Marco A. Muñoz, Joseph R. Prather,
James H. Stronge

Planning and Changing
Vol. 42, No. 3/4, 2011, pp. 241–273

EXPLORING TEACHER EFFECTIVENESS USING HIERARCHICAL LINEAR MODELS: STUDENT- AND CLASSROOM-LEVEL PREDICTORS AND CROSS-YEAR STABILITY IN ELEMENTARY SCHOOL READING

Teacher effectiveness and evaluation using student growth measures is a popular reform strategy in education. Teachers can make a difference in student academic growth, but a question that begs an answer is how to go about measuring this impact. This study examines models of teacher effectiveness and the development of hierarchical linear models (HLM) using fourth grade end-of-year state accountability reading test scores as the outcome variable. An extensive review of literature was conducted to assess the use of HLM in educational settings, particularly as related to teacher effectiveness analyses. Although multiple student variables were explored, previous reading test scores was the most significant student-level variable while teachers' years of experience was used as a classroom-level variable. This model produced a classroom effectiveness index that was notably consistent across three years of data for the same teachers. Implications for policy, practice, and research are discussed.

Teacher effectiveness is an important area of investigation that has emerged in recent years among educational researchers. A growing body of research has shown that teacher effectiveness is a strong predictor of student achievement (Darling-Hammond, 1996; Darling-Hammond, 2000; Hanushek & Lindseth, 2009; Munoz & Chang, 2007; Nye, Konstantopoulos, & Hedges, 2004; Sanders & Rivers, 1996; Stronge, Ward, Tucker, & Hindman, 2008). Yet, teachers receive little formative or summative feedback on their teaching activities; in general, teacher evaluation is a compliance exercise based on non-informative checklists and with the common conclusive rating of "satisfactory" (Weisberg, Sexton, Mulhern, & Keeling, 2009). Furthermore, research indicates that schools serving minority, low-income students have the most difficulty recruiting and retaining effective teachers (Darling-Hammond, 2000). This disparity in teacher effectiveness between schools and districts contributes to the student achievement gap.

In light of these findings, relatively recent educational policies such as *No Child Left Behind* and *Race to the Top* have attempted to address teacher effectiveness at the policy level. The authorization in 2001 of Title 1 of the Elementary and Secondary Act, commonly referred to as *No Child Left Behind* (NCLB), specified that teachers must be "highly qualified" to ensure that all students learn and demonstrate academic proficiency. To be considered "highly qualified," a teacher of a core academic subject is required to

hold a bachelor's degree, have full state certification or licensure, and demonstrate subject matter competence (measured using PRAXIS tests). This focus from NCLB for determining the quality of a teacher is based on teacher preparation—the qualifications or inputs from teachers' training.

This more input-oriented concept of teacher effectiveness was expanded at the national level with *The American Recovery and Reinvestment Act of 2009* to focus on teacher effectiveness in terms of outcomes or what teachers are able to do to improve student achievement. The principle upon which *Race to the Top* (RTTT) is founded calls for teacher effectiveness to be determined from a combination of measures using both students' growth indicators and observation-based assessments. Student growth is based on the change in student achievement for an individual student between two or more points in time rather than on proficiency data. The emphasis placed on student outcomes has served to link teacher effectiveness and quality with teacher evaluations.

In a way, teacher effectiveness has been equalized to student achievement (Stronge, 2010). In research that supports this statement, Sanders and Rivers (1996) found an enormous gap in the achievement levels of students that had three consecutive years of teachers rated as "high" compared to those students that had three consecutive years of teachers rated as "low." According to Stronge and Tucker (2000), "there are distinctive qualities that epitomize good teachers—and one of those qualities is the ability to make a difference in students' lives" (p. 1). When it comes to definitions, Jerald (2003) stated that "we must define good teaching by results, not by personal characteristics or our preconceived notions. When the goal is student learning, seeming to be a good teacher and actually being a good teacher can be very different" (p. 13).

The overarching goal of this research study is to develop a statistical model that will identify those teachers that are "actually being a good teacher." In order to find those "good teachers," the models to be examined for this study are based on "value added" models. The value-added approach is based on the growth a student makes from the time of entering a classroom to the time of their exiting that classroom. This approach is different than most accountability models that examine just the end-of-year achievement scores and do not take into account students' backgrounds, students' prior achievement, or students' effect on each other. Of equal importance, this study looks into value-added fluctuations from year to year as succeeding cohorts of students move through their classrooms. Consistency of effectiveness measures continues to be debated.

Doran (2003) points out that most accountability systems, which are not value-added, are basically (a) invalid or misleading since student achievement scores are affected by external variables that are outside the influence of school and teachers, (b) they fail to recognize the growth of the student since entering the accountable environment, (c) they do not take into account the cumulative effect of prior learning, and (d) they use cut-score

categories that mismeasure academic performance. The strengths of a value-added model have lead Thum and Bryk (1997) to state that "from a purely technical perspective, the arguments seem very clear: Anything other than a value-added-based approach is simply not defensible" (p. 102).

Other authors disagree (Kupermintz, 2002; McCaffrey, Lockwood, Koretz, & Hamilton, 2003). According to Kupermintz (2002) to label a teacher as effective based on the gains of their students is logically faulty for several reasons: (a) teachers with fewer students have less accurate data and their estimates are more likely to be "pulled" toward the district average; (b) there are potentially conflicting explanations, instead of teacher effectiveness, for the performance of students on tests; (c) value-added models do not adequately separate the relationship between teacher characteristics and student characteristics and the observed gains; and, (d) teachers assigned higher performing students are more likely to become labeled as effective, while teachers assigned high-risk students are more likely to be labeled as ineffective. Kupermintz (2002) stated that possible influences on student learning include: "personal propensities and resources (both cognitive and noncognitive), physical and mental maturation, home environment, cultural heritage, institutional and informal community resources" (p.294). McCaffrey et al. (2003) concluded that "the existing research base on VAM [value-added models] suggests that more work is needed before the techniques can be used to support important decisions about teachers or schools" (p. 111).

Hierarchical Linear Models and Value-Added Education

Pedhazur (1997) stated that multilevel analysis uses information from all available levels (e.g., students, classrooms, schools), making it possible to learn how variables at one level affect relations among variables at another level. Moreover, multilevel analysis affords estimation of variance between groups as distinct from variance within groups. Pedhazur further stated that multilevel models "yield more realistic standard errors" (p.692) than Ordinary Least Squares (OLS) estimates.

A number of examples of educational uses of multilevel methods are presented below. For the most part, the aim of the researchers who performed these studies has been to estimate student data and student achievement on school effects (Marsh, Hau, & Kong, 2002; Pituch, 1999), teacher and school characteristics on student achievement (Bankston & Caldas, 2000; Berends, 2000; Guthrie, 2001; Heck & Crislip, 2001; Swanson & Stevenson, 2002), school improvement over time (Mandeville, 1988; Mandeville & Anderson, 1987), student achievement in a specific discipline (Carbonaro & Gamoran, 2002; Lee & Bryk, 1989; Raudenbush, Fotiu, & Cheong, 1998; Wilkins & Ma, 2002), and specific strategies for student achievement (Burns & Mason, 2002; Desimone, Porter, Garet, Yoon, & Birman, 2002; Goldstein, Yang, Omar, Turner, & Thompson, 2000).

Until recently, school effectiveness research has involved an examination of the variables at either the school level (i.e., aggregated from all the students in that school but failing to account for individual effects) or the individual level (i.e., analyzing data at the individual student level but failing to account for group effects). Since the early 1990s, multilevel statistical models, such as Hierarchical Linear Modeling (HLM) and mixed-model statistical analysis, have been gaining popularity in educational research as a means to examine the effects of different levels of grouping on student achievement (e.g., classroom, schools, and districts).

