Educ Asse Eval Acc (2009) 21:209–234
DOI 10.1007/s11092-009-9074-2

Is experience the best teacher? A multilevel analysis of teacher characteristics and student achievement in low performing schools

Francis L. Huang · Tonya R. Moon

Received: 23 September 2008 / Accepted: 30 April 2009 /
Published online: 26 May 2009

Abstract The study investigated several teacher characteristics, with a focus on two measures of teaching experience, and their association with second grade student achievement gains in low performing, high poverty schools in a Mid-Atlantic state. Value-added models using three-level hierarchical linear modeling were used to analyze the data from 1,544 students, 154 teachers, and 53 schools. Results indicated that traditional teacher qualification characteristics such as licensing status and educational attainment were not statistically significant in producing student achievement gains. Total years of teaching experience was also not a significant predictor but a more specific measure, years of teaching experience at a particular grade level, was significantly associated with increased student reading achievement. We caution researchers and policymakers when interpreting results from studies that have used only a general measure of teacher experience as effects are possibly underestimated. Policy implications are discussed.

Keywords Teacher effects · Teacher quality · Reading · HLM · Certification · Teacher experience · Low performing schools

The question of what characteristics make an effective teacher continues to be heavily debated. In recent years, with *No Child Left Behind's* (NCLB) push for highly qualified teachers, the question becomes even more imperative. For policy makers looking at ways to improve K-12 education, one approach has been to focus on teachers as they

F. L. Huang · T. R. Moon
University of Virginia, Charlottesville, USA

T. R. Moon
e-mail: trm2k@virginia.edu

F. L. Huang (✉)
Curry School of Education, University of Virginia, P.O. Box 4829, Charlottesville, VA 22905-4829, USA
e-mail: fhh2b@virginia.edu

continue to be the school system's principal resource (Wayne and Youngs 2003). The logic is that more qualified teachers should be able to help students learn more effectively, as evidenced by student gain scores from year to year. Identifying the determinants of teacher quality in promoting student achievement, however, has been less clear and continues to be an important area of focus in educational research. This study examined a variety of teacher characteristics (e.g., teacher education, certification, and years of teacher experience) and their association with second grade student reading achievement scores in Title I,[1] Reading First (RF) schools in a Mid-Atlantic state. Reading First was enacted as part of NCLB to provide extra assistance to struggling readers in low performing, high poverty schools. Our research contributed to the study of teacher effects by including a more specific measure of experience that has not been explored in-depth: the number of years teaching *at a particular grade level* in addition to overall years of teaching experience. Three-level hierarchical linear modeling (HLM) was used in the analyses, controlling for school, classroom, and student background characteristics, including prior achievement.

From a policy perspective, focusing on the differential returns to student achievement gains based on the type of teaching experience is an important and sensitive issue since teacher compensation has traditionally been based on tenure, rather than performance (Muñoz and Chang 2008). For researchers and policy makers that are investigating the effects of teacher experience on student achievement gains, we caution against strongly interpreting study results that have used a general measure of teacher experience and have reported nonsignificant findings. Studies using a nonspecific measure may underestimate the effect of teacher experience.

1 Review of the literature

Research on the relationship of teacher effects on student achievement goes as far back to the release of the *Equality of Educational Opportunity* report or the "Coleman Report" (Coleman et al. 1966). Even after several decades, debates continue to ensue regarding the highly simplified notion that schools do not matter, interpreted from the statement that "schools bring little influence to bear upon a child's achievement that is independent of his background and general social context" (p. 325). The Coleman Report began a wave of empirical educational production function (EPF) studies that focused on inputs (e.g., school resources, teachers, textbooks) that were associated with increased student achievement. Over the years, several EPF studies have been conducted precisely to analyze various measurable teacher characteristics and their association with increased student achievement (e.g., Hanushek 1992; Kane et al. 2006; Nye et al. 2004; Sanders and Rivers 1996). Central to the study of teacher effects is the notion that the amount students learn in a year is partially a result of their teachers. Teachers have been found to differ enormously in producing achievement gains, with easily-measured teacher characteristics however, producing inconsistent results (Hanushek 1992).

[1] Title I is a federal program that is designated to specific schools with high percentages of poor students and provides schools additional funding opportunities to assist students in meeting academic benchmarks.

1.1 Teachers matter

Several studies (Hanushek 1992; Kane et al. 2006; Nye et al. 2004; Rowan et al. 2002; Sanders and Rivers 1996) have illustrated the positive and cumulative effects of teachers as they relate to student achievement. Hanushek (1992) estimated that the difference between a student having a good teacher and having a bad teacher can be more than one grade-level equivalent in test performance. As a result, students who wind up with good or bad teachers for several years in a row can have very different learning growth trajectories, even just after a few years of schooling. Sanders and Horn (1998) stated that teacher effectiveness was "the most important factor in the academic growth of a student" (p. 250). In another study, Rivkin et al. (2005) analyzed reading and mathematics achievement gains using a large dataset of more than 200,000 students per grade in 3,000 public elementary and middle schools from grades three through seven. Rivkin et al. concluded that teachers mattered and achievement gains were systematically related to teacher and school characteristics. While researchers generally have agreed that teachers matter, education research has failed to reach a consensus as to what readily identifiable teacher characteristics[2] enable teachers to bring about increased student achievement gains (Goldhaber and Anthony 2007; Palardy and Rumberger 2008). In our review of the literature on teacher characteristics and student achievement, we focused on production function studies that have taken into account student background characteristics or some form of prior ability measure (e.g., previous years test scores or IQ).

1.2 Teacher quality

The question of what teacher characteristics influence teacher effectiveness is an important one, given that NCLB calls for the hiring of only "highly qualified"[3] teachers (U.S. Department of Education 2004) on the basis of teacher background characteristics. Screening teachers on the basis of paper-based background qualifications such as educational attainment or certification is appealing in its simplicity and is far less demanding than observing teachers in the classroom or measuring teacher attitudes, practices, and beliefs. The relevance of investigating effective teacher characteristics is important since teachers who demonstrate "better" qualifications are rewarded monetarily (Johnson and Cornman 2008), creating additional incentives for teachers to pursue higher degrees or certification and, as a result, raising costs for school districts.

While teacher quality is a global concern, countries have defined *qualified teachers* differently and only a few studies outside the US have focused on teacher characteristics associated with higher student achievement gains (Akiba et al. 2007). Wayne and Youngs (2003) conducted a review of the literature on teacher characteristics and student achievement gains and only found one study outside of the US that had teachers as the primary research focus with the necessary controls (i.e.,

[2] We refer to readily measured characteristics that form a teacher's credentials and not teacher behaviors, attitudes, and practices.

[3] Highly qualified teachers are defined as those teachers having at least a bachelor's degree, full state certification or licensure, and proof that they know the subject they teach (U.S. Department of Education, 2004).

socioeconomic status and prior ability). However, the presence of large scale datasets such as the Trends in International Mathematics and Science Study (TIMSS), a project of the International Association for the Evaluation of Educational Achievement (IEA), should facilitate more research in this area in the future.

1.3 Teacher certification

One characteristic of a highly qualified teacher under NCLB is having full state certification or licensure. Teacher certification is a topic that has been heavily debated (for an example see Darling-Hammond et al. 2001; Goldhaber and Brewer 2000, 2001). However, despite the existing incentives, empirical research associating certification and student achievement gains has been inconsistent (e.g., Croninger et al. 2007; Goldhaber and Brewer 1996; Kane et al. 2006; Palardy and Rumberger 2008). Palardy and Rumberger (2008) analyzed the nationally representative Early Childhood Longitudinal Study (ECLS) dataset and involved 3,496 first grade students, in 887 classrooms, and 253 schools. Their results suggested that certification mattered for reading achievement in the first grade but not for mathematics. In another study that used the same dataset, Croninger et al. (2007) did not find statistically significant results for certification for both reading and mathematics. In addition, certification in a particular content area may matter for some subjects more than others and at different grade levels (Goldhaber and Brewer 1996). Thirty years ago, Madaus et al. (1979), based on their study on secondary schools in Ireland, hinted at the differential effects of school that were dependent on the subject matter measured such as mathematics that was learned primarily in school, as compared to a language-based skill, such as reading, that can be learned at home.

