Educational Evaluation and Policy Analysis

http://eepa.aera.net

The online version of this article can be found at:
http://epa.sagepub.com/cgi/content/abstract/30/2/111

Published on behalf of

American
Educational
Research
Association

http://www.aera.net

By

$SAGE

http://www.sagepublications.com

Additional services and information for *Educational Evaluation and Policy Analysis* can be found at:

Email Alerts: http://eepa.aera.net/cgi/alerts

Subscriptions: http://eepa.aera.net/subscriptions

Reprints: http://www.aera.net/reprints

Permissions: http://www.aera.net/permissions

Teacher Effectiveness in First Grade:
The Importance of Background Qualifications, Attitudes, and Instructional Practices for Student Learning

Gregory J. Palardy
University of California, Riverside

Russell W. Rumberger
University of California, Santa Barbara

This study uses Early Childhood Longitudinal Study data to investigate the importance of three general aspects of teacher effects—teacher background qualifications, attitudes, and instructional practices—to reading and math achievement gains in first grade. The results indicate that compared with instructional practices, background qualifications have less robust associations with achievement gains. These findings suggest that the No Child Left Behind Act's "highly qualified teacher" provision, which screens teachers on the basis of their background qualifications, is insufficient for ensuring that classrooms are led by teachers who are effective in raising student achievement. To meet that objective, educational policy needs to be directed toward improving aspects of teaching, such as instructional practices and teacher attitudes.

THE No Child Left Behind (NCLB) Act of 2001 mandated that schools hire only "highly qualified" teachers beginning in the 2005 school year (U.S. Department of Education, 2002). This influential federal legislation defines highly qualified teachers in terms of the background characteristics they bring into the classroom, including state certification (not including emergency, provisional, or temporary licenses), a minimum of a bachelor's degree, and—for secondary teachers—demonstrated subject-area competence. The general purpose of this provision is to increase the odds that classrooms are led by teachers who are effective in promoting student learning. Unfortunately, the research literature paints an ambiguous picture of which aspects of teachers are associated with student learning. Indeed, the stipulations established by NCLB may not be the best indicators of teacher effectiveness; rather, aspects such as teacher attitudes and practices may be superior.[1]

An array of studies has concluded that the amount students learn can be traced to aspects of teachers and teaching (Brophy & Good,

1986; Guarino, Hamilton, Lockwood, & Rathbun, 2006; Nye, Konstantopoulos, & Hedges, 2004; Rowan, Correnti, & Miller, 2002; Sanders & Horn, 1995; Wayne & Youngs, 2003; Xue & Meisels, 2004). That is, teachers differ substantially in their impact on student learning. An early review of the literature found that "teachers and schools differ dramatically in their effectiveness" (Hanushek, 1986, p. 1159). More recently, McCaffrey, Lockwood, Koretz, and Hamilton (2003) concluded that "teachers differentially affect student achievement," although the authors also point out that "the literature provides little convincing evidence on the magnitude of the typical teacher effect or relative importance of teachers as a source of variability in student achievement" (p. 113).

Although there is general agreement that teachers make a difference, there is a lack of consensus about which aspects of teachers matter most. Some researchers have focused on the background characteristics of teachers, such as their educational attainment, achievement and intelligence test scores, experience, and credentials (Darling-Hammond, Berry, & Thoreson, 2001; Dunkin & Biddle, 1974; Ehrenberg & Brewer, 1994, 1995; Rowan et al., 2002; Wayne & Youngs, 2003; Wenglinsky, 2002). These characteristics are attractive from a policy perspective, because many of them can be used to screen preservice teaching candidates. NCLB, for example, focuses solely on teacher background characteristics. Despite the widespread interest in such characteristics, however, there is relatively little scientific evidence that these characteristics have a measurable and consistent direct impact on student achievement (Guarino et al., 2006; Wayne & Youngs, 2003).