Over the past two decades HLM has emerged as a well-accepted statistical model to use when conducting a study of school effects within an educational setting. One reason for this growing acceptance is that HLM allows for the effects of the context to be taken into account. Moreover, HLM offers several advantages over other methodologies (Lee, 2000) and solves several difficulties with unit of analysis. Prior to multi-level analysis, researchers attempted to find statistical relationships between school factors and variables measured at the student level. The researchers then had to determine which level of analysis was appropriate. Lee stated that the researchers had to choose between the level where the intervention or effect was administered (school level) or the level where the intervention or effect was believed to occur (student level).

Lee (2000) pointed out that there are three problems when using a single level method, like OLS multiple regression and analysis of variance (ANOVA): (a) aggregation bias, (b) misestimated standard errors, and (c) heterogeneity of regression. The first difficulty is the aggregation bias that can occur when a variable takes on different effects at diverse levels of aggregation. A second difficulty concerns the estimation of the standard errors used for statistical testing; for example, with multilevel data, misestimated standard errors can occur when researchers treat individual cases as though they are independent (a standard assumption of OLS regression) when they are not. A third difficulty concerns heterogeneity of regression slopes, which means that relations between characteristics of students and academic achievement may vary across schools and may be a function of group level variables.

To a substantial extent, HLM solves the problems of aggregation bias, misestimated standard errors, and heterogeneity of regression. First, the problem of aggregation bias is solved since HLM allows for the examination of the data at more than one level of aggregation. Second, the problem of misestimated standard error is avoided since the independence of cases is not an assumption of HLM. Finally, the problem of heterogeneity of regression is solved by HLM since multi-level procedure allows for the investigation of grouping effects.

There is a history of prior application of HLM in school systems. Over the last two decades, the State of Tennessee has adopted and extensively used a value-added model similar to HLM. At the local district

level, the Dallas Texas Public Schools was an early adopter of hierarchical modeling procedures to analyze student data aggregated at the district, school, and teacher levels. The shift in attention from assessing current level of performance to showing progress in learning has refined the way in which policy makers conceptualize educational outcomes.

The Tennessee Value-Added Assessment System (TVAAS) is a statistical model developed to obtain unbiased estimates of the effect of teachers on the academic gains of students (Sanders, 2000). By using a student's previous academic history, the students serve as their own control for extraneous factors. Sanders and Horn (1995) stated that "TVAAS was developed on the premise that society has a right to expect that schools will provide students with the opportunity for academic growth regardless of the level at which the students enter the educational venue. In other words, all students can and should learn commensurate with their abilities" (p. 12). Holland (2001) stated that value-added models, such as TVAAS, can be used by decision-makers to evaluate teachers, the latest innovative curriculum, and teacher preparation programs.

The Dallas Texas Public Schools have used a value-added accountability system for more than two decades (Webster & Mendro, 1997). The current system in Dallas combines the use of multiple regression and hierarchical linear modeling. The Dallas accountability model controls for many preexisting student differences in ethnicity, gender, language proficiency, and socioeconomic status (termed fairness variables).

Purpose of the Study

The purpose of this study is to measure teacher effectiveness using a value-added methodology—namely hierarchical linear modeling. Under this conceptualization, teacher effectiveness is operationally defined as the teacher's impact in reading on statewide assessment. Since school data are an excellent fit with HLM, this study will examine an urban school district's data using multilevel models to determine classroom effectiveness. The study used a multilevel model to control for selected students' demographic effects, previous academic achievement attainment, teacher characteristics, and classroom demographics to obtain a measure of teacher effectiveness. It is a two-level model with individual student characteristics serving as Level 1 variables and classroom characteristics serving as Level 2 variables.

More specifically, the following research questions were addressed: (a) Is there enough classroom variance (as measured by the intra-class correlation) to justify the use of HLM? (b) What student level variables are significant predictors of student achievement as measured by their reading scores? (c) What classroom level variables are significant predictors of student achievement as measured by their reading scores? (d) When combining both students and classroom level variables, what are

significant predictors of student achievement? (e) How consistent are the ratings of classroom level effectiveness over a three-year period of time?

This research examined the usefulness of multi-level models when applied to teacher evaluation or classroom effectiveness. More specifically, the first major objective was to develop an HLM model to identify effective and ineffective classrooms using reading scores. The second major objective was to test the consistency of the teacher scores over a three-year period; in this regard, this study looks into value-added yearly fluctuations as succeeding cohorts of students move through their classrooms.

In this age of accountability, schools are searching for effective methods to identify classrooms that consistently "add value" to a student's education. It is hoped the study will add to the growing body of literature on the use of multilevel models and evaluating classroom level effectiveness.

Methods

The context for the study was elementary schools in the Jefferson County Public Schools (JCPS) in Louisville, Kentucky. The district is located in a large metropolitan area and has about 152 schools serving approximately 100,000 students. JCPS educates a high percentage of at-risk urban students. The district has a student assignment plan based on managed choice, which facilitates the racial desegregation of its schools by providing students with transportation. An additional contextual factor to consider is that School-Based Decision Making (SBDM) is a model employed for setting school policy consistent with district board policy to enhance students' achievement.

Participants

Reading scores of fourth grade elementary school students were analyzed from all elementary schools in JCPS. The data spanned three consecutive years (2001–2003). Two criteria were used when evaluating the data set prior to the analysis. First, the student must have been enrolled in the school system for a minimum of 100 school days (as regulated on the state's accountability system).[a] Second, only classrooms with at least 15 students were included in the analyses (Kreft & De Leeuw, 1998).[b] In total, 2,955 student records were removed from the three-year analyses. As a result of the selection criteria, the number of students included were 5,837 (Year 1), 5,645 (Year 2), and 5,724 (Year 3), the numbers of teachers were 241 (Year 1), 235 (Year 2), and 236 (Year 3), drawn from 81 elementary schools. It should be noted which students tended to be eliminated due to the smaller class sizes (i.e., 15 students per classroom): (a) special needs students, (b) English as a second language (ESL), and (c) gifted and talented.

Instrumentation

The student level variables examined included the following: gender, race, socio-economic status (measured as free/reduced lunch and median income by census tract), parents' education (median education level by census tract), attendance, age, Comprehensive Test of Basic Skills (CTBS) reading (administered in the Spring of the prior school year), and classification (special needs, English as a second language, or gifted and talented). Table 1 describes the Level 1, student-related variables.

Table 1

Level 1 Individual Student Variables

Variables	Description
KCCT reading scale score	KCCT reading scale score (dependent variable)
Student total absences	Number of days absent during the school year
Student days membership	Number of days enrolled during the school year
Parents' education index	0 = completion of grades 0–8, 1 = grades 9–12 no diploma, 2 = diploma or equivalent, 3 = college no degree, 4 = associate degree, 5 = bachelor degree, 6 = master's degree, 7 = professional or doctoral degree
Parents' median education	Census tract median education
Family median income	Census tract median income
CTBS reading scale score	CTBS reading scale score
Student percent attendance	(Days membership – Days absent) / Days membership x 100
Student days old	Number of days old on May 1 of testing year
Student African American	1 = African American, 0 = non-African American
Student White	1 = White, 0 = non-White
Student other	1 = other, 0 = White or African American
Student female	1 = female, 0 = male
Student free/reduced lunch	1 = free/reduced lunch, 0 = full price lunch
Student ESL	1 = ESL student, 0 = non-ESL student
Student special needs	1 = special needs student, 0 = non-special needs
Student gifted	1 = gifted student, 0 = non-gifted student

The teacher/classroom level variables that were examined included the following: classroom aggregated data of selected student level variables, number of students in classroom, teacher's years of experience, teacher's

level of education, and teacher's rank. Table 2 describes the Level 2, classroom-related variables.

