Investigating the "black box" of effective teaching: the relationship between teachers' perception and student achievement in a large urban district

**Marco A. Muñoz · Julie R. Scoskie ·
Diana L. French**

Abstract Given the international need to improve student learning, there is nothing more important than classroom teachers. Obtaining a deeper understanding of effective classrooms is a priority if educational reform efforts are to succeed in any educational system around the world. In the last decade, educational researchers have expanded the knowledge base about the impact of teachers, but far less is known about what effective teachers do that make such a positive difference. This investigation focuses on the relationship between attributes of effective teaching, as perceived by both more and less effective teachers, and fourth grade reading achievement results of their students. This investigation used value-added modeling methods as the first step for identifying the more and less effective fourth grade teachers ($N=261$) and their students ($N=6,962$). Survey research methods were used to compare the perceptions about characteristics of effective teaching for teachers previously identified—in the value-added model—as (a) more effective and (b) less effective. Findings indicated that the more effective teachers value classroom management and organization as the number one characteristic of effective teaching; that, in turn, enables the more effective teachers to focus classroom time on student learning. Implications for theory and practice are discussed.

Keywords Teacher effectiveness · Teacher quality · Value-added assessment · Classroom management · Instructional practices · Student achievement

1 Introduction

Understanding how teachers impact student learning is a critical question for educators. A growing body of research has shown that teacher effectiveness is a strong

M. A. Muñoz (✉) · J. R. Scoskie · D. L. French
Data Management, Planning, and Program Evaluation Services Division,
Jefferson County Public Schools, 3332 Newburg Road, Louisville, KY 40218, USA
e-mail: marco.munoz@jefferson.kyschools.us

predictor of student achievement (Darling-Hammond 2000; Hanushek and Lindseth 2009; Munoz and Chang 2007; Nye et al. 2004; Stronge et al. 2008). Improving student achievement is the primary goal of school reforms, and the importance of the teacher in meeting this goal is well established (Stronge 2007). With the persistent accountability pressure for schools (and consequently for teachers) to increase student academic achievement, improving teacher effectiveness is an imperative in many countries like the USA (American Recovery and Reinvestment Act 2009 and around the world (Hattie 2003).

Internationally, studies indicate that determining effective teaching domains has proved a challenge for education researchers (Hattie 2003; Kyriakides et al. 2006). Kyriakides et al. (2006) found that value-added measures alone did not encompass all the constructs to identify an effective teacher. Teacher perceptions were captured to ascertain effective teacher domains associated with positive educational outcomes; this approach provided evidence of what teachers felt determined effective teaching. The authors stated that focusing only on teacher effectiveness in essence "decontextualized from the school effect" and thus revealed a limitation in measuring teacher effectiveness (p. 19). In another recent international study, Teddlie and Liu (2008) studied rural and urban areas in China while looking at more and less effective schools. Using traditional effectiveness variables as the dependent measure, significant differences were found between more effective and less effective schools as well as between rural and urban schools. Equally interesting was the qualitative findings that helped identify unique characteristics of effective teaching in China. From these international studies, we can ascertain that teachers matter, despite any differences in educational structures and environments.

Within the USA, numerous studies have used value-added models (e.g., multilevel or hierarchical statistics) when examining teachers' influence on student achievement; however, few empirical studies have linked what effective versus less effective teachers do differently (Stronge 2010). Investigations without clear links between teacher characteristics and student achievement are usually referred as "black box" studies. Black box studies are the kind of research that lacks an articulated theory that provides insight as to what is presumed to be causing the outcome (Muñoz 2005). According to Goe (2007), teacher quality consists of two dimensions: (1) the task of teaching and (2) student learning and achievement. Although there is a general agreement that teacher quality matters, there is no consensus on which aspects of teacher quality matter most. Therefore, examining student outcomes in combination with teacher perceptions may help inform the "black box" of what makes an effective classroom teacher.

This study added to existing research linking student achievement and effective teacher qualities by combining two methodological approaches in a sequential mode: (a) value-added modeling to identify more and less effective teachers and (b) survey research to compare their perceptions about what matters most. In the past, when researchers have compared teachers near the top end of the quality distribution versus teachers near the bottom end of the quality distribution, it has been observed that higher quality teachers can lead their students to a full year's worth of achievement (Hanushek and Rivkin 2012; Rivkin et al. 2005). However, we need to understand more about the significant variation associated with specific teacher qualities that might positively or negatively impact student learning as measured by standardized

achievement tests. The next two sections provide background information on a methodology and a theoretical framework used to examine this critical link between teacher qualities and student achievement.

2 Value-added modeling and teacher effectiveness

Until recently, school effectiveness research (Teddlie and Reynolds 2005) 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 1990s, multilevel statistical models, such as Hierarchical Linear Modeling (HLM), have been gaining popularity in educational research (Raudenbush and Bryk 2002) as a means to examine the effects of different levels of grouping on student achievement (e.g., classroom, schools, and districts). Examples of educational uses of multilevel statistical models have increased exponentially in the last decade. 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 et al. 2002), teacher and school characteristics on student achievement (Berends 2000), student achievement in a specific discipline (Wilkins and Ma 2002), and specific strategies for student achievement (Desimone et al. 2002).

However, the value-added approach to teacher effectiveness has been criticized due to the difficulty of capturing the complexity of effective teaching. Kupermintz (2003), for example, argues that to label a teacher as effective based on the gains of their students is logically faulty: (a) teachers with fewer students have less accurate data, and their estimates are more likely to be "pulled" toward the district average and (b) there are potentially conflicting explanations, other than teacher effectiveness, for the performance of students on tests. Despite some criticism, HLM has emerged as a well-accepted statistical model to use when conducting a study of school effects within an educational setting (Raudenbush and Bryk 2002).

One reason for this growing acceptance is that HLM allows for the effects of the classroom context (i.e., students are nested in classrooms) to be taken into account. In a classic work on the topic, Lee (2000) pointed out that there are three problems when using a single level method, like Ordinary Least Squared (OLS) multiple regression and analysis of variance (ANOVA): (a) aggregation bias, (b) misestimated standard errors, and (c) heterogeneity of regression. The aggregation bias can occur when a variable takes on different effects at diverse levels of aggregation. A second difficulty concerns misestimated standard errors which 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 (Lee 2000). 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. Lastly, the

problem of heterogeneity of regression is solved by HLM since HLM allows for the investigation of grouping effects. Although it is not perfect, the aim of research using this value-added methodology has been to identify more and less effective teachers based on student achievement results and trying to move a step forward toward a more empirical-based approach to identifying teaching effectiveness.

3 Theoretical framework on teacher effectiveness

Stronge's book, *Qualities of Effective Teachers* (2007), served as the conceptual framework for this study. Stronge's framework evolved from a comprehensive review and synthesis of research relating to effective teaching. This framework includes 27 research-based qualities for effective teachers grouped in six domains: (a) prerequisites for effective teaching, (b) teacher as a person, (c) classroom management and organization, (d) planning for instruction, (e) implementing instruction, and (f) monitoring student progress (Stronge 2007). Despite the fact there is no universal agreement regarding qualities of effective teachers, there are common elements of teaching logically and empirically linked to student outcomes (Ellett and Teddlie 2003). Although researchers use different terms for similar teacher characteristics and have different conceptual views for the way characteristics are organized, there are similarities and consistent themes (Stronge 2007). Since Stronge's framework (2007) undergirds this study, a review of research in each of the six domains follows.

Prerequisites for teaching Research conducted by Darling-Hammond (2000) found a relationship between teacher quality and student achievement based on the teacher's education, licensing, and professional development; for example, teachers certified or degreed in their teaching field have students who achieved higher in reading and mathematics than teachers who are teaching in areas for which they were not certified. Wenglinsky (2000) found teachers with a major in content areas resulted in higher student achievement especially in mathematics and science; in addition, these teachers engaged in higher-level questioning and utilized more student-centered activities in the classroom. While research has indicated that years of teaching experience and ethnicity have little effect on student achievement (Early et al. 2007; Muñoz and Chang 2007), teachers' verbal ability scores can directly impact student achievement (Coleman et al. 1966; Darling-Hammond 2000). However, school improvement efforts and reform policies focusing on teacher prerequisites alone are not likely to improve student outcomes (Rockoff et al. 2008).