Other scholars have argued that "teaching, not teachers, is the critical factor" (Stigler & Hiebert, 1999, p. 10). That is, the practices that teachers employ in the classroom are more important than their education, credentials, experience, test scores, and other background variables. Two aspects of teaching have been examined in the research literature: teacher attitudes about their ability to teach and about students' ability to learn—sometimes referred to as teacher efficacy (Tschannen-Moran, Hoy, & Hoy, 1998)—and teaching practices or processes (Brophy & Good,

1986; Rowan et al., 2002). Compared to teacher background characteristics, teaching attitudes and practices have received less attention in the research literature, in part because they tend to be more difficult to measure or quantify. But studies that have examined direct measures of teaching practices have found substantial effects on student learning (e.g., Shacter & Thum, 2004). Moreover, these features of teachers may be more alterable once a teacher is in the schools. Teaching practices may be improved through professional development, mentoring, or through better professional training programs, for example, whereas background characteristics are more difficult to change. Moreover, in some academic subjects (e.g., math, science, and special education) where shortages of teachers are common, it may be counterproductive to reject applicants on the basis of background characteristics if they are effective or can learn to be effective by adopting certain attitudes and practices.

Surprisingly little is known about two fundamental aspects of teacher effects: the degree to which teachers matter and the features of teachers and teaching that are most important to student learning. The present inquiry uses data from the Early Childhood Longitudinal Study (ECLS) to investigate three progressively specific research questions. We begin by examining the relative importance of the classroom a child is assigned to versus the importance of individual and school factors to student achievement gains. We then examine the impact of teacher effectiveness on student achievement gains apart from other classrooms effects. Finally, we estimate the relative importance of teacher background qualifications as compared to teacher attitudes, beliefs, and instructional practices, the latter of which we conceptualize as more direct measures of *teaching*. Specifically, we address the following research questions:

1. What proportion of the variation in student achievement gains can be attributed to each of the three general sources: individual differences in student background, classroom effects (including teacher effects), and school effects?
2. To what degree do differences in teacher effectiveness affect student achievement gains?

112

3. What is the relative size of the effect of teachers' background qualifications, attitudes, and instructional practices on student achievement gains in first grade?[2]

Literature and Theoretical Framework

The association between various aspects of teachers and student achievement has been examined in past research dating back several decades. The Coleman report (Coleman et al., 1966) examined the impact of a number of teacher background characteristics, including years of experience, education level, and performance on a vocabulary test, ultimately concluding that teacher background characteristics had a larger effect on student achievement than any other general class of school effects except student body composition. Since then, numerous studies have been conducted on the relationship between various aspects of teacher quality and student achievement. Three aspects that have perhaps received the most attention are teacher background characteristics, teacher attitudes, and teacher instructional practices. We briefly review each of these areas below.

Teacher Background Characteristics

The most widely studied aspect of teacher effectiveness is concerned with teacher background characteristics. Several specific background characteristics have been examined in the research literature, including degrees, coursework, credentials, experience, test scores, and the prestige ratings of teachers' undergraduate institutions. Although individual studies have found that certain aspects of teacher background are associated with student achievement or learning, comprehensive reviews of the research literature have produced inconsistent conclusions, and there does not appear to be a consensus opinion. For example, Hanushek (1986, 1989, 1997) concluded that only a small proportion of the studies examining the effect of teacher characteristics had found positive associations with learning. Greenwald, Hedges, and Laine (1996), on the other hand, found that school "resource variables that attempt to describe the quality of teachers (teacher ability, teacher education, and teacher experience) show very strong relations with

student achievement" (p. 384). A more recent review of 21 studies that controlled for students' prior achievement and socioeconomic status (SES) concluded that "students learn more from teachers with certain characteristics" (Wayne & Youngs, 2003, p. 107). Although the authors found evidence that teachers' college ratings and test scores had consistently positive associations with achievement gains across grade levels and participants, there was less support in the literature for the effects of degrees, coursework, and certification—except in the case of high school mathematics. And although the research literature provides little consistent evidence that teacher background qualifications directly affect student learning, the results from a recent large-scale study of kindergarteners suggest that qualifications may affect learning indirectly (Guarino et al., 2006). For example, these authors found that coursework in reading instruction methods was positively associated with use of various effective reading instruction practices. In summary, although the literature suggests that some teacher characteristics are associated with effectiveness, the evidence is inconsistent and the effects may be indirect.