Table 2

Level 2 Teacher Variables

Variables	Description
Teacher experience in schools	Number of years experience teaching within the school district
Class size	Number of students in the classroom
Teacher degree	4 = bachelor, 5 = 5th year program, 6 = 6th year program, 7 = master's, 8 = master's degree plus 30 graduate hours, 9 = doctorate
Teacher female	1 = female, 0 = male
Teacher White	1 = White, 0 = non-White
Teacher African American	1 = African American, 0 = non-African American
Teacher other	1 = other, 0 = non-African American or non-White
Teacher rank	5 = doctorate, 10 = rank I (master's degree plus 30 graduate hours), 15 = master's +15 graduate credit hours, 20 = rank II (master's degree), 25 = bachelor's +15 graduate credit hours, 30 = rank II (bachelor's degree), 40 = emergency certified

Table 3 reports the Level 1, student-related variables with the number of participants, means, and standard deviations for each year of data. It includes three types of data: (a) demographic (e.g., gender, race), (b) academic (e.g., reading test scores), and (c) non-academic (e.g., attendance).

Table 3

Level 1 Descriptive Statistics (Individual Student Variables)

	Year 1			Year 2			Year 3		
Variable	*N*	Mean	*SD*	*N*	Mean	*SD*	*N*	Mean	*SD*
KCCT reading scale score	5815	541.43	40.78	5560	544.83	36.37	5679	544.63	36.22
Student total absences	5822	7.33	6.93	5560	6.90	6.27	5679	7.64	7.04
Student days membership	5822	174.28	5.31	5560	174.23	6.03	5679	174.33	6.28
Education index	5822	2.70	.75	5560	2.70	.75	5679	2.68	.73
Median education	5822	2.56	.89	5560	2.56	.88	5679	2.52	.86
Median income	5822	39,456.81	18,618.05	5560	40,043.23	18,566.67	5679	39,320.80	18,243.30
CTBS reading scale score	5439	629.04	43.90	5232	631.77	44.31	5339	632.68	44.04
Student percent attendance	5822	.96	.04	5560	.96	.04	5679	.96	.04
Student days old	5821	3,741.99	163.63	5560	3,742.58	165.48	5679	3,739.74	166.45
Student African American	5822	.36	.48	5560	.36	.48	5679	.36	.48
Student White	5822	.60	.49	5560	.60	.49	5679	.58	.49
Student other	5822	.05	.21	5560	.04	.20	5679	.06	.23
Student female	5822	.50	.50	5560	.50	.50	5679	.49	.49
Student free/reduced lunch	5822	.53	.50	5560	.53	.50	5679	.57	.50
Student ESL	5822	.01	.12	5560	.01	.08	5679	.01	.11
Student special needs	5822	.11	.32	5560	.12	.32	5679	.12	.32
Student gifted	5822	.06	.23	5560	.07	.26	5679	.07	.25

Table 4 reports the Level 2, teacher-associated variables associated with the number of classes, means, and standard deviations for each year of data. The focus is on variables related to teachers, such as experience, class size, educational degree, gender, race, and rank. In the school district under study, teachers have a designated rank depending on the level of education (e.g., bachelors, masters, and beyond masters).

Table 4

Level 2 Descriptive Statistics (Teacher Variables)

	Year 1			Year 2			Year 3		
Variable	*N*	Mean	*SD*	*N*	Mean	*SD*	*N*	Mean	*SD*
Teacher experience in school district	241	9.91	9.52	235	9.68	9.10	236	9.59	8.52
Class size	241	25.50	4.13	235	25.22	4.40	236	25.25	4.90
Teacher degree	241	6.10	1.37	235	6.09	1.36	236	6.03	1.37
Teacher female	241	.94	.24	235	.93	.26	236	.90	.30
Teacher White	241	.80	.40	235	.81	.39	236	.81	.39
Teacher African American	241	.18	.38	235	.17	.38	236	.17	.37
Teacher other	241	.02	.13	235	.02	.13	236	.03	.16
Teacher rank	241	20.77	6.12	235	20.60	6.19	236	20.68	5.96

Table 5 reports the Level 2, aggregated-variables associated with the number of classrooms, means, and standard deviations for each year of data. The student level variable that was used as the dependent variable was the state's reading test scores. This spring test is administered to fourth graders and reported using scale scores and proficiency levels. For this study, only scale scores were used.

Table 5

Level 2 Descriptive Statistics (Classroom Aggregate Variables)

	Year 1			Year 2			Year 3		
Variable	*N*	Mean	*SD*	*N*	Mean	*SD*	*N*	Mean	*SD*
Student total absences	241	9.57	8.92	235	6.39	6.03	236	7.11	6.96
Student days membership	241	174.17	6.39	235	174.89	.98	236	174.94	2.73
Education index	241	2.59	.76	235	3.17	.82	236	3.24	.80
Median education	241	2.48	.90	235	3.06	1.16	236	3.16	1.14
Median income	241	56,455.80	20,134.88	235	58,570.99	21,596.38	236	59,807.48	21,407.42
CTBS reading scale score	241	604.98	42.52	235	637.07	50.08	236	634.83	47.49
Student percent attendance	241	.94	.05	235	.96	.03	236	.96	.04
Student days old	241	4,090.11	133.84	235	3,764.24	175.48	236	3,729.74	177.15
Student African American	241	.49	.50	235	.14	.34	236	.17	.37
Student White	241	.48	.50	235	.79	.41	236	.77	.42
Student other	241	.03	.18	235	.07	.26	236	.07	.25
Student female	241	.38	.49	235	.53	.50	236	.42	.49
Student free/reduced lunch	241	.66	.48	235	.62	.49	236	.61	.49
Student ESL	241	.01	.09	235	.01	.09	236	.00	.07
Student special needs	241	.31	.46	235	.14	.34	236	.11	.32
Student gifted	241	.02	.13	235	.08	.27	236	.07	.26

Design and Procedures

Raudenbush's and Bryk's (2002) methodological approach was used as a guide to the model development. This analysis was completed in five stages: (a) the one-way ANOVA with random effects (i.e., unconditional model), (b) the one-way ANCOVA model with random effects (i.e., conditional model at the student level), (c) the regression model with means-as-outcomes (i.e., conditional model at the classroom level), and (d) an intercepts- and slopes-as-outcomes model (i.e., full model), and (e) the analysis of residuals.

To answer the last research question, how consistent are the ratings of classroom level effectiveness over a three-year period, the strategy was to use the residuals from the multilevel models. First, correlations were performed comparing data from all three years. In this sense, the variables for the correlations consisted of data generated from each classroom for the models that yield residuals. For each model, the year-to-year Pearson correlation coefficients were calculated, meaning *r* for Year1 with Year 2, *r* for Year 1 with Year 3, and *r* for Year 2 with Year 3. Second, a comparison was made after categorizing each classroom by quantiles: well-above average (two standard deviations above the mean), above average (between one and two standard deviations above the mean), average (between one standard deviation above the mean and one standard deviation below the mean), below average (between one and two standard deviations below the mean), and well-below average (two standard deviations below the mean). The comparisons were made on the number of individual teachers that remained in the same category compared to the number of teachers that shifted across categories. This is a critical element of this study since it helped assess the consistency of teacher effectiveness ratings across three years for the same teachers.

Results

Results are presented based on the guiding research questions noted earlier in the paper, and using the sequential HLM modeling approach recommended by Raudenbush and Bryk (2002), starting with the simpler models and finalizing with the more complex models: (a) the one-way ANOVA with random effects, (b) the one-way Analysis of Covariance (ANCOVA) with random effects, (c) the regression with means-as-outcomes, and (d) the model with intercepts- and slopes-as-outcomes.

Research question 1: Is there enough classroom variance (as measured by the intra-class correlation) to justify the use of HLM?

Analytic strategy: One-way ANOVA with random effects (fully unconditional model)

The general model is represented by the following equations:

Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$

Level 2 equation: $\beta_{0j} = \gamma_{00} + u_{0j}$

Expanded model: $Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$

For the purpose of this research project, Y_{ij} is the KCCT reading score for the i^{th} student in the j^{th} classroom. β_{0j} is the mean outcome for the j^{th} classroom. The variable r_{ij} is the error term. This model assumes that these error terms are normally distributed with a mean of zero and a constant Level 1 variance, σ^2 (Raudenbush & Bryk, 2002). In this research the Level 2 variable was the classroom. The Level 2 variable γ_{00} is the grand-mean outcome of all the students with u_{0j} the random effect associated with the j^{th} classroom and is assumed to have a mean of zero and a variance of τ_{00}. In other words, the variable u_{0j} is the deviation of classroom j's mean from the grand mean.