If stringent teacher certification requirements raise student academic achievement and deter few potential teachers, then requiring certification may be good policy (Boyd et al. 2007). However, if certification acts as a barrier to entry into the profession, reduces the supply of teachers, and does not produce student achievement gains, then such a policy may be worth revisiting. In reviewing the effect of certification on teacher quality, Boyd et al. concluded that the research evidence was "just too thin to have serious implications for policy" (p. 63). To further complicate matters, certification requirements vary from state to state and teachers who do not meet the requirements in one state may meet the requirement in another state as no single certification test has been used consistently (Birman et al. 2007). The inconsistent results of previous studies warrant the collection and analysis of additional data exploring the impact of teacher certification on reading achievement.

1.4 Teacher education

Under NCLB's definition, a highly qualified teacher should have at a minimum, a bachelor's degree. However, in order to be fully licensed in some states, a master's degree may be required and in some school systems, teachers are required to obtain an advanced degree after a certain number of years of teaching (Goldhaber and Brewer 1998). The emphasis on advanced degrees, like certification requirements, may influence the supply of teachers in the workforce. Almost without exception,

teachers are monetarily rewarded for possessing advanced degrees. For example in 2006, the median base salary for a teacher in Colorado with a bachelor's degree was US$36.7 thousand compared to the median base salary for a teacher with a master's degree which was US$51 thousand, a differential of almost 40% (Johnson and Cornman 2008). Again, the premium given to teachers with advanced degrees, much like licensing requirements, comes into question. In an earlier study of educational productivity factors, Hanushek (1986) reviewed 106 studies that used teacher education level as a variable and found no strong evidence supporting the idea that teachers with advanced degrees performed any differently from those teachers that had only a bachelor's degree.

However, an important assumption is that degree giving bodies are of comparable quality and a possible explanation is that studies may not have differentiated between colleges of differing quality. In an early EPF study, Summers and Wolfe (1977) found a statistically significant association with higher achievement gains if a student were taught by a teacher who received a bachelor's degree from a higher rated college. Another explanation was that the subject focus of a degree (e.g., a degree in mathematics) may matter as well (Goldhaber and Brewer 1996, 1998; Rowan et al. 1997). In a reanalysis of the Hanushek (1986) studies that appeared in peer-reviewed journals but using more statistically advanced meta-analytic methods, Greenwald et al. (1996) found positive results for teacher characteristics, such as education and level of experience, as they related to achievement. They stated that teacher education and teacher experience "show very strong relations with student achievement" (p. 384). Still however, several other recent studies have shown no statistically significant differences with students taught by teachers with advanced degrees (Croninger et al. 2007; Nye et al. 2004; Rivkin et al. 2005). While from a policy maker's point of view, creating a universal incentive for more educated teachers may sound intuitive and appealing, incentives geared at promoting more teacher education may effectively be raising costs without an accompanied increase in student learning gains.

1.5 Teacher experience

Teaching experience, together with certification status and educational attainment, has often been used as a variable in EPF studies. Teacher experience has monetary consequences for policy makers and administrators as teacher tenure has long been the basis of teacher compensation rather than student achievement (Muñoz and Chang 2008). The logic is that the more years of experience a teacher has, the more effective the teacher should be in teaching.

On a cross-national level, average teacher experience can vary widely. Wößmann (2005) examined the TIMSS 1995 dataset and focused on 15 European countries and the US. For mathematics teachers who taught 15 year olds, the average level of teacher experience ranged from a low of 7.94 years for Portugal to a high of approximately 20 years for Belgium, Denmark, France, and Spain, with the US math teachers having an average of 15 years of experience. However, of the countries examined, Wößmann only found statistical significance (at the .05 level) for teacher experience in four countries: Belgium, Denmark, the Netherlands, and Norway. Oddly enough, the coefficients for teacher experience were negative (i.e., the more years of experience,

the *less* effective the teacher became) for England, Germany, and Ireland, though they were not statistically significant. Negative coefficients however have been found in other studies and are not entirely uncommon. In Greenwald et al.'s (1996) meta-analysis, teaching experience was found to be a strong positive predictor of student achievement, with 30% of the studies showing statistically significant positive effects with only three percent showing small, statistically significant negative effects.

A certain amount of teaching effectiveness may also be related to the compositional effect of the teaching cohort (Boyd et al. 2008). Murnane and Phillips (1981) referred to this as selection effects which were defined as "the differences between the average abilities of teachers of a given experience level who choose to remain classroom teachers and those who choose to leave classroom teaching" (p. 456). For example, if ineffective teachers leave the profession early, the remaining teachers on average become "more effective." An experimental study where students were randomly assigned to teachers should alleviate selection effects. In a study of a well-known randomized experiment in Tennessee, Nye et al. (2004) found statistically significant positive results for teacher experience, using three-level multilevel modeling, in second grade reading (but nonsignificant findings for first and third graders).

Rivkin et al. (2005) took into consideration compositional effects and found that beginning teachers did worse than experienced teachers in both reading and mathematics, but there was little evidence that teachers improved after the third year on the job. In another study, Kane et al.'s (2006) New York City public school research found returns to teacher experience to be statistically significant, while controlling for student, classroom, and school characteristics. The researchers also stated that "the issue of heterogeneous returns to different forms of experience has received little attention in the value-added literature, but has important policy implications" (p. 30). If a large part of a teacher's effectiveness comes from having learned subject specific content on the job at a particular grade level, school administrators should, with all things being equal, focus more on retaining their effective teachers as opposed to hiring teachers who have taught at other grade levels.

Other studies that controlled for students' prior achievement, noted no significant results for teacher experience (e.g., Croninger et al. 2007; Link and Ratledge 1979; Summers and Wolfe 1977). Summers and Wolfe's (1977) research found that while on average, teacher experience was unimportant, students above grade level benefited from teachers with more experience and students below grade level were *negatively* affected. While effect sizes for teacher experience on student achievement varied from study to study, research that accounted for the nonrandom matching of teachers to students, suggested that teachers become significantly more effective in the first three to 5 years of their career (Goldhaber 2008).[4]

1.6 Beyond "traditional" background characteristics

Researchers have also studied teacher characteristics that are not so readily measurable or easily captured on a resume. Several other factors such as motivation, enthusiasm, and presentation skills are likely to influence student achievement as well (Goldhaber and Brewer 1998). Other studies that investigated noncognitive

[4] Gains resulting from teacher experience were larger in the area of mathematics.

teacher characteristics such as teacher expectations (Rosenthal and Jacobson 1968), motivation (Rowan et al. 1997), attitudes and instructional practices (Palardy and Rumberger 2008), and other characteristics such as caring, and dedication (Stronge et al. 2008) have shown the importance of affective teacher characteristics. Teacher behavior is an important and significant part of "teacher effectiveness" research (Muijs and Reynolds 2003). Based on their study of 36 English and Welsh primary schools, Muijs and Reynolds found that teacher behavior could account for 5.6% of the variance in adjusted student mathematics achievement and over 50% of between classroom variance. The limited inclusion or lack of noncognitive teacher characteristics in datasets will restrict the amount of explainable variance in student achievement that is attributable at the teacher level.

2 Research questions

The goal of this study was to add to the growing body of literature that examined teacher effects on student reading achievement, with a specific focus on second grade students in low performing, high poverty, RF schools in a Mid-Atlantic state. In addition to including the commonly used proxy measures of teacher "quality" (i.e., certification, education, and experience), we used a teacher's years of experience at a particular grade level as a key independent variable. Overall teaching experience has been investigated in several studies (e.g., Croninger et al. 2007; Link and Ratledge 1979; Murnane and Phillips 1981; Summers and Wolfe 1977) but a more specific measure of experience has important policy implications (Kane et al. 2006). While conceptually, teacher experience should differ depending on the students and subject taught (e.g., teaching high school students is different from elementary school students, teaching math is different from teaching reading), studies and surveys have not effectively differentiated between the different types of teaching experience.