Teacher Attitudes

A number of teacher attitudes and beliefs have been investigated in the research literature: teacher perceptions of school climate and community (e.g., Raudenbush, Rowan, & Cheong, 1992; Rowan, Raudenbush, & Kang, 1991), teacher satisfaction (e.g., Lee, Dedrick, & Smith, 1991), and teacher efficacy (e.g., Hoy & Spero, 2005; Lee et al., 1991; Newmann, Rutter, & Smith, 1989; Raudenbush et al., 1992; Tschannen-Moran et al., 1998). Although much of this literature focuses on how schools influence teacher attitudes or how teacher attitudes change over time, several studies have investigated the association between teacher attitudes and student achievement at the school level of aggregation (i.e., the effect of mean teacher attitudes on mean student achievement), which can result in statistical problems such as aggregation bias and misestimation of standard errors (Raudenbush & Bryk, 2002). Most of these studies have focused on the impact of teachers' self-efficacy, defined in various ways to capture both

113

teachers' perceived ability to teach (sometimes referred to as *personal teaching efficacy*) and teachers' perception of students' ability to learn (sometimes referred to as *general teaching efficacy*).[3] Lee and her colleagues developed a single measure of "teacher collective responsibility for learning," which they found to be predictive of student learning in high school (Lee, Smith, & Croninger, 1997). Goddard, Hoy, and Hoy (2000) also created a single measure of collective teacher efficacy, which was associated with mean student achievement among elementary schools within a single midwestern school district. In contrast, Rumberger and Palardy (2005a) found that the two aspects of teacher efficacy had independent effects on various measures of high school performance. However, as a class of teacher effects, teacher attitudes and beliefs about their ability to teach and students' ability to learn have been underexamined in the literature.

Teacher Instructional Practices

Several studies, both small and large scale, have examined the impact of instructional practices on student achievement. Many of the large-scale studies employed national databases, such as the ECLS of the kindergarten class of 1998–1999 or the National Education Longitudinal Study of 1988. These investigations found significant effects for a number of measures of instructional practices on student learning across one or more grades during both early elementary school (e.g., Guarino et al., 2006; Lee, Burkam, Ready, Honigman, & Meisels, 2006; Xue & Meisels, 2004) and high school (e.g., Carbonaro & Gamoran, 2002; Lee et al., 1997). However, the measures of instructional practices generated by teacher surveys tend to be limited and do not always correlate with more direct measures of instructional practices, such as classroom observations (Burstein et al., 1995; Rowan et al., 2002).

Studies based on direct observations of teacher performance may show much larger effects for instructional practices than studies based on survey data. For example, Schacter and Thum (2004) examined the association between "teaching quality," a construct consisting of several performance-based measures collected from direct classroom observations at the primary school level, and student achievement gains.

They found a 0.91 standard deviation effect size for teaching quality, which is several times larger than the teacher effect size reported in a wide range of prior studies (for a recent review, see Nye et al., 2004, which we summarize below). Although more research is needed to validate the implications of this finding, it does suggest that direct observations show promise for assessing the full extent of the effects of instructional practice and other teacher effects. This may be because direct observations are more sensitive measurements of the actual instructional practices, which may reduce the error in the measurement of those factors and provide a better estimate of the true strength of association between instructional practices and student outcomes.