Teacher as a person Caring is an important quality of effective teachers. Noddings (2001) described a caring relation as a caring connection or encounter between two humans. Under this perspective, the student is viewed as more important than the content matter; thus, relations provide the foundation for successful pedagogical methods. Others support the notion of the caring connection by indicating the need of teacher preparation programs to address building teacher–student relationships as a way of facilitating academic learning (Early et al. 2007). The importance of nurturing the strengths of students in a caring environment is a reoccurring theme in effective teacher research (Lumpkin 2007; Noddings 2006). Teachers do not teach classes, they

teach students. The concept that teachers build academic success on social and emotional learning has been supported by the work of Zins et al. (2004). Muñoz and Vanderhaar (2006) found some evidence associated with a large urban district that indicated it is possible to establish a caring community of learners that enhance academic achievement. In fact, for children displaying difficulties adjusting to the classroom having teachers attending to their social and emotional needs may be more important to their academic progress than instructional practices (Burchinal et al. 2002; Hamre and Pianta 2005).

Classroom management and organization Marzano (2007) reported that quality student and teacher relationships resulted in 31 % fewer discipline problems; effective teachers not only attend to students' emotional needs but also anticipate and help prevent disturbances before they happen. These teachers may spend less time correcting disturbances (Hattie 2003). Similarly, a well-managed classroom is one where "the teacher has clear yet flexible expectations related to the classroom rules and routines. Children understand and follow rules and the teacher does not have to employ many control techniques" (Hamre and Pianta 2005, p. 947); in fact, effective instruction promotes positive student attitudes about learning in a safe, orderly learning environment. Other important aspects of classroom management include: beginning the school year with an emphasis on classroom management, providing classroom arrangements conducive to learning, and communicating, as well as implementing rules and operating procedures (Emmer et al. 2003).

Planning and organizing for instruction When planning and organizing for instruction, teachers' expectations influence student academic achievement (Cotton 2001; Brophy and Good 1986). Similarly, teachers communicating lower expectations for certain students may limit their achievement. Organizing for effective instruction includes developing clear lessons with learning objectives and linking these plans to instructional activities (Cotton 2001). Effective teachers plan ways to implement instruction by blending small- and whole group with individualized instruction (Stronge 2007).

Implementing instruction After carefully planning for instruction, effective teachers utilize a repertoire of instructional strategies to support student engagement and learning. Marzano's (2003) What Works in Schools details three factors impacting teacher effectiveness: (a) instructional strategies, (b) classroom management, and (c) curriculum design. The effective teacher employs instructional strategies that emphasize students' prior knowledge, student engagement, content complexity (e.g., deep learning rather than rote memorization), and inquiry methods (Hattie 2003; Stronge 2007). Hamre and Pianta (2005) state "Notwithstanding the importance of relationships and social support, the nature and quality of instruction is of paramount importance for the value of classroom experience that is intended to produce gains in learning" (p. 951). Focused instruction, quality feedback, and student engagement are important enablers of achievement gains for low socioeconomic status students (Hamre and Pianta 2005).

Monitoring student progress Feedback can have a powerful influence, and its impact can be either positive or negative. Positive feedback can enhance classroom learning and

teaching. Feedback is "information provided by an agent (teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (Hattie and Timperley 2007, p. 81). According to Hattie and Timperley (2007), effective teaching imparts information and understanding to students and involves assessing and evaluating students' understanding so instruction can be targeted taking into consideration the present understanding of the students. Since a typical teacher spends one quarter to one third of his/her work time on assessment-related activities, effective teachers assess accurately, frequently, and use results to help students prosper (Stiggins 2004). Formative assessment processes are useful for providing meaningful feedback to teachers and students so that both parties can adjust their teaching and learning tactics. Effective teaching includes ongoing assessment informing teachers and helping students take responsibility for their own academic success (Stiggins 2004).

In summary, although a review of the literature does not reflect a consensus among researchers regarding the teacher qualities resulting in the greatest gains in student learning, the primary aim of this research study was to determine if there are differences between teacher perceptions of effective teacher qualities in classrooms identified as effective and less effective based on student achievement. This study adds to the existing literature linking teachers' perceptions of teacher qualities with increased student achievement and moves the field of teaching effectiveness toward a deeper understanding of what it takes to make a positive impact on student learning.

4 Research purpose and phases

This investigation follows the tradition of the process-product research, also known as the teacher effects research, by looking at both perceptions about processes as well as outcomes (Berliner 2005; Good and Brophy 1997). Using value-added education approaches (i.e., HLM and the residual scores produced by this multi-level methodology), this investigation studied the measurable impact of teachers on student achievement in fourth grade students as measured by state-wide reading assessment data. In addition, this investigation studied how the perceptions of more effective teachers (i.e., teachers whose students experience more than expected academic achievement) differ from less effective teachers (teachers whose students experience less than expected academic achievement) in an academic year. To address these two critical topics associated with teacher effectiveness, the authors established two research questions, answered separately in a two-phase study.

Phase 1:   What is the measurable impact that teachers have on student achievement—using value-added methodology—as measured by the fourth grade state reading assessment?

Phase 2:   After identifying and grouping more and less effective teachers based on student achievement data, what are the differences in perceptions about effective teacher characteristics?

The significance of this investigation is that it was conducted using two sound methodological approaches that—when combined—offer a unique perspective

towards understanding teaching effectiveness for both more and less effective teachers. By combining the findings from the value-added analysis of teacher effects (i.e., Phase 1) with the results derived from the teacher survey (i.e., Phase 2), this investigation sheds light into important links between attributes of effective teaching and student achievement.

5 Phase 1: Value-added impact of teachers on student achievement overview

Phase 1 of this investigation identified more and less effective teachers by studying the value-added teacher's impact on student achievement (Raudenbaush and Bryk 2002). The research question articulated for this phase focused on studying to what degree do classroom teachers have a measurable effect on student achievement. Hierarchical Linear Modeling (HLM) was used to develop a classroom academic index (CAI) to identify teacher effectiveness using socio-economic status and student achievement data. Information was obtained from a large urban district database and coded to protect confidentiality of schools and teachers.

6 Method

Participants This examination took place in the 26th largest school district in the United States that serves nearly 100,000 students. The information used in this correlational study included data from 90 elementary schools from the district. Demographic data indicates a student population that is 51 % White, 36 % African-American, and 13 % other. Nearly 60 % of the students are from single-parent households and 61 % are identified as low-income families.

The data merged for this study included third-grade 2009–2010 and fourth-grade 2010–2011 Kentucky Core Content Test (KCCT) reading assessment scores into one data set. The population sample included 383 classrooms with 7,833 students. The data were analyzed and cases with incomplete data reduced the sample to 363 classrooms with 7,441 students. Further analysis of data removed cases that had an insufficient (a) number of students per classroom (14 or fewer students) and (b) number of instructional days (less than 100-day attendance). This reduced the population sample to 281 teachers and 6,962 students.

Instrumentation In this study, the Level 1 (student level) predictor variables included: (a) the third-grade KCCT scale score, and (b) socio-economic status (SES) of the student. The Level 2 (classroom level) predictor included the classroom average in SES. The mean scale scores came from the KCCT assessment of fourth-grade student scores that ranged from 400 to 480 (Kentucky Department of Education 2008). The test reliability of .88 (α=.88) derived from the mean scale scores were obtained from the tests' six versions administered to fourth-grade students (Kentucky Department of Education 2008).