Understanding the characteristics and effects of teachers in low performing schools is imperative, even more so at the early stages of a child's academic life. Doing well in the first few years of schooling has been found to be important for long-term success, as the early grades in school make up a "critical period" in a child's overall development (Alexander and Entwisle 1988). Since teachers can make a difference in student achievement, raising the quality of teachers in low performing schools "may well be the single best opportunity to reduce racial and socioeconomic achievement gaps" (Boyd et al. 2008, p. 535). Raising teacher quality however requires policy makers to further understand the significant determinants of teacher quality, as it relates in this case, to both low income and low performing students. Specifically, we asked the following research questions:

1. What proportion of variance in second grade student reading achievement gains were attributable to student background, teacher characteristics, and school effects?
2. What readily measurable teacher characteristics (i.e., teacher certification, possession of an advanced degree, recent attendance of reading conferences, and years of teaching experience) had statistically significant associations with student reading achievement gains, controlling for prior student achievement and other student level characteristics?

3. In terms of years of teaching, were there significant differences between the potential returns of teachers' total years of teaching experience and the years of teaching at grade level?

3 Methods

3.1 Data and subjects

The study focused specifically on low performing, Title I, RF schools. Reading First is based on scientifically-based reading research (SBRR) on "what works" in reading instruction to reduce the number of children who experience reading difficulties in later years (Snow et al. 1998). For a school to be eligible for RF funding in this state, the school had to be a Title I school (with a poverty index of 40%) with a pass rate of less than 60% on the spring 2002 third grade English state test. As a condition of receiving the grant, schools selected a comprehensive reading program that followed the principles of SBRR. Schools also provided a protected block of 90 minutes for reading instruction. In addition, schools had a reading coach and provided a minimum of US$1,000 in professional development for each teacher that worked with K-3 students.

The data for this study came from several sources. Student background and achievement data came from the state RF dataset. Teacher characteristics came from a teacher survey administered to all teachers in RF schools in the state. School-level information was sourced from the Department of Education website.

Assessment data came from students that stayed in the same RF school from 2006 to 2007. Students were tested in the first grade (2006) and second grade (2007) using the Stanford Reading First assessment (Harcourt Assessment, Inc. 2004). A series of filters were used to derive the analytic sample. After the exclusion of students designated as receiving special services (i.e., migrant education, special education, otherwise handicapped, or gifted and talented) and those with special conditions (e.g., autistic, hearing impaired, developmentally delayed) the sample size was reduced from 3,440 students to 2,210 students.

Surveys were administered in May 2007 to all teachers, reading coaches, and school administrators in RF schools in the state. For second grade teachers, 249 surveys were sent out and 195 surveys were received for a response rate of 78.3%. After removing teacher respondents who had joined the school in the middle of the school year, were teaching in newly awarded RF schools,[5] or were from designated control schools, student and teacher data were merged. Merging of assessment data and student data resulted in a final sample size of 1,544 students, 154 teachers, and 53 schools. In the final sample, there was an average of ten students per teacher and three second grade teachers per school.

To test whether the reduced sample of 2,210 students was systematically different (i.e., biased) from the final analytic sample used, two independent *t*-tests were run using spring 2006 and spring 2007 reading assessment results using an α

[5] Over 12 schools only received the RF grant beginning 2006.

of .05. Homogeneity of variance assumptions in both years held; $F_{2006(664,1544)}$= 1.04, p=.53 and $F_{2007(664,1544)}$=1.02, p=.75 and t-test results indicated no statistically significant difference between the group analyzed and the sample excluded; $t_{2006(2210)}$=.27, p=.79 and $t_{2007(2210)}$=−.89, p=.37.

3.2 Variable descriptions

3.2.1 Achievement scores

Reading achievement scores consisted of multiple choice scaled score results from the Stanford Reading First assessment (Harcourt Assessment, Inc. 2004), an edition of the Stanford Achievement Test, Tenth Edition (SAT 10). The assessment focused on the five core components of SBRR (Otaiba et al. 2005): phonemic awareness, phonics, vocabulary development, reading fluency, and comprehension. Scaled scores were linked across different grade levels and represented approximately equal units on a continuous scale, making scaled scores "especially suitable for comparing group performance over time" (Harcourt Assessment, Inc. 2004, p. 33). Based on the Kuder-Richardson formula 20 (K-R 20), the Stanford Reading First multiple choice test had a high degree of internal consistency with a K-R 20 of .92 (Harcourt Assessment, Inc. 2004). Reliability measures of above .80 are considered acceptable (Reynolds et al. 2006). As mandated by the state, the assessment was administered in April of each school year and this study included both first and second grade scores for the same students. First grade reading achievement scores were used as a natural control and allowed the statistical models built to account for various other inputs that occurred prior to the second grade, with first grade scores acting as a proxy for past inputs. Scaled scores were standardized in the analyses (M=0, SD=1).

3.2.2 Teacher qualifications

The sample of teachers was 93% female and 85% white. The independent variables of interest were related to teacher licensing/certification, educational attainment, reading conference attendance, and years of teaching experience (both total years teaching and years teaching at the second grade). Due to the stringent requirements of public school teachers in the state, only a few teachers were provisionally licensed (7%), while the rest were fully licensed (see Table 1). In terms of the highest level of educational attainment, 73% had a bachelor's degree or were enrolled in a master's program, while the rest had at least a master's degree. Around half (50%) of the second grade teachers did not attend any professional reading conference in the past year. In terms of teaching experience, nearly a fifth (19%) of the second grade teachers had less than or equal to 2 years of experience. This group represented the beginning teachers. Two thirds were seasoned teachers (62%), having five or more years of teaching experience and 19% had between two and 5 years of teaching experience. However, when the same teachers were asked how long they had been teaching in the second grade, 40% had only 2 years of experience or less, while 29% had more than 5 years of teaching experience in the second grade. In other words, while the teachers had several years of experience overall, far fewer had actual experience teaching in the second grade.

Table 1 Descriptive statistics for teachers and students (1,544 students, 154 teachers, and 53 schools)

Variables	*M*	*SD*	*Min*	*Max*
School variables				
Percent eligible for free or reduced price lunch	63.958	12.493	40.940	91.850
Classroom (Teacher) variables				
Educational attainment (1 = MA plus)	.273	.447	0	1.000
Licensed (1 = yes)	.935	.247	0	1.000
Experience (1 ≤ 2 years), Total	.188	.392	0	1.000
Experience (1 >5 years), Total	.617	.488	0	1.000
Experience (1 ≤ 2 years), At grade level	.403	.492	0	1.000
Experience (1 >5 years), At grade level	.286	.453	0	1.000
Attended a reading conference (1 = yes)	.506	.502	0	1.000
Class size (1 = 20+ students)	.292	.456	0	1.000
Percent of struggling readers (1 = 15%+)	.429	.496	0	1.000
Percent eligible for free or reduced price lunch	62.104	18.004	14.000	100.000
Student variables				
Minority (1 = yes)	.544	.498	0	1.000
Female student (1 = yes)	.542	.498	0	1.000
Economically disadvantaged (1 = yes)	.576	.494	0	1.000
Age at time of testing	8.078	.406	6.787	10.614
Teacher other race (1 = yes)	.455	.498	0	1.000
First grade reading achievement	590.775	35.985	469	713
Second grade reading achievement	625.668	32.589	532	765

In the analyses, percent eligible for free and reduced price lunch, age, and reading achievement scores are standardized (M=0, SD=1)

3.2.3 Control variables

Several student-level control variables were used, in addition to prior achievement, in an attempt to isolate teacher effects. Gender (1=female, 0=male), economic status (1=economically disadvantaged, 0=not economically disadvantaged), race (1=minority, 0=non-minority), and age at the time of second grade testing as measured in years, were entered. In the student sample, 54% were minority students, 58% were economically disadvantaged, and 54% were female. Student-level free or reduced price lunch (FRPL) status was used as a proxy for SES as represented by the economically disadvantaged variable. The average student age at the time of second grade testing was 8 years old. Age was standardized in the analyses (M=0, SD=1). In addition, previous studies (Dee 1995; Hanushek 1992) have suggested that teachers of the same race as their students performed better in improving achievement. A variable was constructed to reflect a similarity or difference between a teacher and student's ethnicity. A little over half (55%) of the students were of the same race as their teacher.