This model is the fully unconditional model in that no predictors are specified at either Level 1 or Level 2. First of all, when examining teacher effectiveness, the researchers had to determine if a two-level model would be more appropriate to use than an OLS regression model. As stated previously, HLM analysis is appropriate to use if there is sufficient variance at the classroom level. The required procedure is called intra-class correlation (ICC) and it provides information about the outcome variability at each of the two levels; in that sense, ICC measures the proportion of the variance in the outcome that is between the Level 2 units (Raudenbush & Bryk, 2002). Sufficient variance has been interpreted as 10% or more (Ma, 2001). In this study, when the one-way ANOVA with random effects was completed, the amount of variance at the classroom level was 21% for Years 1 and 2, and 17% for Year 3 (see Table 6 for details). Therefore, there was more than sufficient variance at the classroom level to justify the use of a multi-level analysis. The classroom means ranged from 541 to 545. The null hypothesis H_0: $\tau_{00} = 0$ provides a test of whether there is significant variation among classroom means on KCCT reading scores. Table 6 shows that obtained χ^2 statistics for all three data sets were significant at $p < .001$.

At this initial stage of the analysis, deviance is another important statistic to be examined. According to Kreft and De Leeuw (1998):

> The difference in deviation is especially useful to estimate the improvement of fit when the between variance is no longer uniquely defined. For that reason deviances are considered the most important feature in the output, and used for an overall evaluation of models. As a rule of thumb, in order to reach the conclusion that one model is a significant improvement over another, the difference in deviances between two models should be at least twice as large as the difference in the number of estimated parameters. (p. 65)

The deviances for the one-way ANOVA with random effects in the current study (all with two parameters) are reported near the bottom of Table 6.

Table 6

Summary of One-Way ANOVA with Random Effects for Three Years of Data

Model characteristics	Year 1	Year 2	Year 3
Grand-mean of KCCT reading scale score	541.10	544.34	544.52
95% confidence interval for grand-mean	(538.55, 543.65)	(542.03, 546.65)	(542.40, 546.64)
Student level variance	1317.85	1039.66	1083.35
Classroom level variance	353.09	279.88	227.28
95% confidence interval for classroom mean	(504.27, 577.93)	(511.55, 577.13)	(514.97, 574.07)
χ^2, *df* (null hypothesis test)	1,763.41, 230*	1,749.53, 234*	1,433.18, 235*
Intraclass correlation	.21	.21	.17
Deviance	58,754.27	54,863.34	56,218.71
Estimated parameters	2	2	2

* $p < .001$

Research question 2: What student level variables are significant predictors of student achievement as measured by their reading scores?

Analytic strategy: one-way ANCOVA with random effects (conditional at Level 1)

In this model, the Level 1 equation was an explanatory variable that was centered around the grand mean, while the Level 2 intercept equation is random and the slope equations are fixed. The equations are as follows:

Level 1 equation: $Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \overline{X}..) + r_{ij}$

Level 2 equations: $\beta_{0j} = \gamma_{00} + u_{0j}$

$\beta_{1j} = \gamma_{10}$

Expanded model: $Y_{ij} = \gamma_{00} + u_{0j} + \gamma_{10}(X_{ij} - \overline{X}..) + r_{ij}$

This model fixes the coefficient, β_{1j}, to be the same for all of the classrooms. In this model γ_{10} is the pooled within-group regression coefficient of Y_{ij} on X_{ij} (Raudenbush & Bryk, 2002). The variable β_{0j} is the mean outcome for each classroom adjusted for the difference among these groups in X_{ij}. It should be noted that "the Var $(r_{ij}) = \sigma^2$ is now a residual variance after adjusting for the Level 1 covariate, X_{ij}" (Raudenbush & Bryk, 2002, p. 25).

The first major research question was associated with student level variables that are significant predictors of student achievement as measured by their reading scores. An ancillary question was associated with the percent of the variance in reading scores accounted by student level variables. The student level variables that were significant for all three years were CTBS reading scale score, free or reduced lunch, female, advance program, African American, and attendance. The students' CTBS reading scores had the strongest relationship of all the variables. The *t*-values for CTBS reading scores were all > 33, while all the other *t*-values were < 10.

The directionality of coefficients associated with student variables were as follows: The coefficient for days absent was negative (i.e., the more days absent the more likely their KCCT reading score was lower), the coefficient for CTBS reading scale score was positive (i.e., the higher the score on the CTBS the more likely the KCCT reading scale score was higher), the coefficient for African American was negative (i.e. African American students were more likely to score lower on KCCT reading test), the coefficient for female was positive (i.e., female students were more likely to have a higher KCCT reading scale score), the coefficient for free or reduced lunch was negative (i.e., students that qualified for free or reduced lunch were more likely to score lower on the KCCT reading test), and the coefficient for advance program was positive (i.e., advance program students were more likely to score higher on the KCCT reading test). In summary, significant student level variables can be classified into: two academic measures (i.e., CTBS reading scale score and advance program) and four non-academic measures (i.e., gender, race, socio-economic status, attendance). See Table 7 for details.

Table 7

Summary of ANCOVA with Random Effects Coefficients

	Year 1		Year 2		Year 3	
Variables[a]	Coefficient	*t*-value	Coefficient	*t*-value	Coefficient	*t*-value
Intercept	542.72	765.74	544.44	802.09	544.46	874.38
Days absent	-.39	-6.50	-.29	-3.37	-.29	-4.99
CTBS scale scores	.45	33.71	.45	33.61	.44	34.29
African american	-7.95	-8.95	-6.82	-8.49	-5.61	-6.57
Female	6.45	9.59	6.65	9.72	6.48	8.58
Free or reduced lunch	-5.98	-7.00	-4.61	-5.81	-6.90	-6.90
Advance program	10.95	4.78	12.41	7.47	13.35	8.23

[a]All t-values are significant at the $p < .001$ level with the exception of Year 2 Days Absent, which is significant at $p < .01$ level.

The ancillary research question asked about the reduction in the Level 2 variance (classrooms) from the significant Level 1 (student) variables. The Level 2 variance from the ANCOVA with random effects model is reduced to 14% (Year 1), 12% (Year 2), and 9% (Year 3). This model explained 36%, 43%, and 50% of the Level 2 variance in Year 1, Year 2, and Year 3 respectively. Thus, the amount of unexplained Level 2 variance is 64%, 57%, and 50% for Year 1, Year 2, and Year 3, respectively. As stated previously, this indicates that a large portion of the Level 2 variance can be attributed to the differences among the students in these classrooms. The Level 2 error term is still significant, meaning that there is still some Level 2 variance to be explained.

Research question 3: What classroom level variables are significant predictors of student achievement as measured by their reading scores?

Analytic strategy: Regression with means-as-outcomes (conditional at Level 2)

The researchers used the same Level 1 model from Stage 1, but introduced Level 2 variables. Thus the Level 2 equation with one predictor variable can be written as:

Level 2 equation: $\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$

Expanded model: $Y_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + r_{ij}$

W_j is a Level 2 predictor. In this model u_{0j} now represents the residual and u_{0j} and τ_{00} is now the residual or conditional variance in the intercept β_{0j} after controlling for the effects of W_j (Raudenbush & Bryk, 2002). For this research, all Level 2 predictors were tested to determine which were statistically significant. Then various combinations of predictor variables were explored. An intra-class correlation coefficient was calculated to determine the conditional variance in KCCT reading scores controlling for all W_j.