The outcome variable for this study used the fourth-grade residual scores per classroom to establish a CAI. The statistical modeling approach (i.e., HLM)

facilitated comparisons of outcomes to determine and eliminate influences of SES and previous KCCT scores in reading to explain the variance within and between classrooms (Borman and Dowling 2010; Coleman et al. 1966; Raudenbush and Bryk 2002). Previous studies using HLM to identify teacher effectiveness have lent support for the measures used in this study. The student Level 1 variables in this model included free/reduced lunch status (a proxy for SES), and prior achievement as measured by third-grade KCCT 2009–2010 reading achievement scores for each student (Muñoz and Dossett 2001). The classroom Level 2 variable used the classroom average SES as a predictor.

Residual scores were calculated for each student and provided an indication of higher or lower than expected performance. The residual scores were the difference between third-grade KCCT reading scale scores and fourth-grade KCCT reading scale scores. Residual scores per classroom were averaged to standardized scores. Grand-mean centering was used for the variables in this analysis per recommendations derived from other researchers (Stronge et al. 2011). The 281 classrooms listed highest to lowest fourth-grade residuals to create a CAI. The 281 residuals were broken into thirds to differentiate the largest grouping of classrooms for analysis. This approach was similar to Goe (2007) and other researchers such as Geithman (2009), who separated groups in order to identify teacher effectiveness when related to student achievement.

Design and procedures HLM was used to estimate coefficients (predictors) for students in each classroom to predict the expected achievement residual for each student. In this analysis, use of student-level predictors at Level 1 and classroom-level predictors at Level 2 were applied. The student-level predictors included individual SES and third-grade KCCT reading mean scale scores. The classroom-level predictor at Level 2 included the average SES for each classroom.

Using suggestions from Raudenbush and Bryk (2002) as a guide, the following steps were used in the HLM analysis. First, a one-way ANOVA with random effects measured how much variation lies within and between classrooms. Second, a random coefficient regression in HLM used all Level 1 coefficients; Level 1 predictors randomly varied. The Level 2 predictors were unconditional variables set to conduct analysis of the SES and KCCT reading achievement relationship within the 281 classrooms. Next, an Intercepts-as-Outcomes HLM model was calculated. This measure helped explain the variability between classrooms, and indicated the association between SES and third-grade KCCT scores as stronger in some classrooms and not in others. Finally, residuals were calculated to allow classroom identification into more and less effective rankings.

6.1 HLM model 1: one-way ANOVA with random effects

This unconditional model was used to determine whether or not HLM was an appropriate analysis by separating the total variation in the outcome variable (Muñoz and Chang 2007). The intra-class coefficient (ICC) measured the proportion of variance in reading mean scale scores between classrooms. The following equation

represented the ICC calculation: $\hat{\rho} = \tau_{00}/(\sigma^2 + \tau_{00})$, where τ_{00} represented between-class correlation for the outcome measure on the reading mean scale score and σ^2 represented the within-class correlation for the sample. The equations below represented the unconditional means-as-outcome formula:

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

$$\text{Level 2}: \quad \beta_{0j} = \gamma_{00} + u_{oj}$$

In the Level 1 model, Y_{ij} is fourth-grade reading achievement or reading mean scale score for student "i" in classroom "j." β_{0j} is the intercept of the outcome, and r_{ij} represented random error for student "i" in classroom "j." In the Level-2 model, γ_{00} is the mean intercept for all classrooms in the population, and u_{0j} is the unique effect of classroom "j" on the mean intercept.

6.2 HLM model 2: random coefficient model

A random-coefficient regression model was conducted to assess the effects of individual student's SES status and prior achievement on the outcome variable and determined variance at the student level. Mean centering was used to ease levels of interpretation and to remove high correlations between first- and second-level variable interactions (Raudenbush and Bryk 2002). The general form of the model for the random coefficient analysis is shown below. The data analyzed reflected two Level 1 predictors: SES and third-grade reading score.

$$\text{Level 1}: Y_{ij} = \beta_{0j} + \beta_{1j}\left(X_{1ij} - 1_{ij}\right) + \beta_{0j} + \beta_{2j}\left(X_{1ij} - 1_{ij}\right) + r_{ij}$$

$$\text{Level 2}: \quad \beta_{0j} = Y_{00} + u_{0j} \qquad \beta_{1j} = Y_{10} + 1_j$$

In the Level 1 model, Y_{ij} represented the fourth-grade reading achievement scores for student "i" in classroom "j." β_{0j}, the intercept of the outcome and β_{1j} and β_{2j}, the coefficients of the outcome for the Level 1 predictors (individual student SES and prior achievement). X_{1ij} denoted the SES status for student "i" in classroom "j," $\overline{X}_{1 \cdot j}$, the mean SES for all classrooms in the population, and r_{ij} represented random error for student "i" in classroom "j."

6.3 HLM model 3: intercepts-as-outcomes model

The Intercepts-as-Outcomes model shown below determined the relationship between fourth-grade reading scale scores and two Level 1 factors; individual student SES and prior student achievement, while controlling for classroom mean SES at

Level 2 (Ballou et al. 2004; Willms 2010). This model identified how much variation existed by using SES and prior achievement.

$$\text{Level } 1 : Y_{ij} = \beta_{0j} + \beta_{1j}\left(X_{ij} - \overline{X}_{\cdot j}\right) + r_{ij}$$

$$\text{Level } 2 : \beta_{0j} = \gamma_{00} + \gamma_{01}(\text{mean SES}) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{mean SES}) + u_{1j}$$

In the Level 1 model, Y_{ij} represented the fourth-grade reading achievement or reading scale score for student "i" in classroom "j." β_{0j} represented the intercept of the outcome β_{1j}, the coefficient of the outcome for the Level 1 predictor for the individual student SES. X_{1ij} represented the SES status for student "i" in school "j," and $\overline{X}_{1j}$, the mean SES for all classrooms in the population. Grand-mean centering was used in the Level 2 equation to analyze classroom SES by subtracting individual classroom SES from the mean SES of the classrooms in this study. Next, grand-mean centering was used to remove high correlations between Level 1 and Level 2 and cross-level interactions. The proportion of variance was explained by the means-as-outcomes, random-coefficient, and Intercepts-as-Outcomes regression models while controlling for the predictor variables (Raudenbush and Bryk 2002).

6.4 Calculation of residuals

Several value-added studies of student achievement have used averaging of residuals to determine teacher effectiveness (e.g., Stronge et al. 2011). Averaging of all student residual scores for each of the 281 teachers in this study used data from the HLM analysis and measured an estimate of teacher impact on student achievement. Ranking of the individual teachers was based on measures referred to the CAI. This index was determined by ordering fourth-grade KCCT residual scores from highest to lowest to identify classroom teacher effectiveness for analysis in this study. The CAI was calculated by averaging all student residuals for the 281 classroom teachers. This procedure was similar to other value-added models that have used student achievement to obtain an average for each teacher (Bembry and Schumacker 2002). To calculate the residuals, the difference in performance for each student from the sample was compared to the student's fourth-grade KCCT score.

7 Results

The purpose of this study was to identify more and less effective teachers by measuring reading achievement using HLM. Data obtained from the large urban district regarding the sample population required removal of errors and verification of data. HLM models used in analysis of the data included: a random effects one-way ANOVA with the actual Level 1 model: $=\beta 0$ (grade 4 reading)$+R$, and Level 2 Model: $\beta 0 = G00$ (grade 3 reading)$+U0$. The random effects one-way ANOVA represented the null model reflecting how much variance existed among the variables as shown in Table 1.

Table 1 Random effects one-way ANOVA and the final estimation of variance components

Random effects		Standard deviation	Variance component	df	Chi-square	p value
INTERCEPT1	U0	8.71088	75.87950	280	1,988.9981	0.000
Level 1	R	17.67178	312.29179			

Null model (without influence of predictor variable)

The intra-class correlation (ICC) of .195 indicated approximately 20 % of the variance as among or between classrooms ($\hat{\rho}$ = 75.88/(75.88 + 312.29) = 75.88/388.17 = .195). The 20 % ICC indicated further analysis was warranted to determine more specificity between and within-classroom variances. This result validated the continued use of HLM for further analysis of the data (Raudenbush and Byrk 2002). A chi-square test was performed to inform variance and indicated significance at the $p < .01$ level existed among fourth-grade reading classrooms.