Classroom teacher level covariates were also included in the analyses which were class size, percentage of students eligible for FRPL, and percentage of struggling readers in the classroom. A large amount of research (Greenwald et al. 1996; Rivkin et al. 2005;

Word et al. 1990) has discussed the importance of small class sizes and its association with student learning. In the sample analyzed, nearly a third (.292) had class sizes of greater than 20 students. Aside from the possible effect of class size, classroom composition has been found to have an effect on student performance as well. Summers and Wolfe (1977) discussed the importance of peer group effects and stated that being in a student body with more low achievers had a negative effect on learning for all students. More recently, Kainz and Vernon-Feagans (2007) discussed how first grade students in classrooms with higher percentages of below grade level readers exhibited lower reading performance. In the present sample, around 43% of the teachers had a higher than average percentage of struggling readers in the classroom (greater than 15%).

At the school-level, the percent of students eligible for FRPL status in the entire school was also used as a proxy for overall SES. In a three-level multilevel study on reading achievement, Croninger et al. (2007) found a statistically significant positive relationship between average SES at the school-level and student reading achievement. For the present study, on average, schools had an FRPL eligibility status of 64%, which was to be expected as the focus of the analyses was on RF schools, which meant by definition that schools were designated as Title I schools in 2003.

3.3 Analytic strategy

Value-added models were used in the analysis of teacher effects. Value-added models were designed to measure change in student achievement, adjusting for differences in students' prior achievement, home, and social background (Goldhaber and Brewer 1996; Rowan et al. 2002). Achievement has been defined as a function of a student's genetic makeup, SES, teacher quality, non-teacher school quality, and peer group characteristics (Summers and Wolfe 1977). In order to analyze the relationship of teacher quality on achievement, variance from the other variables were controlled for. Assessing the effect teachers have on student achievement has often been reported as "percent of variance" in student achievement accounted for by teachers (Rowan 2004).

In order to partition out the variance at different levels, several three-level multilevel models were used to analyze the results, isolating the attributable amount of variance to the different levels. The use of multilevel modeling disentangled the student, classroom (teacher), and school variance components. Multilevel modeling has become increasingly applied to education studies, explicitly taking into account the nested nature of the data, avoiding issues such as aggregation bias and the misestimation of standard errors (Raudenbush and Bryk 2002). In this study, students (level-1) were nested within classrooms/teachers (level-2) that were in turn nested within schools (level-3). The level-3 model assessed the possible contextual effects at the school-level but a smaller variance between schools was expected as all the schools included in the analyses were similar in the sense that they were Title I, RF schools.

All HLM analyses were done using the PROC MIXED procedure in SAS. We standardized the dependent (2007 reading achievement scaled scores) and continuous independent variables, allowing coefficients to be interpreted as effect sizes. A total of six multilevel models were constructed, allowing the amount of variance explained at each level to be computed with the addition of more variables.

In addition to estimating the statistical significance of the coefficients, a likelihood ratio test (LRT) was also conducted to test model fit when more parameters were added, using the −2 log likelihood (−2LL). The null hypothesis of the LRT was that the change between models was zero, given by comparing the differences in the −2LLs to a χ^2 distribution, with the degrees of freedom corresponding to the number of added parameters between models. A rejection of the null hypothesis indicated that the model provided a better fit than the preceding model.

The first stage of the analyses, in keeping with HLM convention, involved running an unconditional model (i.e., a model with no predictors) or a one-way random effects analysis of variance (ANOVA). The analysis was done to evaluate the importance of the grouping effect through the use of the intraclass correlation (ρ) or the proportion of variance that was between schools ($_k$) and classroom teachers ($_j$). Each child's second grade reading achievement score (Y_{ijk}) could be explained by a classroom mean score (π_{0jk}) and a unique error associated with the child (e_{ijk}). Classroom means were explained by the grand classroom mean (β_{00k}) and the unique error associated with each classroom (r_{0jk}). School means were explained by the grand mean (γ_{000}), and the unique effect for each school (u_{00k}).

The analysis generated answers to the following questions: What proportion of total variance was between schools and what proportion of total variance was between classrooms? How much did individual students vary about their classroom means? Questions were answered by assessing the variance estimates between schools ($\tau_{\beta 00}$), between classrooms ($\tau_{\pi 00}$), and within classrooms (σ^2). Succeeding models with level-2 and level-3 predictors were assessed in comparison to the baseline, unconditional model, assessing a reduction of variance brought about by the inclusion of the covariates.

The second model only included the students' first grade Stanford Reading First achievement score. Prior achievement tended to have the highest association with the dependent variable (e.g., Croninger et al. 2007; Link and Ratledge 1979; Muijs and Reynolds 2003; Nye et al. 2004; Palardy and Rumberger 2008).

The third model estimated a student-level model, which added all the level-1 independent variables. Level-1 covariates included race (i.e., minority status), economic status, gender, and age at time of testing. Controlling for prior achievement scores and student background characteristics was required to reduce the alternative explanations of student achievement differences (Rowan et al. 2002; Wayne and Youngs 2003).While the third model did not answer the research questions directly, it was important to control for student background characteristics, of which policy makers have little influence over. Similarly, the fourth model entered all the contextual variables at the school-level (percent eligible for FRPL: school) and at the classroom level (class size, percent struggling readers, and percent eligible for FRPL: classroom).

The fifth and sixth models estimated the teacher related independent variables: certification/licensing status, highest level of educational attainment, and reading conference attendance in the past year. The fifth model used the total years of teaching experience variables while the sixth model used the years of teaching experience at grade level variables. In addition, interactions between teacher years of experience and economic and minority status were tested to see if disadvantaged and

minority children benefitted more from teachers with greater years of experience. Interactions between classroom size and economic and minority status were tested as well.

4 Results

4.1 School and classroom grouping effects on achievement gains

The results of the unconditional model (see Table 2, Model 1) indicated how much variability in second grade reading was attributable to the grouping of students, within classrooms, within schools. Model results showed that at the end of the second grade, 78% of the variance in reading achievement was between students within classrooms, whereas 17% of the variance was between classrooms within schools, and only 5% of the variance was between schools. All variance components were statistically significant (p=.04 for the school-level and p<.001 for classroom and students). The combined grouped intraclass correlation (of school and classrooms or $u_{00k} + r_{0jk}$) of .22 indicated that around 22% of the variability of achievement scores resulted from school and classroom factors (.053+.168/.999).

Model two included prior achievement as a covariate that acted as a proxy for previous, unaccounted influences. First grade scores were, as expected, strongly associated with second grade scores. The inclusion of first grade scores reduced total variability in student achievement scores from .999 to .494, a large, significant reduction of 51%. The variance was computed by comparing the variance estimates

Table 2 Coefficients and variance components of unconditional model and gain score model (1,544 students, 154 teachers, and 53 schools)

Model	Unconditional model		Prior achievement	
	(1)		(2)	
	Coefficient estimates			
Intercept	.002*		−.003	
Student level				
First grade reading achievement			.687***	
	Variance components			
Variance (Percentage of total)				
School ($\tau_{\beta 00}/\ \tau_{\beta 00}+\tau_{\pi 00}+\sigma^2$)	.053	(5%)	.004	(1%)
Classroom ($\tau_{\pi 00}/\ \tau_{\beta 00}+\tau_{\pi 00}+\sigma^2$)	.168	(17%)	.090	(18%)
Student ($\sigma^2/\ \tau_{\beta 00}+\tau_{\pi 00}+\sigma^2$)	.778	(78%)	.400	(81%)
Total	.999		.494	
	Model fit statistics			
−2 Log likelihood	4191.1		3147.8[a]	

[a] Provided a better model fit compared to the preceding model

# p<.10;*p<.05;**p<.01;***p<.001

of the unconditional model with the current model, which described the proportion of variance accounted for by the predictors that were not included in the base model. Between school-level variability ceased to be significantly different from zero (p=.368), while classroom and student-level variability were still significant (p<.001). The LRT was used to assess whether the addition of prior achievement provided a better model fit based on the additional variable. Results of the LRT indicated a significantly better model fit ($\Delta\chi_{df=1}^2$=1,043.3, p<.001).