Of the three aspects of teachers examined in the present study, instructional practices are conceptualized as having the most proximal association with student learning. That is, instructional practices are theorized to influence student learning directly, whereas teacher background qualifications and teacher attitudes are theorized to influence learning indirectly through their association with instructional practices. For example, better trained teachers and teachers with higher levels of self-efficacy may engage in more effective instructional practices, and those more effective practices may directly affect student learning. Although instructional practices are believed to be more directly associated with learning, some research suggests the association is moderated by certain aspects of the classroom, such as class size (e.g., Betts & Shkolnik, 1999; Stasz & Stecher, 2000) and classroom composition (e.g., Burns & Mason, 2002; Connor, Morrison, & Katch, 2004; Stipek, 2004). Classroom composition, sometimes referred to as *peer effects*, can also have direct effects on classroom learning (e.g., Betts, Zau, & Rice, 2003; Hanushek, Kain, Markman, & Rivkin, 2003; Hoxby, 2000).

The Magnitude of Teacher Effects

Existing research has examined not only which aspects of teacher quality matter but also how much teachers matter. That is, studies have attempted to determine the proportion of the variation in student achievement and learning that can

114

be attributed to classroom or teacher effects as opposed to other sources (e.g., school effects and the effects of individual and family background characteristics). A recent review of this literature by Nye et al. (2004) examined 18 analyses from seven studies.[4] They reported that the proportion of the variance in student achievement gains due to teacher effects ranged from about 0.07 to 0.21. They also attempted to quantify these findings into effect sizes, concluding that the average magnitude of teacher effects was about 0.32 standard deviations. Although that would be considered a small effect size by general references (e.g., Cohen, 1988, chapter 9), it is substantial relative to the size of some other school-based effects on student achievement gains, such as class-size reduction (Nye et al., 2004).[5]

Nye and her colleagues (2004) also analyzed data from the Tennessee class-size reduction experiment, which randomly assigned students to classrooms. They found that approximately 65% to 73% of the variance in achievement gains was between students within classrooms.[6] For mathematics achievement gains during first grade, 65% of the variance was between students within classrooms, 21% was between classrooms within schools (teacher effects), and 14% was between schools.[7] For the reading achievement gains outcome, 73% of the variance was between students within classrooms, 11% was between classrooms within schools, and 16% was between schools. These findings are generally consistent with other multilevel studies of achievement gains (Rumberger & Palardy, 2005b; Scheerens & Bosker, 1997). In the second and third grades, the proportion of the variance between classrooms within schools was substantially greater than the proportion of the variance between schools, leading the authors to conclude that the choice of teachers has a greater impact on student learning than the choice of schools (Nye et al., 2004).[8]

Limitations

Although the literature on teacher effectiveness is fairly extensive, there are a few shortcomings. Two problems are the breadth of the theoretical frameworks guiding studies and the appropriateness of the statistical models employed. Many studies employ a theoretical framework narrowly focused on teacher background, neglecting to examine teaching practices and attitudes. Although teacher background characteristics are easier to measure than teacher attitudes and practices, a comprehensive theoretical framework of teacher effectiveness should include measures of all three aspects outlined above, because they are likely interrelated. Similarly, statistical models including variables representing all three aspects of teacher effectiveness may be necessary for unbiased parameter estimates, because they are likely intercorrelated to some degree and may moderate the effects of one another. Omitting one or more types from the statistical model may therefore result in the omitted variable problem, which may bias the teacher effects that are estimated.

Employing the most appropriate statistical model for studying teacher effects is important because it affects estimates of the magnitude of teacher effects as well as their standard errors. Although most recent studies on teacher effects have used multilevel models to partition the variance in student learning into student and classroom components or into student and school components, most do not correctly partition it into all three important components. Employing a multilevel model that includes only student and classroom levels will typically result in overestimating the classroom variance component, because between-school variation in the outcome will be absorbed primarily by the classroom component. Omitting the classroom level of analysis from the model has even greater statistical consequences on estimates of teacher effects. Because teacher effects are manifested at the classroom level, estimating them at the school level results in aggregation bias and other statistical problems.