The random coefficient model represented the second HLM technique used to measure the lunch or SES variable. SES was found to be a strong predictor regarding proportion of variance on fourth-grade scale scores. The proportion of variance in student test scores explained by Level 1 variables (312.29–146.39/312.29) was 53 % (.531). Therefore, third-grade KCCT reading scores and free/reduced lunch status (SES) accounted for 53 % of the variation in student tests scores. The random coefficient model strongly predicted scores on the fourth-grade KCCT reading at 58 % (.576) and indicated both predictor variables (third-grade reading and SES) were significant in predicting fourth-grade KCCT performance of students.

The Intercepts-as-Outcomes model was used to calculate the intercepts average lunch and grade 3 reading scale score variables. This model combined the random-coefficient regression model at Level 1 and the means-as-outcomes model at Level 2. For intercept (means on fourth-grade reading) = β_{oj}, the proportion of variance explained (using the random coefficient variance as a comparison) was 84 % (82.48–13.53)/82.48 = 68.95/82.48 = 83.5). Therefore, 84 % of the variance in the fourth-grade achievement scores in reading was explained by the classroom mean third-grade reading and SES (free/reduced lunch). As shown in Table 2, the one-way ANOVA results explained the within- and between-classroom components.

SES and prior achievement were both significant predictors and indicated that these variables can be predictive regarding mean scores obtained on reading

Table 2 Random coefficient model and final estimation of variance components

Random effects		Standard deviation	Variance component	df	Chi-square	p value
INTERCEPT1	U0	9.08186	82.48017	261	3,987.10859	0.000
G3READ1 slope	U1	0.09796	0.00960	261	386.45303	0.000
LUNCH21 slope	U2	1.24967	1.56167	261	238.95781	>.500
Level 1	R	12.09935	146.39418			

The chi-square statistics reported are based on only 262 of 281 unites that had sufficient data for computation

achievement tests ($p<.01$). The Intercepts-as-Outcomes model at Level 2 is indicative of a positive correlation between SES and prior achievement. This positive relationship between SES and prior achievement to determine future achievement scores in this manner indicated plausibility. The benefit of the HLM model indicated an efficient statistical process to compare multiple variables at one time.

The final HLM model (see Table 3) was judged to be a useful representation of the data. The Level 1 residuals for each student were calculated, and a residual file was aggregated by classroom using the mean of the residuals for each classroom. The Level 2 predictor SES (free/reduced lunch) was centered on the grand mean. The highest residuals indicated over performance with regard to expected student achievement scores that resulted in higher rankings. Based upon the analysis of this data, the remaining unaccounted variance determined by the three HLM models regarding teacher effectiveness was linked to create the CAI.

For the purpose of this study, classroom residuals were organized from highest to lowest on the index. The residual sort represented 281 classrooms; this was the total of classrooms after excluding (a) classes of 14 or less students and (b) students with less than 100 instructional days). The CAI was differentiated into thirds. The top third classrooms ($n=94$, classrooms 1–94) and bottom third classrooms ($n=92$, classrooms189–281) represented the more effective and the less effective teacher components to examine. The residual mean score ($M=133.49$) for the more effective teachers when group averaged equaled $M=1.42$ overall. The residual mean score ($M=-159.54$) for less effective teachers when group averaged equaled $M=-1.73$. The residual mean score for the 6,962 students indicated $M=143.42$. The range of residual score means

Table 3 Intercepts and slopes as outcomes model

Final estimation of variance components:						
Random effects		Standard deviation	Variance component	*df*	Chi-square	*p* value
INTERCEPT1	U0	3.67787	13.52674	259	827.47221	0.000
G3READ1 slope	U1	0.09420	0.00887	261	385.90590	0.000
LUNCH21 slope	U2	1.16644	1.36058	261	238.32664	>.500
Level 2	R	12.09935	146.39418			
Final estimation of fixed effects:						
Fixed effect		Coefficient	Standard error	*T* ratio	Approx. *df*	*p* value
For INTERCEPT1, β0						
INTERCEPT2	G00	446.278622	0.263255	1,695.232	278	0.000
LUNCH22	G01	6.806683	1.546361	4.402	278	0.000
G3READ1	G02	0.768222	0.045086	17.039	278	0.000
For G3READ1 slope, β1						
INTERCEPT2	G10	0.658381	0.010119	65.061	280	0.000
For LUNCHR21 slope, β2						
INTERCEPT2	G20	3.795988	0.374557	10.135	280	0.000

The chi-square statistics reported are based on only 262 of 281 unites that had sufficient data for computation. Fixed effects and variance components are based on all the data

indicated 9.93 and 145.15 for the more and less effective groups, respectively. A significant difference was found between the less effective groups when compared to the overall residual means.

The classroom characteristics generated through the HLM models identified the residual scores for the 281 classrooms. The range of residuals indicated a high of 3.57 for more effective teachers and a low of −4.53 for the less effective teachers. The average residual mean for the top third more effective teachers was 1.42. The average residual mean for the bottom third less effective teachers was −1.73. These differences resulted in a 3.15 residual point difference among the two groups.

Descriptive statistics for the residual sort yielded an average fourth-grade reading achievement residual mean of 453.64 for more effective teachers and 433.38 for less effective teachers. The Level 1 third-grade reading yielded a mean of .01 and a standard deviation of 18.32 points. The fourth-grade mean score yielded 446.60 and a standard deviation of 19.72. A wider dispersion of variation existed among the fourth-grade residuals than the overall group. However, the difference in standard deviation (SD=1.4) between the third-grade group and fourth-grade group reflected a small to moderate significance level at $p<.01$.

Upon further examination of the residual sort, the top five and bottom five classroom data were compared (Table 4). The sample population was taken from the residual sort to indicate the significance found between the more effective and less effective teachers. For example, classroom 346 had a residual of 3.57. As the residual sort was structured from highest to lowest, this meant that classroom 346 had the highest residual and represented the more effective teachers. The average mean of the top five teachers indicate $M=3.30$ residual mean and $M=453.64$ on the fourth-grade reading achievement mean.

The bottom residual for less effective teachers was −4.53 and associated with classroom 180. This meant that classroom 180 had the lowest residual and were

	Case number	Classroom ID	Residual mean	Grade 4 mean
Table 4 Comparative analysis of residuals of the top five more effective teachers vs. the bottom five less effective teachers on reading achievement	Case summaries sampling—top five "more effective" teachers			
	1	346	3.57	460.39
	2	256	3.45	442.95
	3	157	3.34	444.19
	4	306	3.10	454.42
	5	121	3.06	466.24
	Average		3.30	453.64
	Case summaries sampling—bottom five "less effective" teachers			
	277	340	−3.66	435.11
	278	33	−3.69	435.18
	279	60	−4.34	430.94
	280	337	−4.38	428.45
	281	180	−4.53	437.20
	Average		−4.12	433.38

considered the less effective teachers. The bottom five classrooms average residual mean indicated $M=-4.12$ and $M=433.38$ fourth-grade achievement score mean. The differences between the 281 classrooms and 6,962 students versus the 10 classrooms and approximately 200 students representing 1 % of the sample allowed comparison of the extreme ranges found in the CAI. The mean scores obtained through closer examination indicated a stark difference exists between highly effective and less effective teachers for this sample; this, in turn, could lead to an inference that targeting of less effective teachers to receive additional support to improve student achievement outcomes may be warranted, particularly if the individual teacher effect would be consistent across multiple years (Munoz et al. 2011).