4.2 Effects of school, classroom, and student covariates

Models three and four (see Table 3), while they did not directly answer the research questions, were tested in order to isolate how much student background and the school/classroom (but not teacher) context contributed to student achievement gains. Student background characteristics, and to an extent, school and classroom contextual variables, are not within the control of the classroom teacher. Economic status (p<.01), minority achievement (p<.001), and age (p<.001) were statistically significant and negatively associated with achievement gains, while gender and having a teacher of a different race were not found to be significant (ps>.05). While prior achievement overall had the largest effect size (.65), minority status was the second largest factor (effect size=−.33). The addition of the student background variables, compared to the model with only prior achievement, proved to be a better model fit ($\Delta\chi_{df=5}^2$=83.9, p<.001).

Model four entered the school and classroom contextual variables and showed the significant effect of the concentration of students with FRPL eligibility status at the classroom level (effect size=−.08), which had a significant association with achievement, controlling for prior achievement and student background characteristics. However, the coefficients for FRPL eligibility status at the overall school-level, classroom size, and the concentration of struggling readers, were all not significantly different from zero (all ps>.05). The slight improvement was mirrored in the LRT which was not statistically significant compared to model three ($\Delta\chi_{df=4}^2$=7.2, p=.126).

4.3 Effects of teacher characteristics

Models five and six represented the main models of interest in the study which were the teacher only characteristics. Both models used the traditional measures of "highly qualified teachers" that included licensing status and highest level of educational attainment. Attendance to a professional development conference was also used to represent recent training in the past year. While the coefficients were in the expected direction (positive), none of them were statistically significant at even the .10 level. Models five and six directly answered research questions two and three as to which teacher characteristics were significantly associated with student achievement and if the type of teaching experience mattered.

The main difference between models five and six was that model five used the teachers' total teaching experience and model six used teaching experience at the second grade. For model five, teachers who had total teaching experience of less than 2 years (i.e., the novice teachers), did worse (effect size=−.06) than teachers with two to 5 years of total teaching experience, however the results were not

Table 3 Coefficients and variance components of observable teacher characteristics on second grade reading achievement (1,544 students, 154 teachers, and 53 schools)

Model	Student covariates (3)	Contextual covariates (4)	Total years teaching (5)	Teaching at grade level (6)
	Coefficient estimates			
Intercept	.189***	.184**	.071	.062
School Level				
Percent eligible for FRPL		.003	.002	.001
Classroom level				
Percent eligible for FRPL		−.084*	−.066*	−.073*
Percent struggling readers (1=15%+)		.003	−.003	.000
Class size (1=20+)		−.012	−.021	.000
Educational attainment (1=MA plus)			.073	.074
Licensed (1=yes)			.031	.030
Reading conference attendance (1=yes)			.012	.026
Experience, Total (1≤2 years)			−.063	
Experience, Total (1>5 years)			.127#	
Experience, At grade level (1≤2 years)				−.035
Experience, At grade level (1>5 years)				.273***
Student level				
First grade reading achievement	.652***	.650***	.649***	.650***
Minority (1 = yes)	.326***	−.323***	−.329***	−.324***
Economically disadvantaged (1 = yes)	−.073*	−.078*	−.081*	−.088*
Female student (1 = yes)	.043	.044	.043	.041
Age at testing	−.038*	−.037*	−.038*	−.037*
Teacher other race (1 = yes)	.017	.023	−.033	−.037
	Variance components			
Variance (Percentage of total)				
School ($\tau_{\beta 00}/\ \tau_{\beta 00}+\tau_{\pi 00}+\sigma^2$)	.001 (0%)	.005 (1%)	.003(1%)	.006 (1%)
Classroom ($\tau_{\pi 00}/\ \tau_{\beta 00}+\tau_{\pi 00}+\sigma^2$)	.077 (17%)	.068 (15%)	.063 (14%)	.051 (12%)
Student ($\sigma^2/\ \tau_{\beta 00}+\tau_{\pi 00}+\sigma^2$)	.382 (83%)	.382 (84%)	.382 (85%)	.382 (87%)
Total	.460	.455	.448	.439
	Model fit statistics			
−2 Log likelihood	3063.9[a]	3056.7[b]	3046.1[b]	3030.3[c]

Models (5) and (6) differ in the type of experience variable used (total experience compared to experience teaching at the second grade). Model fit was tested using a Likelihood Ratio Test

FRPL Free and Reduced Price Lunch

[a] Provided a better model fit compared to the preceding model

[b] Did not provide a better model fit compared to the preceding model

[c] Provided a better model fit compared to model 4

# $p<.10$;*$p<.05$;**$p<.01$;***$p<.001$

statistically significant. Teachers who had more than 5 years of total experience had a positive association with increased student achievement (effect size=.13) but this was only statistically significant at the .10 level. Technically, none of the variables of interest in model five were of statistical significance and the lack of model improvement was also mirrored in the LRT, comparing model five with model four ($\Delta\chi_{df=5}^2=10.6$, $p=.06$).

However, model six results indicated a statistically significant improvement in model fit when compared to model four ($\Delta\chi_{df=5}^2=26.4$, $p<.001$). While the teacher variables of licensing, education, and professional development attendance were all still not statistically significant, more than 5 years of teaching at the grade level was statistically significant. While novice teachers did poorer (effect size=−.04) than teachers with two to 5 years of experience, the result was also not statistically significant. However, when seasoned grade level teachers were considered (i.e., those with more than 5 years of teaching at the second grade), gains were both positive (effect size=.27) and statistically significant ($p<.001$). The bulk of variance in achievement gain scores still resided among students (87%), but the variance as a result of grouping (schools context plus teachers within classrooms) was reduced to 13% [(.006+.051)/(.006+.051+.382)]. This represented a variance reduction of 41% compared to the fully unconditional model with no predictors [(.22–.13)/.22]. The proportion of variance of achievement gains between classrooms was 12% [.051/(.006+.051+.38)] and represented a reduction of 30% [(.17–.12)/.17] compared with the unconditional model. In addition, all interactions tested were not statistically significant (all *ps*>.05).

5 Discussion

5.1 The importance of contextual and background characteristics

A large proportion of variation in student achievement in the unconditional model (model one) indicated that the bulk of variance was attributable to student characteristics (78%). This was in line with other studies that have partitioned the variance among students, classrooms, and teachers (Croninger et al. 2007; Palardy and Rumberger 2008; Rowan et al. 2002). Variance partitioned to schools was small in comparison—possibly because all schools were of a similar nature as all were Title I schools with a large amount of struggling readers as of 2002. Adding students' prior achievement almost effectively halved the variance—significantly improving the model fit and also eliminating any variance at the school-level. Using students' prior achievement showed the importance of controlling for previous achievement in assessing value-added models, since a large part of the variance can be attributed to factors that occurred in prior years. The effect size also of a student's prior achievement remains sizable throughout all the models (effect sizes=.65–.69) and was also similar to other models that have used previous year's scores (e.g., Croninger et al. 2007; Nye et al. 2004).

The entry of additional student background characteristics in model three indicated that older, minority, and economically disadvantaged students scored significantly lower than younger, Caucasian, noneconomically disadvantaged

children. The directions of the coefficients were in the expected direction. Nonsignificant findings included gender and having a teacher of another race. All the student background coefficients retained their direction and significance in all the other models. Of all the characteristics, the largest factor in all the models, second to prior achievement, was race, which accounted for around a third of a standard deviation difference. While the current study did not explore the widening or narrowing of the gaps by race, it does indicate that more effort needs to be placed in improving the scores of minority students, being an important element of NCLB. Economic status continued to remain to be an important factor in student achievement, though it had a smaller effect size compared with race. An additional model was constructed to test for interactions between race and economic status but the interaction was not significant. The use of both students' prior achievement and background characteristics illustrates why it is important to include these variables in assessing teacher effects, as these variables account for a large proportion of variance, which needs to be attributed correctly.

Model four showed the importance of the contextual effects. At the school-level, only FRPL status was used, though this was not found to be significant. Between school variance was already nonsignificant with model two, which included students' prior achievement. What was found to be significant at the classroom level was the proportion of FRPL students. The size of the effect is comparable to that found at the student-level. Again, the inclusion of contextual and peer effects indicates the importance of controls at the classroom level and was similar to findings of Croninger et al.'s (2007) study using the ECLS dataset. The number of struggling readers and class size, however, were not found to be significant. Research on poverty and SES effects goes as far back as the Coleman report (1966) and the effects of SES remain persistent even until today. While the use of SES in relation to academic achievement has been scrutinized in the past (White 1982), SES remains as one of the most commonly used contextual variables in education research (Sirin 2005). While the sample investigated were homogenous since they were all low income, RF schools, even within the schools, the concentration of low SES students made a difference in the classrooms.