Although our review above suggests that teacher effectiveness has received considerable attention in the research literature, much of the quantitative work in this area is provisional because it is based on data and models with various shortcomings. The most common and critical limitation is perhaps failing to use a data set that includes repeated achievement test scores within a single school year. Two other common shortcomings are samples with insufficiently large numbers of classrooms and data sets that do

not include a comprehensive set of measures of a teacher's background and instructional practices. These sample shortcomings lead to imprecise estimates of achievement gains that occurred while students were members of specific classrooms, to low statistical power, and to underspecified models of teacher effectiveness that are prone to biased estimates of associations. As critical as these sampling characteristics are for studying teacher effects using quantitative models, ECLS is the first large-scale National Center for Education Statistics (NCES) database that was designed to circumvent these problems. Yet it only does so in kindergarten and first grade.[9] Another noteworthy shortcoming is meticulous quantitative studies on teacher effects in the early grade levels. Not coincidentally, with the release of ECLS, this shortcoming is being alleviated, particularly at the kindergarten level (see, for example, Guarino et al., 2006, or Xue & Meisels, 2004).

This study makes a few noteworthy contributions to the literature on teacher effects. First and foremost, we develop a conceptual framework based on three components of teacher effects— teacher background qualifications, instructional practices, and attitudes—and examine the independent contribution of each. This conceptual framework not only allows relative comparisons but also puts forth a more comprehensive model with important and timely policy implications. A second contribution is the breadth of teacher variables that are examined, particularly, measures of instructional practices, which allow us to make more fine-grained inferences about teacher effects. Finally, the results are strengthened by the use of the three-level hierarchical linear model (HLM) to disentangle child, classroom, and school variance components while controlling for prior achievement. This is a precise model of teacher effects in that it isolates the variance in student learning that can be attributed to teacher effectiveness.

Methods

Data Source

This study used data from the ECLS, which sampled approximately 20,000 kindergarteners enrolled in more than 1,000 public and private schools in the fall of 1998 (NCES, 2002) and followed them as they progressed through fifth grade. Achievement tests were administered to students near the beginning and end of kindergarten and first grade as well as near the end of the third and fifth grades. Surveys with questions about a wide range of family, school, and community characteristics; about teacher background, attitudes, and practices; and about classroom and school composition were collected from parents, teachers, and principals at the same points in time. The sampling design involved oversampling certain subgroups (e.g., Asian children). NCES also developed weights for various samples, designed to make them nationally representative.

This study uses a longitudinal sample of first graders from the ECLS. One issue with the first-grade ECLS data is that the fall data collection was limited to an approximately 30% sample of schools. This resulted in a first-grade longitudinal sample of 5,034 students;[10] the present study uses a subsample of this group. The students not included in this subsample were omitted for a few reasons. First, students who changed teachers during first grade were omitted. This was necessary because the purpose of this study was to estimate teacher effects based on student achievement gains, and if the child was not in a single teacher's classroom for the duration of the school year, it was not possible to determine what part of the child's achievement gain was attributable to a specific teacher. Students without teacher or school IDs and students repeating kindergarten during the 2nd year of the ECLS were also omitted, as were a small percentage of students who met the above criteria but had missing test scores. Our final sample included 3,496 students, 887 classrooms, and 253 schools.[11]

To investigate whether these selection criteria biased our sample, a comparison of the weighted, full first-grade longitudinal sample and the weighted sample used in this study was conducted on key variables. Table 1 shows comparisons of means and standard deviations from these two samples on the achievement variables and SES. Both the means and standard deviations of each variable are highly similar across samples, suggesting that the final sample can be considered approximately nationally representative. The *t*-test results do suggest, however, that

116

TABLE 1

Comparison of Variable Means and Standard Deviations From Full First-Grade Longitudinal Sample and Final Sample