8 Discussion

The use of HLM provided the results for the research question for Phase 1 (i.e., "To what degree do classroom teachers have a measurable effect on student achievement?"). The three HLM models used in this study provided a logical and systematic approach to isolate predictor variables and determine teaching effectiveness. The range of residuals from 3.57 to −4.53 reflected a significant difference of the aggregated residual points. SES and prior achievement using third-grade KCCT scores represented 84 % of the remaining variance determined by the HLM models. The predictor variables of third grade KCCT reading scale scores and SES were found to be significant and allowed the remaining variance to be linked to teacher effectiveness via a residual file.

The residual file created through HLM was aggregated by classroom using the mean of the residuals for each classroom and grand mean centering for the SES predictor variable. The classroom level predictor results indicated a significant difference among the two groups of teachers identified as more effective and less effective. The three HLM models used in this study provided a logical and systematic approach to isolate predictor variables and determine teaching effectiveness. The wide range in residuals reflected a significant difference among the aggregated residual points. SES and prior achievement using third-grade KCCT scores represented 84 % of variance. The remaining variance determined by the HLM models provided the basis to analyze teacher effectiveness. The predictor variables of third-grade KCCT reading scale scores and SES were found to be significant and allowed the remaining variance to be linked to teacher effectiveness via a residual file. The residual file created through HLM was aggregated by classroom using the mean of the residuals for each classroom and grand mean centering for the SES predictor variable. The classroom level predictors were similar to those in Stronge et al.'s (2011) study. Finally, the resultant residuals obtained in this study were used to form a CAI similar to the Teacher Academic Index and trichotomizing classrooms (i.e., three groupings) based on teacher effectiveness and as outlined in this phase of the study.

9 Conclusion

The practicality of this approach incorporated simple HLM models using two accepted predictor scores: SES and prior achievement. HLM allowed analysis of multiple variables at one time to determine within-groups and between-groups variance (Coleman et al. 1966; Goe 2007). In part, the findings of this study replicated Stronge et al. study (2011). Although this step is foundational, all of Phase 1 is merely background to the main focus of the study—studying the "black box" of effective teaching. A caveat that needs to be considered is to what degree the residuals are reliable measures of teacher effectiveness since there was no corroboration from other data or from multiple years in this particular study (Munoz et al. 2011).

10 Phase 2: comparison of teachers' perception on effectiveness characteristics

10.1 Overview

Phase 2 of this investigation collected perception data about the characteristics of effective teaching by using a survey developed on the basis of Stronge's (2007) theoretical framework. Phase 2 helped explore the relationship between perceptions of effective teacher qualities and student achievement.

11 Method

Participants The data source consisted of approximately 380 reading teachers in classrooms (meeting criterion discussed in the method section for Phase 1) across 90 elementary schools in the 26th largest district in the nation. The entire population of fourth-grade teachers was targeted in order to obtain an adequate statistical power sample size to reach valid, reliable, and useful conclusions to inform the field of education. The survey administration began in early 2012 and continued until achieving a stratified random sampling (Trochim and Donnelly 2008) of approximately 29 teachers in each effectiveness group (i.e., more effective, less effective). This involved dividing our population into homogeneous subgroups and then taking a simple random sample in each subgroup; stratified random sampling has more statistical precision than simple random sampling if the strata or groups are homogeneous. Targeted teachers were asked, based on their knowledge and experience, to rank order teacher qualities from one to a maximum number of eight (1 = strongest impact on student achievement; 8 = lowest impact on student achievement) across five of the six domains. The prerequisites for teachers, included in section two of the survey, were used as moderating variables to address influences of participants' characteristics (e.g., gender, years of experience teaching, years of experience teaching reading, ethnicity, and level of education) on perceptions. In addition, another moderating school designation variable was included (Title I versus non-Title I).

From a statistical power perspective, distributions of surveys continued until achieving a minimum sample size of 29 for each level (highest CAI=1 and lowest CAI=2) of the independent variables. The sample size was obtained using Table C12 (Hinkle et al. 2003, p. 654) with a power of .80, two levels of treatment levels (k), population error variance of .75 σ and alpha level of .05. The actual sample size of participants in phase II exceeded the minimum requirement ($N=76$).

Instrumentation A synthesis of research-based effective teacher qualities affecting student achievement provided the framework for the survey administered in Phase II (Stronge 2007). Information was collected using Williams' survey (2010), which in turn was based on Stronge's meta-review of qualities of effective teachers (2007). In the survey, fourth-grade reading teachers in a large urban district were asked to rank five qualities of effective teachers based on their perceptions of qualities affecting student achievement; also, participants also ranked teacher behaviors serving as indicators of those qualities. The survey consisted of four sections. First, participants ranked indicators of teacher quality in five general categories (a) classroom management and organization; (b) planning for instruction; (c) implementing instruction; (d) monitoring student progress; and (e) teacher as a person. Second, participants ranked these five categories. Third, respondents listed any additional indicators of qualities not represented in the survey. Fourth, teacher/school characteristics information were requested: (a) gender; (b) years of experience; (c) years of experience teaching reading; (d) education level; (e) ethnicity; and (f) school designation.

Design and procedures Phase 2 focused on 2010–2011 fourth-grade teachers and collected survey responses to answer the following: Are there differences between teacher perceptions of effective teacher qualities in classrooms identified as effective and less effective? The Cronbach's alpha was used to determine item-total correlations verifying reliability of the survey. Surveys completed were matched with Phase I data ($N=121$) and grouped accordingly: (a) more effective ($n=44$); (b) effective ($n=37$); and (c) less effective ($n=32$). Most self-contained special education teachers in the district teach less than 15 students; therefore, due to sample size issues associated with value-added methodology, special education teachers in self-contained classrooms were not included in the survey sample ($n=8$). This reduced the total number of respondents from 121 to 113. Furthermore, the primary focus of the Phase II analysis consisted of 76 survey respondents (i.e., 37 effective teachers were excluded) from the two comparison groups (i.e., more effective teachers, less effective teachers). Descriptive statistics were used to analyze survey data linked to the top third and bottom third teachers based on the CAIs.

Based on their knowledge and experience, teachers were asked to rank order indicators of teacher quality impacting student achievement. The number of items in each category determined the range of numbers used for ranking purposes with eight items being the maximum number of items in any category (e.g., strongest impact=1; lowest impact=8). Mean ranks were used to ascertain the indicators of quality teachers perceived having the greatest influence on student achievement.

Due to the ordinal nature of the survey items, the nonparametric analog to ANOVA Kruskal–Wallis test statistic was used to determine if one group was different from at least one other group (Hinkle et al. 2003). The survey rank ordered items relating to

teacher characteristics influencing student achievement were coded and were compared using SPSS.

12 Results

The most statistically significant finding was in the Classroom Management and Organization category with the item "Maintains a physically and emotionally safe environment for students." Teachers in the more effective group rank ordered this item as being more important than less effective teachers. A cross-tabulation was performed for this survey item revealing more than half (52.5 %) of the more effective teachers ranked this item as having the most importance.

After reviewing survey item results showing statistically significant differences between groups on the Kruskal–Wallis test, an ANOVA was used to further explore these items using a parametric test comparing survey rankings of respondents from three teacher groups (more effective, effective, and less effective). The homogeneity of variance was non-significant for this research; therefore, since this was a normal distribution and there was independence of observation, the ANOVA procedure was considered robust with respect to the violations of assumption for the purposes of this research (Hinkle et al. 2003).

Again, results demonstrated a significant difference in the Classroom Management and Organization category, specifically with survey item "Maintains a physically and emotionally safe environment for students" between more effective and less effective teachers [$F(2,110)=4.08$, $p<.05$]. Therefore, ANOVA results helped validate findings using the Kruskal–Wallis test. A Tukey's HSD was used to determine the differences between teacher groups. This analysis revealed teachers who were not effective rated this survey item not as important to student achievement ($M=2.34$, SD$=1.24$) than the more effective teachers ($M=1.84$, SD$=1.08$). In summary, the more effective teachers placed higher value on the survey item "Maintains a physically and emotionally safe environment for students" than the less effective teachers.

