achievement growth rates are not linear across segments but, rather, systematically lower in kindergarten compared with first grade. As a result, when the L-CREGM is used on these data rather than the better fitting P-CREGM, the variation of the classroom slopes about the mean school trajectory is overestimated. Note that a similar problem will inflate the teacher effect estimate if nonlinear data is fitted to the L-CREGM when only one achievement test score is collected per year, which is the data condition introduced by Raudenbush (1993). In that case, the distribution of classroom means will not be centered on

the respective mean school trajectories, and thus, the average teacher residual will be larger than for the piecewise model.

This rationale for why the L-CREGM inflates the teacher effect is supported by the findings of a recent study on the consequence of violations of the linearity assumption. Bauer and Cai (2009) found that random slope variation increases with the degree of unmodeled curvature in the association between the covariate and outcome, but only when the average value of the covariate varied across Level 2 units. In the context of a growth model, this translates to the mean of the time variable varying across Level 2 units. Hence, growth models tend to be immune to this problem because repeated measurements typically cover similar ranges on the time variable and tend to be somewhat balanced on time across individuals. However, for the CREGM, the means on the time variable will vary considerably across classrooms. For example, the kindergarten classrooms will have lower mean values on the time variable than the first-grade classrooms. Bauer and Cai also note that the problem is more acute when the number of Level 2 units is small, which will generally be the case with teacher effects estimates, because typically only a small number of classrooms are sampled per school.

Unmodeled Summer Learning Can Bias Teacher Effects

The summer learning differential is a concern not only because the learning rate is expected to drop, creating nonlinearity that may inflate the teacher effect estimates, but also because failing to control for the summer can result in inequitable estimates of teacher performance. This problem arises because summer learning is more closely linked to student demographics than school year learning (Alexander et al., 2001; Burkam et al., 2004). Hence, failing to partition the summer period from school year learning may introduce a negative bias against teachers serving disadvantaged students who tend to learn at similar rates as their affluent peers during the school year but lag behind during summer. Although adding covariates to the model that control for student background characteristics will reduce the magnitude of this bias, controlling for the summer period provides a much more comprehensive protection against this threat to the internal validity of the teacher effect estimates.

Implications of Results to Data for Estimating Teacher Effects

The findings of this study have implications for future efforts to build data systems for studying or evaluating teacher effectiveness. The results suggest that at least two annual measurements of student achievement are necessary for reasonably accurate estimates of teacher effects. Two annual measurements are the minimum number to estimate the P-CREGM or to control for the summer period. Although two is the minimum, additional annual measurements will provide greater modeling flexibility and more information for studying teacher effects. For example, the P-CREGM with random summer effects and the P-CREGM with nonlinear growth within grade levels both require more than two repeated measurements annually.

Summary and Conclusions

The L-CREGM has received considerable attention for estimating teacher effects. However, the results of this study suggest the model overestimates teacher effects when even minor nonlinearly in student growth trajectories is left unmodeled. This raises concerns about the appropriateness of the model for estimating teacher effects because some degree of nonlinearity in students learning trajectories apart from the teacher effect is anticipated across grade levels. Although most growth models are immune to this problem because the time variable is balanced across higher level units (Bauer & Cai, 2009), the L-CREGM is not. The problem is amplified when within-group samples (i.e., the number of teachers within each school) are small, which is prevalent in databases for estimating teacher effects. The P-CREGM is proposed for modeling nonlinearity when there are discrete periods (i.e., grade levels) and typically a very limited number of repeated measurements per period restricting the complexity of the form of the growth curve within segments. The results suggest the P-CREGM performs well in this situation and thus is recommended as a replacement for the L-CREGM for estimating teacher effects. The results also suggest the accurate estimation of teacher effects requires a minimum of two repeated achievement measures per school year.

Acknowledgments

The author is grateful to Stephen Olejnik, Robert Hanneman, Myung Hwa Koh, and three anonymous reviewers for their helpful comments.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

Funding

This research was supported by a grant from the American Educational Research Association which receives funds for its AERA Grants Program from the National Science Foundation and the National Center for Education Statistics of the Institute of Education Sciences (U.S. Department of Education) under NSF Grant REC-0310268. Opinions reflect those of the authors and do not necessarily reflect those of the granting agencies.

References

Alexander, K. L., Entwisle, D. R., & Olsen, H. R. (2001). Schools, achievement, and inequality: A seasonal perspective. *Educational Evaluation and Policy Analysis, 23*, 171-191.

Bauer, D. J., & Cai, L. (2009). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics, 34*, 97-114.

Browne, W. J. (2003). *MCMC estimation in MlwiN, Version 2.0*. London: Centre for Multilevel Modeling.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1*, 473-549.

Burkam, D. T., Ready, D. D., Lee, V. E., & Logerfo, L. F. (2004). Social-class differences in summer learning between kindergarten and first grade: Model specification and estimation. *Sociology of Education, 77*, 1-31.

Clayton, D., & Rasbash, J. (1999). Estimation in large crossed random effect models by data augmentation. *Journal of the Royal Statistical Society: Series A, 162*, 425-436.

Draper, D. (2008). Bayesian multilevel analysis and MCMC. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 77-140). New York: Springer.

Garner, C. L., & Raudenbush, S. W. (1991). Neighborhood effects on educational attainment: A multilevel analysis. *Sociology of Education, 64*, 251-262.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika, 74*, 430-31.

Goldstein, H. (2003). *Multilevel statistical model* (3rd ed.). London: Edward Arnold.

National Center for Educational Statistics. (2002). *User's guide to the longitudinal kindergarten-first grade public-use data file*. Washington, DC: U.S. Department of Education.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237-257.

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in the first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis, 30*, 111-140.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational and Behavioral Statistics, 18*, 321-350.

Raudenbush, S. W. (2008). Many small groups. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 207-236). New York: Springer.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*, 1525-1567.

Seltzer, M. H., Wong, W. H., & Bryk, A. S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics, 21*, 131-167.

Spiegelhalter, D. J., Best, N. G., Carlin B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B, 64*, 583-640.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education, 15*, 345-422.