In this model, the outcome is predicted by classroom-level characteristics used as predictors (i.e., conditional at Level 2) and answers the second major research question. The second major research question was concerned with significant variables at Level 2 and—as an ancillary question—the amount of Level 2 variance accounted for by those variables. Although the three years originally produced slightly differing models, they all accounted for approximately 32–39% of the Level 2 variance. There were three Level 2 variables that were significant for all three years: teacher years experience, mean class education index, and mean class CTBS reading scale score. These three Level 2 variables explained approximately 29–32% of the Level 2 variance in school means on KCCT reading. The conditional ICC for the three variable models were .155 (Year 1), .161 (Year 2), and .126 (Year 3).

Research question 4: When combining both student and classroom level variables, what are significant predictors of student achievement?

Analytic strategy: Intercepts- and slopes-as-outcomes (full model)

This model is the same as the random-coefficients regression model, except now are included Level 2 predictors W_j. The equations that represent this model are:

Level 1 equation: $Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \overline{X}_{.j}) + r_{ij}$

Level 2 equations: $\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}$

$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}$

Expanded model:

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + \gamma_{10}(X_{ij} - \overline{X}_{.j}) + \gamma_{11}W_j(X_{ij} - \overline{X}_{.j}) + u_{1j}(X_{ij} - \overline{X}_{.j}) + r_{ij}$$

The third major research question was concerned with developing a full prediction model. This is a natural step after identifying the significant Level 1 and Level 2 variables. Answering this research question involved constructing several models. When originally combining the models, there were six Level 1 variables (CTBS reading scale score, African American, female, advance program, total absences, and free and reduced lunch), three Level 2 variables predicting the level one intercept (teacher years experience, classroom education index, and classroom CTBS reading scale score), and one random coefficient (i.e., for CTBS reading scale score). All of these variables, in addition to the Level 1 error term, the intercept, and the Level 2 random component, were included in the full prediction model. The rationale for only one random slope is that none of the other Level 1 variables had a significant residual term, which suggests that there is no significant variation to explain after controlling for the Level 1 variables.

When combining the significant Level 1 and Level 2 variables into the same model, two of the Level 2 variables, classroom education index and classroom CTBS reading scale score, were no longer significant for all three years. Therefore, a revised model dropped the latter Level 2 variables. After dropping the two variables it was noted that for Year 2 and Year 3 there was little difference when examining the deviance, but there was a small difference in the Year 1 models. The revised model explained between 70% and 77% of the Level 2 variance of the intercept, which represented the mean KCCT reading score. See Table 8 for details.

Table 8

Comparison of Intercepts- and Slopes-as-Outcomes Models for Coefficients of β_2

	Year 1†	Year 2†	Year 3†
Variable	Deviance	Deviance	Deviance
Parents' education index	50486**	48473*	49984*
Student total attendance	50492**	n.s.	n.s.
Student advance program	50483**	n.s.	49987**
Teacher years experience	n.s.	n.s.	49995**
Median education	50488**	n.s.	49985*
Teacher non-Caucasian	n.s.	n.s.	49988**
Teacher race other	n.s.	n.s.	49989**
Student free and reduced	n.s.	n.s.	49986**
Student percent attendance	50482**	n.s.	n.s.
Intercepts- and slopes-as-outcomes Model 2	50483	48477	49986

Note. n.s. = not significant,
† Parameter = 4, * $p < .01$, ** $p < .05$

Research question 5: How consistent are the ratings of classroom level effectiveness over a three-year period of time?

Analytic strategy: Analysis of residuals

The last major research question deals with the consistency of the teacher effectiveness ratings using the residuals from year to year. To examine the models and their consistency, three steps were taken. First, the researchers examined three models. The first model (A) used only the CTBS reading scale score as a Level 1 predictor with the coefficients fixed and a random intercept. The second model (B) used the six significant Level 1 variables: absences, CTBS reading scale score, African-American, female, student free or reduced lunch, and advance program and had fixed coefficients and a random intercept. The third model (C) is the same as the second model except the intercept has a predictor variable of Teacher Years Experience.

When examining the correlations, all three models are highly correlated with each other. All correlations are above .92. Thus, all three models appear to be viable options for obtaining teacher effectiveness indices. Table 9 shows the Pearson correlation coefficient between the models for each year. The number of teachers was 241, 235, and 236 for Year 1, Year 2, and Year 3, respectively.

Table 9

Correlation of Models

	Model A/Model B	Model B/Model C	Model A/Model C
Year 1	.98	.95	.92
Year 2	.97	.98	.95
Year 3	.96	.98	.93

Note. All correlations are significant at the .001 level.

The second part of this analysis compared the correlation between the models and the null model across the three years to determine which model was most consistent. Table 10 shows the Pearson correlation coefficient of the residuals from the different models across the three years. The number of teachers was 150, 149, and 116 for Year 1, Year 2, and Year 3, respectively. All four models provided a significant correlation at the $p <$ .001 level, but the Null model and Model A provided the most consistent residuals over time.

Table 10

Correlation of Residuals by Year

	Null model	Model A	Model B	Model C
Year 1	.72	.63	.56	.56
Year 2	.72	.54	.44	.42
Year 3	.68	.59	.52	.50

Note. All correlations are significant at the .001 level.

The third part of this analysis included a comparison of the rankings for teachers that were in all three years of the study ($n = 100$). Following the established practice in value added models, an "effective" teacher would be associated with a positive residual and an "ineffective" teacher with a negative residual. Based on the residuals from Model C, all teachers were given a ranking of 2 = well above average (two or more standard deviations above the mean), 1 = above average (the standard deviation is between one and two standard deviations above the mean), 0 = average (between one standard deviation above and below the mean), −1 = below average (the standard deviation is between one and two standard deviations below the mean) and −2 = well below average (two or more standard deviations below the mean). Table 11 shows a summary of the patterns of teacher rankings over the three years, including the sum frequency of their residual rankings (for example, for the first two teachers, it is noted that their residuals are 2,1,1 and 1,1,2 which means it equals 4). Only one teacher crossed through the average range (i.e., had both a positive and a

negative residual ranking). Four teachers had three different levels of performance (i.e., a different level each year).

Table 11

Summary of Individual Teacher Rankings

Sum (frequency)	Individual teacher combinations of residuals–year 1, year 2, year 3 (frequency)				
4 (n = 2)	2, 1, 1 (n = 1)	1, 1, 2 (n = 1)			
3 (n = 6)	1, 1, 1 (n = 3)	2, 0, 1 (n = 1)	0, 2, 1 (n = 1)	0, 1, 2 (n = 1)	
2 (n = 10)	1, 1, 0 (n = 5)	0, 1, 1 (n = 3)	2, 0, 0 (n = 2)		
1 (n = 17)	1, 0, 0 (n = 5)	0, 1, 0 (n = 5)	0, 0, 1 (n = 7)		
0 (n = 40)	0, 0, 0 n = 39)	1, −1, 0 (n = 1)			
−1 (n = 14)	−1, 0, 0 (n = 2)	0, −1, 0 (n = 2)	0, 0, −1 (n = 10)		
−2 (n = 7)	−1, −1, 0 (n = 2)	−1, 0, −1 (n = 1)	0, −1, −1 (n = 1)	−2, 0, 0 (n = 1)	0, 0, −2 (n = 2)
−3 (n = 2)	−1, −1, −1 (n = 2)				
−4 (n = 2)	−2, −1, −1 (n = 1)	−1, −1, −2 (n = 1)			

Note. Under Individual Teacher Combinations, within the grouped numbers, first number refers to Year 1 ranking, second number refers to Year 2 ranking, and third number refers to Year 3 ranking.

In summary, HLM models were constructed to assess predictors of student performance and to rank-order classroom performance using residuals (see Table 12).