5.2 No detectable differences with licensing status and educational attainment

The results of models five and six suggested that the traditional measures of teacher quality used by NCLB did not contribute significantly to student achievement, at least for second grade students in Title I, RF schools. Neither certification status nor level of educational attainment was significantly associated with increased student achievement. In other words, licensed teachers performed no better in raising student achievement compared with provisionally licensed teachers. The same can be said with regards to teachers' level of educational attainment: there was no significant difference between a teacher with a master's degree and a bachelor's degree as it related to increasing student achievement. However, in the case of NCLB, the minimum level of educational attainment was a bachelor's degree and comparing teachers with a bachelor's degree and without one was not performed. Most of the teachers in the sample could be considered highly qualified (93%) compared to the national average in 2004–2005 which stood at 75% (Birman et al. 2007). Our

findings echo much of the previous research on certification status, educational attainment, and student reading achievement (e.g., Croninger et al. 2007; Goldhaber and Brewer 1998; Hanushek 1986, 1992; Kane et al. 2006; Link and Ratledge 1979; Nye et al. 2004; Rivkin et al. 2005; Summers and Wolfe 1977)

5.3 Type of experience matters

While all other teacher background characteristics used in the models were not significant, only years of teaching at grade level was statistically significant. Total years of teaching experience was only significant though at the .10 level, much like what was reported by Croninger et al. (2007). When teacher experience is examined, it becomes apparent that type of experience, not just years of experience, is important. The effect of seasoned teachers (effect size=.27), who had more than 5 years of experience teaching at a particular grade level, is meaningful in comparison with the other effect sizes. The effect size for seasoned grade level teachers is three times larger than the effect size of economic status (effect size=−.09) and nearly as large as the effect of minority status (effect size=−.33). This suggests that experienced teachers teaching at a specific grade level could have a large effect in countering the effects of the widening achievement gaps.

While the effect size for seasoned teachers may be considered small by Cohen's standards (Cohen 1992), the effect is for one school year only. In the current sample, a one SD difference represents the learning growth between grades one and two or 35 scaled score points. Assuming that teacher effects are cumulative (Coleman et al. 1966; Hanushek 1992; Sanders and Rivers 1996) and effects are similar across grades, students taught by seasoned grade level teachers for 4 years in a row may wind up scoring more than one standard deviation (.27×4=1.08) higher or approximately one grade level more than students taught by beginning teachers.

A possible confounding factor in the study was the possibility of the nonrandom distribution of students to teachers. More experienced teachers may be assigned to higher achieving students (e.g., as a privilege of teacher seniority) or lower achieving students (e.g., to help compensate for weaker student ability). The phenomenon of teacher sorting may also be apparent, where minorities and economically disadvantaged children are more likely to be taught by the least skilled teachers (Lankford et al. 2002). To test the hypothesis that students were not systematically sorted among teachers, the sample was split in two: those taught by teachers with more than 5 years of grade level experience (i.e., the seasoned teachers group) and those taught by teachers with five or fewer years of experience at grade level (i.e., the beginning teachers group). A three-level multilevel model was run using prior year's achievement, the Stanford Reading First assessment in grade one, as the dependent variable. The independent variable of interest was grouping (i.e., taught by seasoned teachers=1 or beginning teachers=0) while controlling for gender, age, economic, and minority status. Results indicated that there was no statistically significant difference (p=.35) in baseline test scores between the students that were assigned to either type of teacher. In other words, the hypothesis that student assignments to teachers was based on prior ability was not supported and both groups had students with average scaled scores of 590 at the end of grade one, with fluctuations only arising out of chance or sampling error. This finding lends further

support to the analyses done, reducing the possible explanation of student gains resulting from selectivity bias (i.e., the nonrandom assignment of students to teachers).

Teaching experience has been found to exhibit nonlinear effects, with teachers becoming more effective in the first few years of teaching with learning gains decreasing or tapering off (Goldhaber 2008; Kane et al. 2006). While we used dummy variables for teacher experience, results should not be interpreted simply to mean that after 5 years of teaching at grade level, student achievement gains will materialize. To test for nonlinear effects for teaching experience at grade level, an additional three-level multilevel model was run using nonstandardized dependent and independent variables using a second-degree polynomial. A continuous years for teaching at grade level variable (min=1, max=40), as opposed to the original dummy variable specification, was used and a squared experience at grade level variable was introduced. In the model, both the coefficients (i.e., $ExpGL$=1.32 and $ExpGL^2$=−.031) for teacher experience at grade level were statistically significant (ps<.05). Teachers constantly improved teaching effectiveness until the 21st year and declined beyond that (see Fig. 1). The most effective teachers had 19–24 years of experience at grade level and were associated, holding all other variables constant, with increased student reading achievement of 14 scaled score points (effect size=.40). Teachers with at least more than 5 years of experience were associated with an increase of at least seven scaled score points (effect size=.20). Given that teacher effectiveness declined in succeeding years, experienced teachers were still more effective than rookie teachers, which explained why the dummy coded variable in model six was statistically significant. The results showed that teachers did get better over time but learning growth slowed over the years, with the largest growth coming from the first few years of teaching. To further test the model, a third-degree polynomial was introduced (i.e., a cubic function) but was not found to be statistically significant.

5.4 What did experienced teachers do differently?

While the focus of our study was not on finding what seasoned grade level teachers did differently from beginning grade level teachers, our teacher survey did collect information on how often teachers used grouping arrangements in their classroom (see Table 4). Teachers were asked to estimate, on average, what

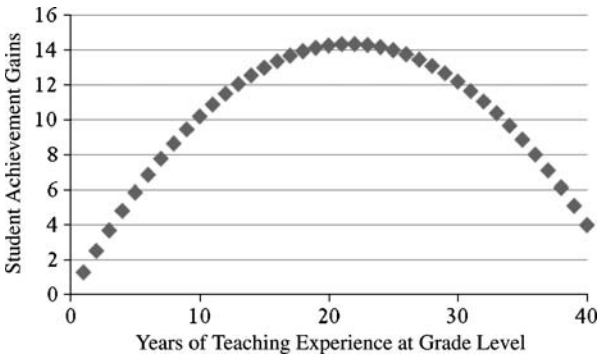

Fig. 1 Student achievement gains per additional year of teaching experience at grade level

Table 4 Average percentage of time a grouping arrangement was used in the classroom (142 teachers)

Grouping arrangement	Beginning grade level teachers (n=101)	Experienced grade level teachers (n=41)	Mean difference	t statistic	p value
Independent (students working by themselves)	17.46	14.32	−3.14	1.82#	0.07
One-on-one with teacher	7.86	7.55	−0.31	0.36	0.72
Small group (grouped by instructional need)	23.53	23.86	0.33	0.00	0.99
Small group (mixed ability)	9.70	7.80	−1.90	1.32	0.19
Pairs	8.09	7.52	−0.57	0.67	0.50
Whole-class	25.67	33.41	7.74	−2.90***	<.01

Homogeneity of variance assumptions held for all groups (p>.05)

# p<.10;*p<.05;**p<.01;***p<.001

percentage of time they used different types of grouping strategies. The grouping choices were: independent (students working by themselves), one-on-one with teacher, small group (grouped by instructional need), small group (mixed ability), pairs, and whole-class. Of the teachers that returned our survey, 142 or 92% supplied their grouping strategies ($n_{experienced}$=41; $n_{beginning}$=101). We used our dummy coded variable (1=seasoned grade level teachers and 0=beginning grade level teachers) to conduct independent sample t-tests to detect any differences in the proportion of time a grouping strategy was used. A prior study by Rowan et al. (2002) hypothesized that teachers who spent more time actively teaching in whole group lessons would have a positive effect on student achievement. By the same token, teachers who let their students use more individualized time (i.e., working independently) would indicate a *lack* of active teaching, resulting in negative effects on student achievement growth. For elementary school reading, Rowan et al.'s study produced significant results in the expected direction for both the use of whole group and individualized work. Analyses of our data indicated that, on average, beginning and experienced teachers allocated their time differently when it came to whole group instruction (t=−2.9, p<.01). Seasoned grade level teachers spent approximately a third of their time (33%) using whole-group instruction as compared to beginning grade level teachers who reported using it for only a little over a quarter of the time (26%). Muijs and Reynolds (2003) stated that a teacher actively teaching a whole class can be more effective than letting students work on their own for most of the lesson. The result of letting students work on their own independently was in the expected direction, where more experienced teachers used the strategy less, but the result was only significant at the .10 level (t=1.82, p=.07). Our results generally supported Rowan et al.'s findings. While other grouping strategies (e.g., ability grouping, one-on-one instruction, pairing) can be more effective depending on the variations in student ability levels and size of the class (Good and Brophy 1994), whole class instruction may be more effective in homogenous classes. An assumption made with whole group instruction was that the teacher was actively engaging their students and that lessons were well-structured, something that the required reading

program series was supposed to provide. Our findings however were exploratory in nature and warrant further study. Others studies (e.g., Good and Brophy 1994; Marzano 2007; Stronge et al. 2008) have identified different effective instructional behaviors and practices of teachers in much more detail.