The second most statistically significant area demonstrating differences between groups was the Planning for Instruction category, survey item "Limits interruptions and focuses classroom time on teaching and learning" [$F(2,110)=3.72$, $p<.05$]. The Tukey's HSD revealed differences between more effective teachers ($M=3.84$, SD$=1.86$) and less effective teachers ($M=2.88$, SD$=1.91$). The importance of classroom management and organization is closely linked to this item found in the *Planning for Instruction* category. This seems to reinforce the finding of how classroom management matters for the more effective teachers. As shown on Table 5, teachers ranked Classroom Management and Organization domain as having the most influence on student achievement and monitoring student progress as the least influence on student achievement.

13 Discussion

Phase 2 builds on both Stronge's (2007) and Williams' (2010) previous work. Stronge (2007) identified research-based characteristics of effective teacher

Table 5 Survey respondents mean score

Survey item	More effective	Effective	Less effective	Total
Classroom mgt. & organization (1 = highest priority to 4 = lowest priority)				
Physically and emotionally safe environment	1.84	2.57	2.34	2.22
Order and routines	2.34	2.24	1.91	2.19
Preparation	2.70	2.41	2.16	2.45
Discipline	2.84	2.46	2.66	2.66
Planning for instruction (1 = highest priority to 6 = lowest priority)				
High expectations	2.39	2.14	2.00	2.19
Links instruction to objectives	2.82	3.08	3.28	3.04
Considers student learning styles	2.89	2.68	2.59	2.73
Limits interruptions and focuses class time	3.84	4.00	2.88	3.62
Relevance	3.91	3.92	3.53	3.81
Pacing	4.56	4.54	3.88	4.36
Implementing instruction (1 = highest priority to 6 = lowest priority)				
Engagement	1.98	1.70	2.00	1.89
Variety	2.68	2.27	2.23	2.42
Guided practice	2.84	3.35	2.60	2.95
Higher-order skills	3.59	3.92	3.17	3.59
Variety	4.23	4.35	4.03	4.22
Grouping	5.02	4.89	4.43	4.82
Monitoring student progress (1 = highest priority to 5 = lowest priority)				
Assessment	2.39	2.57	2.41	2.45
Using data	2.45	2.54	2.44	2.48
Re-teaching	2.57	2.16	2.34	2.37
Feedback.	2.75	2.76	2.44	2.66
Homework	4.41	4.65	3.94	4.35
Teacher as a person (1 = highest priority to 8 = lowest priority)				
Excitement	2.98	2.84	2.69	2.85
Interaction	3.20	3.59	2.53	3.14
Concern	3.23	4.05	3.38	3.54
Fairness	3.64	4.43	3.56	3.88
Commitment	3.80	3.70	3.63	3.72
Content knowledge	4.80	4.62	4.25	4.58
Respect	5.43	5.22	4.63	5.13
Reflection	6.32	6.16	5.50	6.04
Teacher qualities (1 = highest priority to 5 = lowest priority)				
Classroom management and organization	1.95	2.05	2.28	2.08
Implementing instruction	2.52	2.89	2.19	2.55
Planning for instruction	2.75	2.81	2.47	2.69
Teacher as a person	3.43	3.32	3.03	3.28
Monitoring student progress	3.66	3.57	3.28	3.52

categories and associated items used as the foundation of Williams' (2010) survey. This study expands Williams' work by examining student achievement linked to survey results. Williams' study found a general agreement among teachers and administrators, and our study found a similar agreement between more and less effective teachers. In general, since only two differences are the main findings of this study, this indicates that there was actually a high level of a agreement between effective and less effective teachers, more than we originally hypothesized. This is an interesting finding by itself and probably cautions the use of single metrics to classify more and less effective teachers. This calls to the need of using multiple measures for teacher evaluation. Another possibility could be that the teacher effectiveness variables may not be highly robust.

The most significant finding of this research was the item "Maintains a physically and emotionally safe environment for students" contained in the Classroom Management and Organization category. The more effective teachers ranked this indicator as having greater importance than less effective teachers did. Similarly, Williams' (2010) research found administrators and teachers both ranked "Maintaining a physically and emotionally safe environment for student" as the most important component of classroom management and organization.

Previous research helped understanding of the findings. As far back as the mid-twentieth century, Maslow articulated one of the most well-known theories of human motivation. Maslow (1943) identified a hierarchy of needs beginning with the physiological needs of food, shelter, and clothing with the highest human priority. Personal safety including physical and emotional needs is the second highest foundation of human motivation (Maslow 1943). Since the human brain helps ensure our survival by maintaining safe conditions and avoiding danger, fulfilling a child's basic needs is critical before the child's mind is capable of learning (Jensen et al. 2006). Blum (2005) also recognized physical and emotional safety as a critical requirement for students indicating unsafe schools or poorly managed classrooms cannot provide a stable learning environment. In fact, results show children's perceived exposure to violence has a significant negative impact on their reading performance (Nettles et al. 2000). Although this may seem like a soft approach in an era of increased accountability, school connectedness (including physical and emotional safety) can impact student achievement (Blum 2005). Similarly, results of this study indicate more effective teachers recognize how important it is to maintain a physically and emotionally safe environment for students.

Another significant finding was "Limits interruptions and focuses classroom time on teaching and learning" contained in the Planning for Instruction category. The more effective teachers ranked this indicator as having less importance than the less effective teachers indicated. Although focusing classroom time on teaching and learning is important, effective teachers also tend to emotional needs and anticipate disturbances spending less time correcting disturbances (Hattie 2003; Stronge 2007).

Teachers are frustrated by increasing student achievement demands and regular classroom interruptions (Leonard 2001). The erosion of instructional time due to interruptions is not a new concept; however, Leonard (2001) found evidence to suggest intrusions on instructional time is largely contextual in the way teachers respond. Furthermore, educators who are aware of these conditions of interruptions

will not disregard the environment but plan and adapt accordingly (Leonard 2001). Since the more effective teachers in this survey ranked this indicator differently than less effective teachers, one may speculate the more effective teachers may view interruptions to be innocuous.

Regardless of whether the value-added measure is a valid indicator of effectiveness, there is no question that the low-performing teachers are failing to raise test scores and are likely feeling more stress. Thus, these teachers might feel more pressure to obtain immediate results rather than attend to social–emotional learning and other indirect precursors of student achievement. In addition, a question remains about as to what degree school variables might influence results, like having an ineffective principal leading the school reform effort or the effects of school climate.

14 Conclusion

Taken together, these findings support the notion that effective teachers focus on meeting students' basic physical and emotional needs understanding that if these are not met the students' brains are not likely to engage in cognitive thinking. Likewise, less effective teachers place a greater value on limiting interruptions and focuses classroom time on teaching and learning perhaps at the detriment of recognizing the importance of addressing basic needs. One cannot assume that each child comes to school every day with a mindset fully prepared to learn subject matter knowledge.

This study also analyzed the influence of teaching experience, experience teaching reading, education level, gender, and ethnicity on teacher effectiveness and teacher perceptions of quality attributes. However, the analysis revealed none of these teacher characteristics showing significance in improving teacher effectiveness for the fourth-grade teachers participating in this study. The Tukey HSD post hoc analysis also revealed no statistically significant differences.

15 General discussion

This study has important practical implications for the teacher preparation programs in colleges and universities as well as for the human resource functions in school districts as we think about teacher quality and student achievement (Darling-Hammond 2000). It is relevant to identify the knowledge, skills, and predispositions that pre-service teachers need. In this particular study, the authors collected empirical evidence about the importance of classroom management courses in teacher preparation programs. Furthermore, this study has multiple practical implications for school districts in the areas of teacher (a) recruitment, (b) selection, (c) induction, (d) compensation, (e) in-service professional development, and (f) evaluation.