Table 12

Key Results of HLM Models for KCCT Reading Years 1–3

Model	Predictors		Results
	Level–1 significant	Level–2 significant	
One-way ANOVA with random effects (unconditional model)	n/a	n/a	Intraclass correlation between .17 and .21
ANCOVA with random effects (conditional at level-1 or using student-level predictors)	1) Days absent 2) CTBS reading scale score 3) African American 4) Female 5) Free or reduced lunch 6) Advance program	n/a	Intraclass correlation between .09 and .14
Regression with means-as-outcomes (conditional at level-2 or using classroom-level predictors)	n/a	1) Mean teacher experience 2) Mean education index 3) Mean CTBS reading scale score	Conditional intra-class correlation between .13 and .16
Intercepts- and slopes-as-outcomes models (full model or using student- and classroom-level predictors)	1) Days absent 2) CTBS reading scale score 3) African American 4) Female 5) Free or reduced lunch 6) Advance program	Teacher years of experience	This model explains between .70 and .76 of the variance of the level-2 intercept

Discussion

Despite the black box associated with evaluations of teaching-and-learning processes (Munoz, 2005), researchers and practitioners are coming to an agreement that teachers are the most important factor determining a child's educational success. Teachers affect student learning in multiple ways (Stronge, 2007, 2010), particularly how students learn (i.e., pedagogy), what students learn (i.e., curriculum design), and how much students learn (i.e., achievement). This is particularly important as the re-authorization of the Elementary and Secondary Education Act (ESEA) looms in our high-stakes accountability environment across our nation's public schools.

Application of Various Models to Teacher Effectiveness Analyses

From the start of this research the overriding question has been whether there is a statistical model that will identify effective teachers. The approach decided upon for the study was a value-added approach. There are several advantages of using the value-added approach. These advantages include taking into account the student's prior achievement, student characteristics, classroom characteristics, and the effect students have on each other (Palardy & Rumberger, 2008). The best model for determining the teacher's value-added score was a regression-based, multi-level model known as HLM.

Multiple models were obtained to determine the teacher effectiveness ratings. The null model was only used to assess whether the ICC was enough to appropriately use HLM. The first model, which is similar to the one used for TVAAS, is based on a student's previous scale score (CTBS reading scale score). This is a very efficient model due to the reliance on previous test scores. The second and third models are similar in many respects to the Dallas model because they also include student characteristics.

For this study, the third model was the one the researchers used to obtain the teacher effectiveness ratings. The third model uses teacher years experience to predict the level one intercept. An argument could be made to use the second model that examines teacher effectiveness regardless of their experience. However, the researchers decided to use the model that included teacher experience as a predictor. The reason is that the more experienced teachers could potentially obtain better results and the researchers did not want to over-identify newer teachers as less effective, because they could be very effective for their stage of career. The researchers would recommend using the first (only previous test scores at the student level) or the third model (previous test scores at the student level and teachers' years of experience at the classroom level). Only previous test scores as predictor is definitely more efficient (Sanders, 2000; Sanders & Horn, 1995; Sanders & Rivers; 1996).

The multi-level model was a natural fit for school data, since students are nested within classrooms and since there was enough level-two variance (i.e., ICC). The proposed model is consistent with previous research in that prior achievement is the definitively strongest predictor to current achievement. In the proposed model, the CTBS reading scale score was significantly the largest predictor of the KCCT reading scale score. This supports the TVAAS in their claim that students' prior test scores is the best control to determine teacher effectiveness.

Previous research on years of experience still has many unanswered questions. In this study, teacher experience was a valuable predictor of student learning at the classroom level, but the knowledge base shows mixed findings on the strength of this predictor (Munoz & Chang, 2007; Rockoff, 2004; Rivkin, Hanushek, & Kain, 2005; Wayne & Youngs, 2003). For example, Munoz and Chang (2007) indicated that years of experience have little

effect on student achievement in high school reading achievement gains in a large urban district. Contrasting with the above finding, a study by Wayne and Youngs (2003) found that most of the studies they reviewed on student achievement and teacher experience yielded a positive effect of teaching experience. Also, more experienced teachers are more likely to be drawn to higher performing schools (Cochran-Smith & Zeichner, 2005). Generalizations of these findings are limited as teacher experience is often impacted by the subject area and the market forces.

Stability of Teacher Effectiveness Classifications

In terms of the consistency of "classifications," the following observations were made from the examination of the teacher rankings. There were no teachers that changed over two levels. The largest changes in classification from one year to the next was two levels, from well above average to average (n = 2), average to well above average (n = 1), well below average to average (n = 1), average to well below average (n = 2), and above average to below average (n = 1). As already noted, it is interesting that only one teacher crossed over the average range (i.e., the ranking went from above average to below average).

With the limited movement of teachers across classifications, especially the restricted number of cases with movement across more than one classification (above, average, or below), the findings tend to suggest stable and trustworthy teacher effectiveness ratings over multiple years. In fact, of the 100 teachers included in the study across all three years, 35 teachers remained in the above average classification for all three years, 39 teachers remained on the average classification for three years, 25 teachers remained on the average classification, and only one teacher crossed the below and above classification for the three years. Given this relatively high level of consistency, it appears that a single year rating for a teacher is relatively stable (Sanders & Horn, 1994). When three years of data produce three-year average classifications, the stability for classifying teachers' effectiveness is even more trustworthy.

Nonetheless, we do not advocate using value-added modeling as a single source for identifying teacher effectiveness. Rather, we strongly encourage using measures such as those produced in this examination as merely one source among multiple data sources to assess teacher effectiveness (Peterson, 2006). Moreover, we encourage due caution when interpreting and applying findings from value-added modeling (or, for that matter, any applicable data source, such as classroom observations, student surveys, etc.) in judging teacher evaluation. Multiple data points are always needed.

Policy Implications

The policy implications of this study are analyzed using two seminal frameworks for connecting teacher evaluation to student achievement: (a) nine guidelines developed by James Stronge and Pamela Tucker (2000), and (b) four guidelines developed by Jason Millman (1997). The researchers understand that there is no perfect way for evaluating policy implications for teacher evaluation systems like the one investigated in this study. Currently, and regardless of the approach, teacher evaluation systems are deemed by most teachers (and sometimes even for administrators) to be extremely stressful and of little value.

Evaluation of proposed model using guidelines from James Stronge and Pamela Tucker. Stronge and Tucker (2000) provided nine guidelines for using testing data models as part of teacher evaluations, including (a) student learning as only one component, (b) the importance of context, (c) the value of growth, (d) longitudinal perspective on students, (e) gain scores limitations, (f) time frame, (g) fairness and validity of measures, (h) alignment issues, and (i) teaching to the test issues. These guidelines provide discussion points for evaluating the strengths and weaknesses of the current model for policy implications.

The first guideline is to "use student learning as only one component of a teacher evaluation system that is based on multiple data sources" (Stronge & Tucker, 2000, p. 53). The current model uses only the KCCT reading score and that is problematic. In high-stakes testing environments, it is often easy to equate single-subject test scores to teacher effectiveness. This model should be interpreted as teacher effectiveness on teaching reading as measured by the KCCT reading test. The second guideline is "when judging teacher effectiveness, consider the context in which the teaching and learning occur" (p. 54). There are contextual issues that may inflate or deflate student scores beyond the student's previous attainment and the teacher's effectiveness. The models used for this study examined some of those variables (e.g., absenteeism), but did not examine many of the other contexts mentioned. The third guideline is to "use measures of student growth versus a fixed achievement standard or goal" (p. 55). This is one of the strengths of this model in that students' performance is examined in terms of their own previous academic attainment. By using previous achievement scores, teachers are more likely to be judged on their own ability to increase student learning.

The fourth guideline is to "compare learning gains from one point in time to another for the same students, not different groups of students" (p. 55). This is a notable strength of the proposed model. Teachers' effectiveness is determined by comparing students' gains to their own prior achievement, not by comparing their gains to a previous group of students' achievement. The fifth guideline is to "recognize that gain scores have pitfalls that must be avoided" (p. 56). Many models, including the proposed model, do

not take into account "the regression to the mean" phenomenon which is the likelihood that those students that scored high on the previous test are more likely to score lower on the next test and conversely. Another pitfall can be a ceiling effect if the pre- and post-test measures are not sufficiently robust to account for growth of high ability students. This would be a problem particularly if there are large numbers of high ability students grouped in the same classes. The sixth guideline is to "use a time frame for teacher evaluation that allows for patterns of student learning to be documented" (p. 56). The proposed model produced stronger results when examining teacher scores over a period of three years instead of a single year, "snapshot," model.