5.5 Implications of teaching experience

The finding that seasoned grade level teachers can make a difference on student achievement may not seem startling. While Summer and Wolfe's (1977) possible explanation of why beginning teachers can be more effective in teaching below grade level students, because they possessed "undampened enthusiasm" compared to more experienced teachers may be plausible, other researchers (Goldhaber 2008; Hanushek 1992; Murnane and Phillips 1981) over three decades have suggested that teachers learn their trade on the job and become more effective through the years. In the sample, overall years of teaching experience and teaching experience at grade level were highly correlated (r=.66) but they had differential effects on student achievement. The average years of teaching experience in the sample of teachers used was 12.19 years (SD=10.83) while the years of teaching at grade level was less than half at 5.86 years (SD=7.22). While the two forms of experience are highly related, they are not the same thing.

A cause for concern is that Birman et al. (2007) found that high poverty and high minority schools had teachers who were considered under NCLB to be highly qualified, but had on average, three or fewer years of experience than their counterparts. In addition, highly qualified teachers in high poverty schools were less likely to have a degree in their field compared to teachers in low poverty schools. The phenomenon of nonrandom teacher sorting has been described in more detail by Lankford et al. (2002). If the more at-risk students (i.e., economically disadvantaged, minority children) are taught by less experienced teachers and teacher effects are cumulative, achievement gaps can be expected to worsen over time—exactly the opposite of what NCLB was supposed to address. Furthermore, a teacher's choice to move schools or leave the teaching profession showed a stronger association if the teacher taught in a high minority school with students from low income families (Shen 2001).

As suggested by Rivkin et al. (2005), policies should be investigated that promote the retention of more senior teachers and reduce the impact of inexperienced teachers. The issue of the retention of public school teachers has been a continuing concern (Shen 2001). Teaching, compared to other professions, has a relatively high turnover and almost 39% of new teachers in the US leave within the first 5 years on the job due to job dissatisfaction or the desire to explore other careers (Ingersoll 2002). While policy makers have suggested that merit pay may work in attracting and retaining good teachers, teachers themselves have not embraced the concept with the same vigor as policy makers (Matthes et al. 1990).

In our study's sample of teachers, 40% had taught at grade level for 2 years or less and 70% had taught at grade level for 5 years or less. If experienced teachers at grade level, not just overall years of teaching experience alone, are essential to increasing student achievement, the high turnover represents an even more urgent and pressing problem to be addressed. While school districts may prefer hiring

teachers with experience (with all other things being equal), more attention can be paid to matching teaching experience with grade level experience and promoting incentives that reward that matching. For example, in a hypothetical school, a certified teacher with a master's degree, that has been teaching for 5 years in middle school but has moved to teach grade two students will be paid more than a noncertified teacher, with a bachelor's degree, that has already been teaching in grade two for 5 years. However in the example, the noncertified teacher may be more effective in raising student achievement scores compared to the certified teacher. In addition, the current incentive system encourages the noncertified teacher to become licensed and pursue graduate studies—not necessarily negative factors, but effectively increases costs for the school which pays more for better credentialed teachers, while not necessarily increasing student learning gains.

5.6 Caution in interpreting teacher experience studies

An important finding of our research is that studies that have used only a general variable for teacher experience, as opposed to a specific measure such as teaching experience at grade level, may be underestimating the actual effect of experience when reporting nonsignificant findings. Total years of teaching experience sets the upper boundary of experience since teaching experience at grade level is less than or at most equal to total years of teaching experience (i.e., teaching experience at a particular grade level is a subset of the total years of teaching experience). It is understandable that a general measure of teaching experience has been used for convenience but study results using an undifferentiated teaching experience variable can be misinterpreted. In our study, if we had just used "years of teacher experience" without a more refined measure, a simple interpretation would have been "experience does not matter" since the findings were not statistically significant.

Standard, undifferentiated years of teaching experience variables are not uncommon. For example, in TIMSS, one of the largest educational data collection efforts worldwide, the teacher questionnaire asks, "3. By the end of the school year, how many years will you have been teaching altogether? [Number of years taught]" (International Association for the Evaluation of Educational Achievement, 2007). A general form of teaching experience variable does not necessarily distinguish between a teacher who has been teaching fourth grade mathematics for 10 years compared to a person who has been teaching eighth grade science for 9 years and fourth grade mathematics for 1 year: both have 10 years of experience but one may be a more effective mathematics teacher. If that were the case, it may not be surprising to find nonsignificant results for studies using a "years of teaching experience" variable. Wößmann's (2005) study used the TIMSS dataset and found nonsignificant findings for teacher experience for the majority of the countries he analyzed, including the US and 12 European countries. While the statistical methods employed may be correct, the type of experience variable used limits the interpretation of results. Researchers and policy makers are cautioned on interpreting nonsignificant empirical findings if only a "years of teaching experience" variable is used.

6 Conclusions

This study focused on the differential returns of two types of teaching experience variables, overall teaching experience and grade level teaching experience, in relation to second grade student reading achievement gains in Title I, RF schools in a Mid-Atlantic state. In addition, we investigated the commonly used measures for teacher quality such as certification status and level of educational attainment. Value-added models were used to analyze data from 1,544 students, 154 teachers, and 53 schools using three-level HLM, controlling for prior ability along with contextual and student background characteristics. Results indicated that a highly qualified teacher was not necessarily a highly effective teacher. Certified and uncertified teachers as well as teachers with master's degrees or bachelor's degrees performed no differently from each other in terms of raising student achievement. While school systems may reward certified teachers with advanced degrees more than uncertified teachers with bachelor's degrees, the result could lead to higher teacher costs without an accompanied increase in student achievement. In addition, requirements may also act as barrier to entry into the profession, limiting the supply of teachers. Reading conference attendance by teachers in the past year also did not have much of an association with student outcomes.

The differential effect of teaching experience was tested as one model used total years of teaching experience as an independent variable, while another model used a more specific variable, total years of teaching experience at grade level. Years of teaching experience at a particular grade level was statistically significant, but not total years of teaching. The model that tested the association of total years of teacher experience indicated no significant difference in terms of student achievement between a novice and a seasoned teacher (at the .05 level). More importantly however, a teacher who had been teaching at a particular grade level for more than 5 years was positively and significantly associated with increased student achievement (effect size=.27). While not a large effect size by Cohen's (1992) standards, in comparison to other statistically significant variables in the model, grade level experience was sizable compared to race (minority status effect size=−.33) and SES (economically disadvantaged effect size=−.08). As research has shown that teacher effects can be cumulative (Coleman et al. 1966; Hanushek 1992; Sanders and Rivers 1996), effects of teacher grade level experience can be compounded over the years, leading to larger effect sizes.

Teacher oriented policies and interventions that focus on both retaining effective teachers and minimizing the impact of beginning teachers (e.g., effective mentoring, coaching, team teaching) should be further explored. As teaching experience at a particular grade level was found to show a statistically significant association with academic achievement gains, rewarding years of teaching at grade level, not just years of teaching alone, becomes more important to consider in efforts to improve student reading achievement.