Variable name	Full sample[a] (N = 5,034)		Final sample (N = 3,496)		t statistic[b] p value
	M	SD	M	SD	
SES	–0.08	0.78	–0.08	0.77	0.994
Math (fall)	32.43	9.60	32.74	9.58	0.054
Reading (fall)	37.99	12.60	38.12	12.55	0.544
Math (spring)	42.99	9.50	43.26	9.44	0.089
Reading (spring)	54.68	14.21	54.98	14.25	0.216

Note. All means and standard deviations are weighted using the normalized first-grade panel weight (normalized C3C4cw0). SES = socioeconomic status.
a. The first grade panel is an approximately 31% random sample of schools.
b. The one-sample *t* test compares final sample means with population values based on full sample. No means differ significantly at $\alpha = .05$.

students who scored low on the math achievement test tend to be underrepresented in the final sample, although this does not significantly bias the means at the $\alpha = .05$ level. Although this comparison suggests that the sample used in this study is nationally representative, some caution is in order when making inference from models using the math outcome to the population of U.S. first graders.

Theoretical Framework

Variable selection and model building were guided by a multilevel theoretical framework that recognizes the variation in student achievement gains due to three distinct and nested levels: the school, classroom, and student levels. This framework also divides the schooling process into three sequential stages, including inputs, processes, and outputs. In Figure 1, we provide examples of the measures representative of each level and each stage, and Appendix A provides a comprehensive list of variables that were used in this study to explore the association between aspects of teachers and achievement gains.

The horizontal arrows shown in Figure 1 indicate a sequential flow from left to right. That is, inputs affect both processes and outcomes, whereas processes affect only outcomes. Inputs are conceptualized as aspects of students' and teachers' backgrounds that they bring into the school with them. Inputs include classroom and school contextual factors, resources, and structures that may be associated with effectiveness. In other words, inputs may be considered aspects of the students, classrooms, and schools that potentially affect the achievement gains of students and are largely beyond the control of teachers and schools. Processes are the practices and behaviors employed at each level as well as the values, attitudes, and beliefs that may contribute to the learning climate. Outputs are the outcome measures used to gauge classroom and school effects. In general, the main interest in classroom and school effects research is in the association between the processes and outputs. Yet because those associations may depend on inputs to some degree, inputs are an important aspect of the framework.

The vertical arrows indicate interlevel influences. For example, school processes may affect the attitudes and practices of teachers at the classroom level. Note that there are solid and dashed interlevel arrows. The solid arrows indicate a potential causal influence, and the dashed arrows indicate an association due to aggregation. An example of the aggregation type is the arrow between the student and classroom outputs. The classroom outputs are aggregate measures of the student outputs (e.g., mean achievement gains).