The seventh guideline was to "use fair and valid measures of student learning" (p. 56). The test used for this study has undergone severe scrutiny to determine validity and reliability. The eighth guideline is to "select student assessment measures that are most closely aligned with existing curriculum" (p. 57). The curriculum that was being tested was well documented for teachers. The curriculum included content guides, often called content standards, which explicitly stated what content items the students should be taught and would be tested. The ninth guideline is to "not narrow the curriculum and limit teaching to fit a test" (p.58). The weakness of the proposed model is that it only examines the results from the KCCT test and, therefore, is more likely to positively recognize teachers that may be narrowing the curriculum to fit the KCCT test, commonly referred to as "teaching to the test."

Evaluation of proposed model using guidelines from Jason Millman. Millman (1997) provided four criteria for use when examining high-stakes assessment systems and their relationship to teacher evaluation: (a) fairness, (b) comprehensiveness, (c) competitiveness, and (d) consequential validity.

The first criterion is that of fairness. According to Millman, "the single most frequent criticism of any attempt to determine a teacher's effectiveness by measuring student learning is that factors beyond a teacher's control affect the amount students learn" (Millman, 1997, p. 244). The proposed model controls for some of these factors, including attendance, race, socio-economic status, and prior achievement.

The second criterion is that of comprehensiveness. Millman stated that "this criterion refers to the degree the range of intended learning outcomes are represented in the assessment measures" (p. 245). The problem here is the potential loss of teaching an enriched curriculum as well as teachers finding themselves "faced with disincentives to follow their perceived best practices if doing so would detract from student performance according to the accountability system" (p. 245). The third criterion is that of competitiveness. Millman's question is: how competitive is this method of teacher evaluation compared to other methods used? There is currently no known perfect evaluation method. Although different teacher evaluation methods have advantages and disadvantages, the proposed model appears to be com-

petitive in comparison to other methods. The fourth criterion that Millman proposed is that of consequential validity. This criterion examines the desired and undesirable effects of a teacher evaluation method, even if the method is accurate and appropriately used. At this point, since the proposed model has not been used as an evaluation tool, the consequential validity has not been determined. Assuming the model is used, the potential consequential validity could be positive if teachers view the model as their personal impact on student learning (i.e., adding value) and that student characteristics are taken into account. The potential negative side could be that teachers are reduced to just a number and the variables do not sufficiently account for all student characteristics that could impact student learning.

Implications for Practice

The value-added method explored in this study gives administrators an objective measure to be able to identify the most and least effective teachers as determined by the KCCT reading scale scores. By using the index scores (i.e., teachers rated as well above average to well below average), administrators can identify which teachers need assistance in helping their students achieve. Similarly, administrators can identify those teachers whose students are achieving at higher than expected rates. However, it should be noted that value-added models such as those tested in this study only identify those teachers that are being successful or not; they do not identify the behaviors or instructional techniques that are making them successful.

In an era of high stakes testing, any additional information educators and policy makers can obtain to know which teachers are adding value and which are not is useful information that can be used to assist in developing strategies to increase student success. It is recommended that until further validation has occurred, these scores should be used in connection with numerous other evaluation tools already at teacher evaluators' disposal, such as teacher observations, teacher portfolios, and student work. For example, the proposed model could be used to assist principals or other teacher evaluators in identifying low performing teachers and give them some assistance and guidance in how to become a better teacher. Again, this should be done with multiple data sources (Peterson, 2006).

Research Limitations

As we hope is evident, the proposed value-added model has several advantages, but there also are numerous limitations. Teacher effectiveness is a complex phenomenon when it comes to defining it or—even more—on how to measure it. Selected limitations have already been mentioned when evaluating the proposed model by the guidelines provided by Stronge and Tucker (2000) and the criteria provided by Millman (1997). Additionally, another limitation and a primary concern about the use of this model is that

it identifies effective teachers based solely on the students' Kentucky KCCT Reading scores. If one is not careful, one could easily over-simplify this result and apply the label of "effective" or "ineffective" to a teacher in general. It is important to remember that there is more to being an effective teacher than just producing reading results on a point-in-time test, even if the test has extensive validation information to support its use.

A second limitation is that the objective of this research was to identify the most, and least, effective teachers. This is not to be confused with determining effective, and ineffective, teaching strategies. The proposed model can only be used to assist in identifying teachers who are achieving greater, or lower, gains than expected. A third limitation is that the teachers are assigned to students they have in their classes during testing. The data do not account for high mobility students within a school or classroom. As noted earlier, students were excluded from the data set if they were not enrolled within the school district for at least 100 days, but the data set does not take into account whether the student changes schools within the district. Related to this issue, each year of data was treated independently and later compared on a yearly basis. Another approach could be to use an HLM growth model that links all three years of data (Raudenbush & Bryk, 2002).

Recommendations for Future Research

At least four recommendations can be presented for future research. This study focused solely on the Kentucky Reading Test for fourth graders. As a result, one recommendation for future studies is to develop similar models to include other tested content areas as well as other tested grade levels. The second recommendation is to validate the teacher rankings by having principals complete a rating scale on teachers and then calculate the correlation between the principal's ratings and the model indexes. In addition, trained observers could study teachers who were identified as most effective and least effective by the HLM model to determine productive and counter-productive instructional strategies (similar to the study conducted by Stronge et. al., 2008). A third recommendation is to study the impact of this kind of modeling when it is applied to Title I schools or schools with high teacher turnover.

The fourth and final recommendation for future research is to explore the role of teacher-to-teacher collaboration and the emergence of professional learning communities as related to classroom-based student achievement. Schools that emphasize distributive leadership and shared decision-making have produced notable gains in student achievement (Leithwood, Louis, Anderson, & Wahlstrom, 2004). Additionally, past research in urban elementary schools indicated that teachers who work together on school improvement efforts tend to increase student achievement results (Goddard, Goddard, & Tschannen-Moran, 2007).

Conclusion

The most important school-related contributing factor to student achievement is the quality of teaching. While traditional approaches to teacher effectiveness have focused heavily on input and process variables, the purpose of this study was to explore the use of HLM to determine teacher effectiveness as measured with reading outcomes. Results identified high-, average-, and low-performing classrooms, significant predictors of student learning gains, and addressed stability of teacher effectiveness ratings. The study found limited value-added fluctuations from year to year as succeeding cohorts of students move through these teachers' classrooms. Although not conclusive due to the need for replication, the instability of classroom effectiveness indices found in this study was minimal and indicates the potential usefulness of value added as an indicator of future performance.

Multi-level models hold substantial promise as a tool for helping determine teacher effectiveness. Despite multiple potential drawbacks, multi-level analysis may provide valuable information to use in personnel decisions and teacher compensation systems. The fact that value-added estimates will never measure precisely the quality of instruction in a classroom, to a certain extent, recognizing the methodological limitations can facilitate more informed uses of standardized test results and the development of stronger value-added models.

Most certainly, many questions remain unanswered. For instance, Rothstein (2010) raised issues about statistical bias when using value-added estimates, even when adjusting for students' prior achievement measures. Rothstein argued that some teachers might be assigned students that are systematically different (e.g., motivation, parental engagement) which affect their performance and these differences may not be captured by prior achievement or demographic variables. Ravitch (2010) added that value-added assessment has potential problems such as narrowing the curriculum, promoting teaching to the test, and discarding social factors that influence student scores.