Finally, policy makers and researchers are cautioned when interpreting studies that use nonspecific teacher experience measures which may possibly underestimate the explanatory power of experience. Pronouncing that "experience does not matter" based on nonsignificant empirical results sends a wrong signal to educators and policymakers, when in fact, the opposite may be true, as shown in our study.

7 Limitations

As with all studies, there are several limitations of this paper. The first limitation was that the sample was limited to low performing, high poverty schools in one state. While RF schools have high poverty levels (greater than 60% of the student population), it is not clear how RF schools are different from eligible, non-RF schools, aside from the additional resources that the schools receive as provided by the grant. Research (Huang and Moon 2008) has noted no overall reading achievement differences in RF schools and eligible, non-RF schools based on the state reading assessment used in the third grade from 2003 through 2006. A second limitation is that the current study focused on a single grade level for one subject area (i.e., second grade reading). Teacher qualifications may have a stronger impact at different grade levels (e.g., high school) and in different subject areas (e.g., mathematics). Third, while this study may shed light on the strength of association of certain teacher characteristics on student achievement scores, it is unclear how these qualifications work in the classroom. Policy makers must keep in mind that easily measured teacher characteristics (i.e., those found on a resume or teacher credentials) only form a small part of what makes teachers effective and a large proportion of variance at the teacher-level was still unaccounted for. Much of what makes a teacher effective may be observed in the classroom and calls for more qualitative research or studies that deal with the affective aspects of teaching. Lastly, the evidence presented is primarily correlational and was not experimentally manipulated, despite the statistical methods used in the analyses. While the term *effect* may be used to describe the statistical relationship between variables, a true causal relationship was not established. Subsequent experimental studies can be done to verify and validate findings in this study.

References

Akiba, M., LeTendre, G., & Scribner, J. (2007). Teacher quality, opportunity gap, and national achievement in 46 countries. *Educational Researcher, 36*, 369–387.

Alexander, K., & Entwisle, D. (1988). Achievement in the first two years of school: Patterns and processes. *Monographs of the Society for Research in Child Development*, *53(2)*.

Birman, B., Le Floch, K., Klekotka, A., Ludwig, M., Taylor, J., Walters, K., et al. (2007). *State and local implementation of the NCLB Act: Volume II- Teacher quality under NCLB: Interim report*. RAND Corporation. Retrieved on August 1, 2008, from http://www.rand.org/pubs/reprints/RP1283/.

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The effects of certification and preparation on teacher quality. *The Future of Children, 17*, 45–68.

Boyd, D., Lankford, H., & Wyckoff, J. (2008). Increasing the effectiveness of teachers in low-performing schools. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 535–550). Routledge.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

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

Croninger, R., Rice, J., Rathbun, A., & Masako, N. (2007). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review, 26*, 312–324.

Darling-Hammond, L., Berry, B., & Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. *Educational Evaluation and Policy Analysis, 23*(1), 57–77.

Dee, T. (1995). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review, 95*, 158–165.

Goldhaber, D. (2008). Teachers matter, but effective teacher quality policies are elusive. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 146–165). Routledge.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics, 89*, 134–150.

Goldhaber, D., & Brewer, D. (1996). Evaluating the effect of teacher degree level on educational performance. In W. Fowler (Ed.), *Developments in school finance* (pp. 197–210). Washington, DC: NCES.

Goldhaber, D., & Brewer, D. (1998). When should we reward degrees for teachers? *Phi Delta Kappan*, 134–138.

Goldhaber, D., & Brewer, D. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129–145.

Goldhaber, D., & Brewer, D. (2001). Evaluating the evidence on teacher certification: A rejoinder. *Educational Evaluation and Policy Analysis, 23*(1), 79–86.

Good, T., & Brophy, J. (1994). *Looking in classrooms* (6th ed.). New York, NY: Collins.

Greenwald, R., Hedges, L., & Laine, R. (1996). The effect of school resources on student achievement. *Review of Educational Research*, *3*, 361–396.

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *The Journal of Economic Literature, 24*, 1141–1177.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *The Journal of Political Economy, 100*, 84–117.

Harcourt Assessment, Inc. (2004). *Stanford reading first: Technical data report*. San Antonio, TX: Harcourt Assessment, Inc.

Huang, F., & Moon, T. (2008). *Reading First's impact on reading outcomes on second grade students: One county's experience in a Mid-Atlantic state*. Poster session presented at the annual meeting of the American Educational Research Association, New York, NY. (March).

Ingersoll, R. (2002). Holes in the teacher supply bucket. *School Administrator, 59*(3), 42–43.

International Association for the Evaluation of Educational Achievement (2007). *Trends in International Mathematics and Science Study: Teacher Questionnaire.* Retrieved February 25, 2009, from http://timss.bc.edu/TIMSS2007/PDF/T07_TeacherQ_G4.pdf.

Johnson, F., & Cornman, S. (2008). *Findings from the Pilot Teacher Compensation Survey: School year 2005–06 (NCES 2008-440).* National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, D.C. Retrieved July 5, 2008, from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2008440.

Kainz, K., & Vernon-Feagans, L. (2007). The ecology of early reading development for children in poverty. *The Elementary School Journal, 107*, 407–427.

Kane, T. J., Rockoff, J. E., & Staiger, D. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City.* NBER Working Paper Series, Vol. w12155. Retrieved March 5, 2008 from http://ssrn.com/abstract=896463.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*, 37–62.

Link, C., & Ratledge, E. (1979). Student perceptions, I. Q. and achievement. *The Journal of Human Resources, 14*, 98–111.

Madaus, G., Kellaghan, T., Rakow, E., & King, D. (1979). The sensitivity of measures of school effectiveness. *Harvard Educational Review, 49*, 207–230.

Marzano, R. (2007). Using action research and local models of instruction to enhance teaching. *Journal of Personnel Evaluation in Education, 20*, 117–128.

Matthes, W. A., Tollerud, T. R., & Langeveldt, W. (1990). Teacher beliefs regarding career ladders and merit pay. *Education, 110*, 535–40.

Muijs, D., & Reynolds, D. (2003). Student background and teacher effects on achievement and attainment in mathematics: A longitudinal study. *Educational Research and Evaluation, 9*, 289–314.

Muñoz, M., & Chang, F. (2008). The elusive relationship between teacher characteristics and student academic achievement growth: A longitudinal multilevel model for change. *Journal of Personnel Evaluation in Education, 20*, 147–164.

Murnane, R., & Phillips, B. (1981). Learning by doing, vintage, and selection: Three pieces of the puzzle related to teaching experience and teaching performance. *Economics of Education Review, 1*, 453–465.

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

Otaiba, S., Kosanovich-Grek, M., Torgesen, J., Hassler, L., & Wahl, M. (2005). Reviewing core kindergarten and first-grade reading programs in light of No Child Left Behind: An exploratory study. *Reading and Writing Quarterly, 21*, 377–400.

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.

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

Reynolds, C., Livingston, R., & Willson, V. (2006). *Measurement and assessment in education*. Boston, MA: Allyn & Bacon.

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

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York, NY: Holt, Rinehart and Winston.

Rowan, B. (2004). *Teachers matter: Evidence from value-added assessments. Research Points*. Washington, DC: American Educational Research Association.

Rowan, B., Chiang, F. S., & Miller, R. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education, 70*(4), 256–284.

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

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*, 247–256.

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

Shen, J. (2001). Teacher retention and attrition from public schools: Evidence from SASS91. *The Journal of Educational Research, 91*, 81–88.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*, 417–453.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, D.C.: National Academy.

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

Summers, A., & Wolfe, B. (1977). Do schools make a difference? *The American Economic Review, 67*, 639–652.

U. S. Department of Education. (2004). *New No Child Left Behind flexibility: Highly qualified teachers*. Retrieved July 5, 2008, from http://www.ed.gov/nclb/methods/teachers/hqtflexibility.html. March.

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

White, K. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*, 461–481.

Wößmann, L. (2005). Europe's schools. *Education Policy*, 445–504.

Word, E., Johnson, J., Bain, H., Fulton, D., Zaharias, J., Lintz, M., et al. (1990). *Student/Teacher Achievement Ratio (STAR): Tennessee's K-3 class-size study*. Nashville, TN: Tennessee State Department of Education.