16 Implications for practice

The methodology used to rank teacher effectiveness offers district leaders and policy makers an opportunity to design teacher professional development aimed to improve

student achievement. District leaders could differentiate/customize professional development according with teachers ranking by the CAI. Every school could rank each grade level using district assessments as well as achievement test results. This ranking could also be helpful to identify teacher leaders within schools who could mentor less effective teachers. Moreover, linking student achievement outcomes to classroom teachers may serve to identify students in need of differentiated instruction, intervention, or remediation. This information may assist the teacher in meeting individualized student needs in order to promote learning and raise student achievement scores.

Furthermore, by focusing on teacher effectiveness, teacher pre-service programs may strengthen teacher's pedagogical skills on building positive relationships with students. Teachers' priority would be ensuring a physically and emotionally safe learning environment to facilitate academic learning. Strategies may include: enforcing a zero tolerance towards bullying, teaching emotional intelligence skills, discussing threat of school violence, and maintaining a caring attitude (Jensen et al. 2006). Better understanding of effective teacher characteristics may assist in selection of teachers, teacher retention, and teacher development. School improvement efforts cannot focus on teacher prerequisites alone to improve student outcomes. Research indicates teachers' education and experience affect student achievement, but the effectiveness of experience levels off after 5 years (Nye et al. 2004). Teacher application forms may provide information by asking specific questions pertaining to effective teacher characteristics. This process would allow school administrators who are responsible for (a) hiring, (b) retaining, and (c) developing effective teachers to recognize effective teacher qualities in both prospective and incumbent teachers. Professional development offerings may focus on effective teacher practices resulting in improved student outcomes. Finally, communicating effective teacher qualities may provide a teacher self-reflection tool used to increase teacher's own awareness to personal growth and improvement of student outcomes.

17 Limitations and recommendations for future research

In this study, not all teachers were included in the CAI due to insufficient data (e.g., 14 or less students and less than 100 attendance days) to calculate residuals. The resultant 262 classrooms represented 93 % of the classrooms. While the percentage was strong enough to use in this study, making a determination on every teacher's effectiveness is beyond the focus of this research. Even though this study used HLM to identify teacher effectiveness related to student achievement scores, the sole use of this type of data is not suitable for teacher evaluation purposes.

This study considered data from only 1 year and other factors beyond teacher effectiveness may have been attributable to student achievement. Stronge et al. (2011) suggested longitudinal studies along with other measures such as teacher observations to determine teacher effectiveness for evaluation purposes. Several researchers (Stronge et al. (2011)) have warned that the residual index developed in their studies did not necessarily reflect teacher's effectiveness. A low-performing classroom or school could show large positive gains in achievement scores while still being an underperforming school. Thus, this study offers the use of value-added methodology to differentiate classroom performance levels. Value-added data can help make

informed decisions to improve teaching and learning processes; however, it should not be used as a single source to decide on retaining or dismissing teachers.

More inclusive approaches to evaluate teacher effectiveness have been proposed. Some of them contemplate using (a) four domains of professional practice along with value-added outcomes (Danielson 2007) or (b) six-domain rubric of teacher performance (Marshall 2009). These domains include planning and preparing for instruction, classroom management, delivery of instruction, monitoring, assessment and follow-up, family and community outreach, and professional responsibilities (Marshall 2009).

In addition, Stronge's (2007) framework identifying effective teachers incorporates many elements expressed by Danielson (2007) and Marshall (2009). This framework offers a more inclusive and balanced way to incorporate value-added data in identifying effective teachers. Future research may consider inclusion of all students (e.g., special needs, English language learners) in elementary schools; this would offer an opportunity to better understand the "black box" of effective classrooms for diverse students.

The presence of increased accountability at classroom level and implementation of high-stakes testing supports the need of more in-depth studies on the complexities of teacher effectiveness (Goe 2007; Stronge et al. 2011; Williams 2010). It is recommended to include a qualitative component by interviewing more effective and less effective teachers to gain a better understanding of survey responses. In addition, this study could be expanded by including classroom observations to provide valuable information about what happens in the classrooms. These classroom observations can be triangulated with teacher perception and value-added data.

There is no perfect research. This section raises some concerns about simply using HLM and value-added methods to categorize teachers. As a result, it suggests that the authors could be overinterpreting the survey findings to some degree. First, they assume that the effectiveness categories are valid and, second, they are isolated findings that could have emerged by chance.

18 General conclusion

This study successfully answered to what degree the sample classroom teachers had a measurable effect on student achievement in a given year. The development of a CAI using residual scores helped link student achievement scores to classroom teacher effectiveness. This investigation first identified teachers who were successful and unsuccessful in producing evidence of learning gains beyond expectations. Then, the investigation collected survey data of these two groups of teachers to assess differences in their perceptions of what constitutes attributes for effective teaching. With this investigation, after analyzing empirical data, the authors attempted to look inside the "black box" of classrooms (Brophy and Good 1986; Sanders and Rivers 1996; Webster and Mendro 1997) by asking one of the primary actors (i.e., teachers) about what constitutes good teaching. The authors responded to the call for more research beyond variance decomposition models (like HLM) that estimate the random effects of classrooms on student achievement (Munoz and Chang 2007; Rowan et al. 2002). Even though information gleaned from this study offered a way to identify teachers for

possible professional development aimed to raise student achievement scores in reading, it was not intended nor recommended to evaluate teachers participating in this study.

This research reported significant differences between the perceived importance of the more effective teachers and the less effective teachers relating to the survey item "Maintains a physically and emotionally safe environment for students." Consistent with previous research (Blum 2005; Jensen et al. 2006; Maslow 1943), more effective teachers consider that basic needs need to be met before students brains are ready to learn. Conversely, less effective teachers place a significant higher importance on the survey item "Limits interruptions and focuses classroom time on teaching and learning" than more effective teachers do. Since eliminating interruptions is not feasible, more effective teachers recognize this environment planning and adapting accordingly. This new study, along with the Williams' (2010) and the Stronge et al. study (2011), is indicating a strong focus on the importance of classroom management. In general, researchers might believe that instructional practices (e.g., differentiation, higher order thinking) would be a more pronounced teacher effectiveness characteristic. By no means, this set of studies is saying that instructional practices are unimportant. Rather, studies seem to be showing that classroom management remains a mainstay of effective teachers. It is not possible to effectively teach if classroom management is not in place first.

The findings of this study also align with recent results from surveys of K-12 students covering a range of classroom characteristics linked to teacher quality and how these surveys can successfully predict student achievement (Ferguson 2012). The measures of teaching quality in the Tripod surveys are gathered under seven headings called 7Cs: Control, Care, Clarify, Challenge, Captivate, Confer, and Consolidate. These attributes are grounded in the work of many educational researchers over several decades and capture what is considered important in determining how well teachers teach and how much students learn.

Control, the main attribute aligned with the findings of this study, pertains to classroom management (Ferguson 2012). Teachers need skills to manage student propensities toward off-task or even disruptive behavior in order to foster conditions that allow for effective teaching. Control helps to maintain order and supplements caring by making the classroom calm and emotionally safe. Well aligned with our current study, Ferguson (2012) found that Control is the strongest predictor of value-added achievement gains. In particular, these are the three Control items that most strongly predict gains in student learning: (a) students in this class treat the teacher with respect, (b) my classmates behave the way my teacher wants them to, and (c) our class stays busy and doesn't waste time. These items are similar to our items associated with classroom management: (a) maintains a physically and emotionally safe environment for students and (b) limits interruptions and focuses classroom time on teaching and learning.

Our findings based on teacher perceptions as well as Ferguson (2012) results based on student perceptions suggest that the highest-achieving classrooms (as measured by value-added methodologies) are respectful and orderly environments, with students who stay busy and learn to focus classroom time on teaching and learning. This article raises questions about HLM and value-added methods to accurately identifying effectiveness. HLM data need to be used in combination with multiple measures (e.g., peer observations, student surveys, and principal ratings). Furthermore, multiple

years of value-added scores are better than a single year so that results can be triangulated. Multiple indicators collected in various occasions are the best way to measure teacher effectiveness. Value-added measures cannot be the sole source of information in any teacher evaluation system, although these measures do play an important role if we want to value student learning gains as a source of information.