Teacher effectiveness is at an important crossroad. Perhaps more work is needed before the value-added techniques can be used to support important decisions about teachers; in particular, we need models that allow researchers to build on the tradition of process-product research (Good & Brophy, 1997) which means that we need to peer inside the black box of classrooms and not only focus on outcomes (Rowan, Correnti, & Miller, 2002). If these models are used when evaluating teachers, they should be used along with other measures of teacher performance in our schools (Peterson, 2006; Stronge, Ward, & Grant, 2011). These models will also have to take into consideration the collaborative nature of teacher-to-teacher relationships in professional learning communities and not focus on rewarding the already archaic individuality of the teaching profession. A better identification of effective teachers and an improved understanding of what characterizes good teaching have significant implications for deci-

sion-making regarding the preparation, recruitment, selection, compensations, in-service professional development, and evaluation of teachers. No matter the perspective on assessing teacher effectiveness—our children deserve the best teachers, teachers who can make a difference in their success in school, college, and life.

End Notes

[a] Student records deleted due to the 100 day requirement were 108 (Year 1), 71 (Year 2), and 108 (Year 3).

[b] This classroom size requirement caused the removal of 979 (Year 1), 831 (Year 2) and 858 (Year 3).

References

Bankston C. L., & Caldas, S. J. (2000). White enrollment in nonpublic schools, public school racial composition and student performance. *The Sociological Quarterly, 41*, 539–550.

Berends, M. (2000). Teacher reported effects of New American School designs: Exploring relationships to teacher background and school context. *Educational Evaluation and Policy Analysis, 22*(1), 65–82.

Burns, R. B., & Mason, D. A. (2002). Class composition and student achievement in elementary schools. *American Educational Research Journal, 39*(1), 207–233.

Carbonaro, W. J., & Gamoran, A. (2002). The production of achievement inequality in high school English. *American Educational Research Journal, 39*(4), 801–827.

Cochran-Smith, M., & Zeichner, K. M. (2005). *Studying teacher education: The report of the AERA panel on research and teacher education.* Mahwah, NJ: Lawrence Erlbaum Associates.

Darling-Hammond, L. (1996). What matters most: A competent teacher for every child. *Phi Delta Kappan 78*(3), 193–200.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1), 1–44.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis, 24*(2), 81–112.

Doran, H. C. (2003). Adding value to accountability. *Educational Leadership, 61*(3), 55–59.

Goddard, Y. L., Goddard, R. D., & Tschannen-Moran, M. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary schools. *Teachers College Record 109*(4), 877–896.

Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Applied Statistics, 49,* 399–412.

Good, T. L., & Brophy, J. E. (1997). *Looking in classrooms* (7th ed.). New York: Addison-Wesley.

Guthrie, J. T. (2001). Benefits of opportunity to read and balanced instruction on the NAEP. *Journal of Educational Research, 94*(3), 145–162.

Hanushek, E., & Lindseth, A. (2009). *Schoolhouses, courthouses, and statehouses: Solving the funding-achievement puzzle in America's public schools.* Princeton, NJ: Princeton University Press.

Heck, R. H., & Crislip, M. (2001). Direct and indirect writing assessments: Examining issues of equity and utility. *Educational Evaluation and Policy Analysis, 23*(1), 19–36.

Holland, R. (2001). How to build a better teacher. *Policy Review, 106,* 37–47.

Jerald, C. (2003). Beyond the rock and the hard place. *Educational Leadership, 61*(3), 12–16.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling.* London: Sage.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 287–298.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35*(2), 125–141.

Lee, V. E., & Bryk, A. S. (1989). A multilevel model of the school distribution of high school achievement. *Sociology of Education, 62,* 172–192.

Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning: Review of research.* St. Paul, MN: University of Minnesota Center for Applied Research and Educational Improvement.

Ma, X. (2001). Health outcomes of elementary school students in New Brunswick: The education perspective. *Evaluation Review, 24,* 435–456.

Mandeville, G. K. (1988). School effectiveness indices revisited: Cross-year stability. *Journal of Educational Measurement, 25*(4), 349–356.

Mandeville, G. K., & Anderson, L. W. (1987). The stability of school effectiveness indices across grade levels and subject areas. *Journal of Educational Measurement, 24*(3), 203–216.

Marsh, H. W., Hau, K., & Kong, C. (2002). Multilevel causal ordering of academic self-concept and achievement: Influence of language of instruction (English compared with Chinese) for Hong Kong students. *American Educational Research Journal, 39*(3), 727–763.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: Rand.

Millman, J. (1997). How do I judge thee? Let me count the ways. In J. Millman (Ed.), *Grading teachers,grading schools: Is student achievement a valid evaluation measure?* (pp. 243–247). Thousand Oaks, CA: Corwin.

Munoz, M. A. (2005). Black box. In S. Mathison (Ed.), *Encyclopedia of Evaluation* (pp. 34-35). Thousand Oaks, CA: Sage.

Munoz, M. A., & Chang, F. C. (2007). The elusive relationship between teacher characteristics and student academic growth: A longitudinal multilevel model for change. *Journal of Personnel Evaluation in Education, 20,* 147–164.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257.

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in first grade: The importance of background qualifications, attitudes, and instructional practices for student learning. *Educational Evaluation and Policy Analysis, 30*(2), 111–140.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction.* London: Wadsworth.

Peterson, K. D. (2006). Using multiple data sources in teacher evaluation. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (2nd ed., pp. 212–232). Thousand Oaks, CA: Corwin Press.

Pituch, K. A. (1999). Describing school effects with residual terms. *Evaluation review, 23*(2), 190–211.

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

Raudenbush, S. W., Fotiu, R. P., & Cheong, Y. F. (1998). Inequality of access to educational resources: A national report card for eighth-grade math. *Educational Evaluation and Policy Analysis, 20*(4), 253–267.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education.* Philadelphia, PA: Basic Books.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review, 94*(2), 247–252.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics 125*(1), 175–214.

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 elementary schools. *Teacher College Record, 104*(8), 1525–1567.

Sanders, W. L. (2000). Value-added assessments from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*(4), 329–339.

Sanders, W. L., & Horn, S. P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8,* 299–311.

Sanders, W. L., & Horn, S. P. (1995). Educational reassessment reassessed: The usefulness of standardized and alternative measures of student achievement as indicators for the assessment of educational outcomes. *Educational Policy Analysis Archives, 3*(6). Retrieved from http://epaa.asu.edu/ojs/article/view/649.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Stronge, J. H. (2007). *Qualities of effective teachers.* Alexandria, VA: Association for Supervision and Curriculum Development.

Stronge, J. H. (2010). *Effective teachers = student achievement: What research says.* Larchmont, NY: Eye on Education.

Stronge, J. H., & Tucker, P. D. (2000). *Teacher evaluation and student achievement.* Washington, DC: National Education Association.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*(4), 339–355.

Stronge, J. H., Ward, T. J., Tucker, P. D., & Hindman, J. L. (2008). What is the relationship between teacher quality and student achievement? An exploratory study. *Journal of Personnel Evaluation in Education, 20*(3–4), 165–184.

Swanson, C. B., & Stevenson, D. L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. *Educational Evaluation and Policy Analysis, 24*(1), 1–27.

Thum, Y. M., & Bryk, A. S. (1997). Value-added productivity indicators: The Dallas system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 100–109). Thousand Oaks, CA: Corwin.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Educational Research, 73,* 89–122.

Webster, W. J., & Mendro, R. L. (1997). The Dallas Value-Added Accountability System. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 81–99). Thousand Oaks, CA: Corwin.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* New York, NY: The New Teacher Project.

Wilkins, J. L. M., & Ma, X. (2002). Predicting student growth in mathematical content knowledge. *Journal of Educational Research, 95*(5), 288–292.

Marco Munoz is an Evaluation Specialist in the Accountability, Research, and Planning Department at Jefferson County Public Schools, Louisville, Kentucky.

Joseph A. Prather is an Evaluation Specialist in the Accountability, Research, and Planning Department at Jefferson County Public Schools, Louisville, Kentucky.

James H. Stronge is the Heritage Professor in the Educational Policy, Planning, and Leadership Area at the College of William and Mary, Williamsburg, Virginia.