Statistical Models

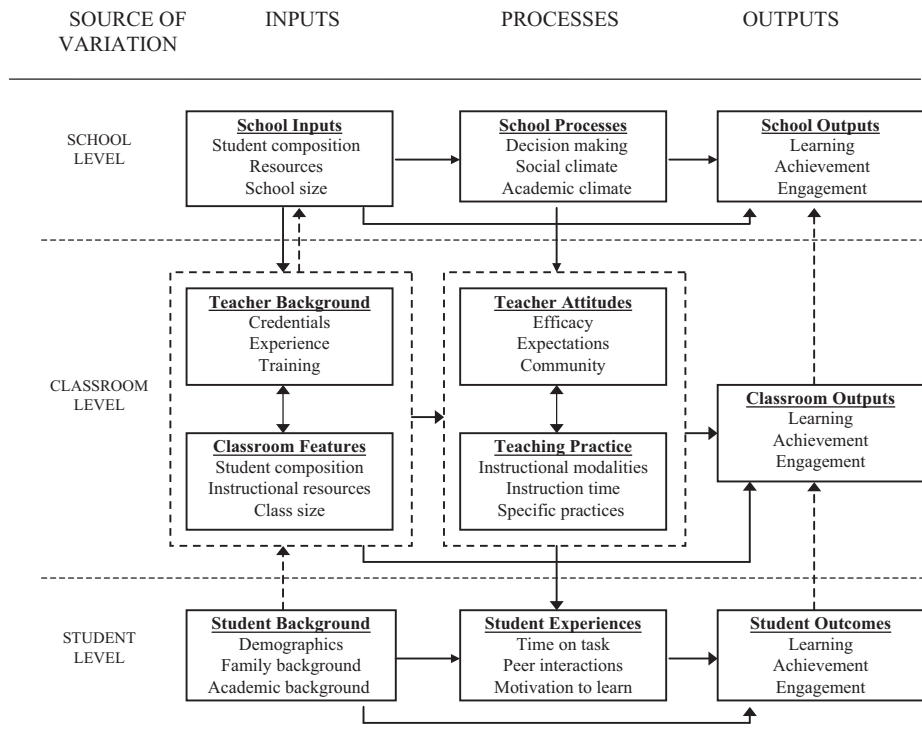Because students in the ECLS data are nested within classrooms and classrooms are nested in

117

| | SOURCE OF VARIATION | INPUTS | PROCESSES | OUTPUTS |

FIGURE 1. *A multilevel theoretical framework of classroom and school effects.*

schools, we used a three-level HLM.[12] HLM methods have been developed in the past 20 years to deal with issues specific to nested or multilevel data sets including aggregation bias, misestimation of errors, and the unit of analysis problem (Raudenbush & Bryk, 2002). HLM is highly suitable for isolating the variation in student achievement gains due to classroom or teacher effects, and isolating such variation is necessary for correctly modeling the association between aspects of teacher and student achievement gains.[13] It is important to note that the HLMs estimated in this study are consistent with our conceptual framework shown in Figure 1.

Two outcomes—math and reading achievement—were used to study the effects of teacher quality on student achievement gains in first grade. ECLS provides achievement test scores on these variables at two points—near the beginning and near the end of first grade—which allows for the estimation of achievement gains or learning but presents limitations for the use of the more desirable growth model.[14] We use the

spring achievement test score as the outcome and the fall score as a covariate, which is sometimes referred to as a *residual gain score model*.[15] Both the outcome and the prior achievement covariate were standardized to a mean of 0 and standard deviation of 1. Similarly, we also standardized all the continuous predictor variables used in the analysis. Standardizing the variables in this manner provides the benefit of slope coefficient estimates that are in units of effect size (i.e., the expected standard deviation gain in achievement per standard deviation increase in the predictor). Also, because the time between fall and spring test administrations was not uniform across students, a variable we call *assessment gap* was developed, which measures the time between fall and spring testing dates for each student.[16] Adding this covariate to the model adjusts the achievement gains for difference in the time between assessments across students. For a few other recent applications of this model for estimating teacher effects, see Nye et al. (2004) and Xue and Meisels (2004).

118

Model building was carried out in steps that are consistent with our theoretical framework and purposefully ordered to address our research questions. Variable selection was informed by our theoretical framework as well as previous research. Using multilevel modeling convention, we began with a fully unconditional model, which uses the spring achievement test score as the outcome but does not include covariates. This model is for the purpose of estimating the proportion of the variance in the achievement outcomes at each level of the model—within classroom, between classrooms within schools, and between school levels.[17] The second model is a residual gain score model, which includes fall achievement and assessment gap covariates at the within-classroom level. This model is used to estimate the proportion of the variance in achievement gains that are at each level of the analysis. This information can be used to assess the relative contributions of student, classroom, and school factors on achievement gains. The variance estimates from this model are used as the baseline when computing the proportion of the variance in achievement gains explained by successive models.