Our findings also indicate that most teachers (whether having high or low value-added scores) regard the importance of various teaching skills much the same way. This is not surprising. Teaching is a complex science and art that include multiple domains. We need to be realistic about the complexity of teaching-and-learning, particularly if we consider that different cohorts of students might need different approaches to teaching-and-learning. We need to be flexible enough and adjust our teaching "tool box" to adapt to changing cohorts of students.

To make a difference in the quality of education, we need effective teachers in every classroom. We need to learn what good teachers do that actually produce student learning gains beyond expectations. This investigation is a contribution to this critical area of research by asking the more and the less effective teachers what effective teaching is about. Teachers and students are the closest to the teaching-and-learning action—we need to uncover the "black box" of actual student learning by studying the real actors. We need to incorporate teachers and student voices in the important debate around teacher effectiveness and evaluation. Teacher evaluation systems need to focus on improving practice in addition to just measuring student performance.

References

American Recovery and Reinvestment Act (2009). *Race to the top program executive summary*. Retrieved from www.ed.gov/recovery.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessments for teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–65.

Bembry, K., & Schumacker, R. (2002). Establishing the utility of a classroom effectiveness index as a teacher accountability measure. *Journal for Effective Schools, 1*(1), 61–79.

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.

Berliner, D. C. (2005). The near impossibility of testing teacher quality. *Journal of Teacher Education, 56*, 205–231. doi:10.11770022487105275904.

Blum, R. W. (2005). A case for school connectedness. *Educational Leadership, 62*(7), 16–20.

Borman, G., & Dowling, M. (2010). Schools and inequality: a multilevel analysis of Coleman's equality of educational opportunity data. *Teachers College Record, 112*(5), 1201–1246.

Brophy, J., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York, NY: Macmillan.

Burchinal, M. R., Peisner-Feinberg, E., Pianta, R., & Howes, C. (2002). Development of academic skills from pre-school through second grade: family and classroom predictors of developmental trajectories. *Journal of School Psychology, 40*, 415–436. doi:10.1016/S0022-4405(02)00107-3.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity*. Washington, DC: Government Printing Office.

Cotton, K. (2001). *Expectations and student outcomes*. Retrieved from http://www.nwrel.org/scpd/sirs/4/cu7.html.

Danielson, C. (2007). *Enhancing professional practice: a framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

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.

Early, D. M., Maxwell, K. L., Burchinal, M., Bender, R. H., Ebanks, C., Henry, G. T., et al. (2007). Teachers' education, classroom quality, and young children's academic skills: results from seven studies of preschool programs. *Child Development, 78*(2), 558–580.

Ellett, C. D., & Teddlie, C. (2003). Teacher evaluation, teacher effectiveness and school effectiveness: perspectives from the USA. *Journal of Personnel Evaluation in Education, 17*(1), 101–128.

Emmer, E. T., Evertson, C. M., & Worsham, M. E. (2003). *Classroom management for secondary teachers*. Boston, MA: Allyn & Bacon.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24–28.

Geithman, B. W. (2009). *Examining principal perceptions, and teacher and school effectiveness through a value-added accountability model* (Doctoral dissertation). Retrieved from ProQuest database. (UMI No. 3355255).

Goe, L. (2007). *The link between teacher quality and student outcomes: a research synthesis*. Retrieved from http://secc.sedl.org/orc/resources/LinkBetweenTQandStudentOutcomes.pdf.

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

Hamre, B. K., & Pianta, R. C. (2005). Can instruction and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949–967.

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

Hanushek, E., & Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics, 4*, 131–157.

Hattie, J. (2003). *Teachers make a difference: what is the research evidence?* Retrieved from http://www.aiceonline.com/Resources/RC2003_HattieTeachersMakeA-Difference.pdf.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*. doi:10.3102/003465430298487.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston, MA: Houghton Mifflin Company.

Jensen, E., Dabney, M., Markowitz, K., & Selsor, K. (2006). *Environments for learning*. Cheltenham:VIC: Hawker Brownlow Education.

Kentucky Department of Education (2008). *Kentucky core content test*. Retrieved from www.education.ky.gov.

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.

Kyriakides, L., Demetriou, D., & Charalambous, C. (2006). Generating criteria for evaluating teachers through teacher effectiveness research. *Educational Research, 48*(1), 1–20.

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

Leonard, L. J. (2001). From indignation to indifference: teacher concerns about externally imposed classroom interruptions. *The Journal of Educational Research, 95*(2), 103–109.

Lumpkin, A. (2007). Caring teachers: the key to student learning. *Kappa Delta Pi Record, 43*(4), 158–160.

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.

Marshall, K. (2009). *Rethinking teacher supervision and evaluation: how to work smart, build collaboration, and close the achievement gap*. San Francisco, CA: Jossey-Bass.

Marzano, R. J. (2003). *What works in schools*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2007). *The art and science of teaching: a comprehensive framework for effective instruction*. Alexandria, VA: Association for Supervision and Curriculum Development.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review, 50*, 370–396.

Muñoz, M. A. (2005). *Black box. Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

Muñoz, 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. doi:10.1007/s11092_008_9054-y.

Muñoz, M. A., & Dossett, D. (2001). Equity and excellence: the effect of school and socio-demographic variables on student achievement. *Journal of School Leadership, 11*, 120–134.

Muñoz, M. A., & Vanderhaar, J. E. (2006). Literacy-embedded character education in a large urban district: effects of the child development project on elementary school students and teachers. *Journal of Research in Character Education, 4*(1&2), 27–44.

Muñoz, M. A., Prather, J. R., & Stronge, J. H. (2011). Exploring teacher effectiveness using hierarchical linear models: student- and classroom-level predictors and cross-year stability in elementary school reading. *Planning and Changing, 42*(3–4), 241–273.

Nettles, S. M., Mucherah, W., & Jones, D. (2000). Understanding resilience: the role of social resources. *Journal of Education for Students Placed at Risk, 5*(1&2), 47–60.

Noddings, N. (2001). Care and coercion in school reform. *Journal of Educational Change, 2*(1), 35–43. doi:10.1023/A:1011514928048.

Noddings, N. (2006). Educational leaders as caring leaders. *School Leadership and Management, 26*(4), 339–345.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257. doi:10.3102/01623737026003237.

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

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

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2008). *Can you recognize an effective teacher when you recruit one?* Cambridge, MA: National Bureau of Economic Research.

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., & 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.

Stiggins, R. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan, 80*(2), 22–28.

Stronge, J. H. (2007). *Qualities of effective teachers* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

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

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. doi:10.1007/s11092-008-9053-2.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A cross-case analysis of the connections between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*(4), 339–355. doi:10.1777/0022487111404241.

Teddlie, C., & Liu, S. (2008). Examining teacher effectiveness with differentially effective primary schools in the People's Republic of China. *School Effectiveness and School Improvement, 19*(4), 387–407.

Teddlie, C., & Reynolds, D. (2005). Contemporary issues in school effectiveness research, with emphasis on applications (Part one). *Educational Administration Research, 5*, 22–33.

Trochim, W. M., & Donnelly, J. P. (2008). *The research methods knowledge base* (3rd ed.). Mason, OH: Atomic Dog/Cengage Learning.

Webster, W., & Mendro, R. (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.

Wenglinsky, H. (2000). *How teaching matters: bringing the classroom back into the discussion of teacher quality.* Princeton, NJ: Educational Testing Service.

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

Williams, R. E. (2010). *Administrator and teacher perceptions of the qualities of effective teachers* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3438556).

Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teachers College Record, 112*(4), 1008–1037.

Zins, J. E., Weissberg, R. P., Wang, M. C., & Walberg, H. J. (2004). *Building academic success on social and emotional learning: what does the research say?* New York, NY: Teachers College Press.