Third, we estimated a student model. It is important to note that in this study we are not specifically interested in the impact of student background characteristics on achievement gains, but rather, we wish to statistically control for differences in the background characteristics of students across classrooms, because those differences may contribute to classroom effects over which the teacher has little control. We include variables measuring SES, gender, and ethnicity at the student level. As mentioned above, all continuous variables were standardized and indicators were dummy coded so that slope coefficients were in units of effect size.

Fourth, we estimated a classroom composition model, which adjusts for the effect of several aggregated student characteristics at the classroom level. As was the case with the student-level model, we control for classroom composition for the purpose of equalizing classrooms on compositional factors that are believed to affect the learning environment but are not necessarily related to teacher effectiveness. We control for the mean SES, the number of students below grade level on initial (fall) achievement levels,

and variation in initial achievement. It is hypothesized that a more diverse student body in terms of initial achievement represents a more challenging teaching environment, which will undermine achievement gains. The classroom context model reduces the variance in mean classroom achievement gains down to what could be roughly expected if classrooms had similar student inputs. This model provides a foundation for estimating teacher effects because it adjusts for differences in both the individual and compositional effects of the students enrolled, which can account for a substantial amount of the observed differences in achievement gains but are not necessarily due to teacher performance (Raudenbush & Willms, 1995).

The fifth model was the teacher background qualifications model. Background qualifications are conceptualized as aspects of the teachers' effectiveness related to their training and level of teaching experience. These characteristics are typically accumulated before teachers enter the classroom. For this reason, the background model was estimated prior to the teacher attitude and practice model. In this model, we estimated the impact of variables such as years of experience, advanced degree,[18] and certifications held.

The sixth and final model was use to investigate the impact of teacher attitudes and practices on achievement gains after controlling for student background, classroom composition, and teacher background. The results of this model are used to determine which attitude and practice variables matter most as well as to determine whether teacher attitudes and practices can mediate the impact of teacher background and classroom composition. We used three classes of instructional practice variables. The first class measured the amount of time spent on general instruction in reading and math, including homework assigned. The second class measured instructional modalities, including whole-class instruction, small groups, mixed ability groups, peer tutoring, and so on. The final class consisted of variables measuring frequency of use of specific instructional approaches. By including all three classes in the model, we are testing whether variables in each specific class are associated with achievement gains while controlling for the variables in the other classes. This model-building approach provides stronger evidence

119

that significant associations are not because of omitted variables or spurious correlations. For example, if frequency of phonics instruction has a positive association with reading achievement gains, we know that that effect is not because of differences in instructional modes or the amount of reading instruction, which may be correlated with frequency of phonics instruction.[19]

The formulations for the hierarchical linear models used in this study are presented in Appendix B.

Limitations

This study uses a large, nationally representative data set to examine teacher quality. Secondary data have certain limitations, including predetermined questions that may be less than optimally focused or measured for addressing the research problem at hand, which tends to undermine the magnitude and significance of effects. Moreover, the design of the data collection generally precludes firm causal inferences. Related to this, although an extensive number of the variables were included in our statistical models to control for the nonrandom assignment of students and teachers to classrooms and schools,[20] there are still possible biases in coefficient estimates because of omitted variables. Moreover, it is well known that respondents to survey questionnaires tend to provide answers that are biased toward what is socially acceptable.[21] Although secondary data have certain limitations, ECLS is, overall, an outstanding data source for studying teachers and schools, given the repeated assessments of achievement within a single grade level— which is necessary for modeling achievement gains—and the large number of relevant teacher and classroom survey items.

Results

Classroom Effects on Achievement Status and Achievement Gains

The unconditional model results (Table 2) show that at the end of first grade, 72.5% of the variance in reading achievement and 75.3% of the variance in math achievement are between students within classrooms, whereas only 7.4% and 7.9% of the variance in reading and math achievement, respectively, are between classrooms within schools, and 20.2% of the variance in reading and 18.2% of the variance in math are between schools. Note that these estimates are not suitable for making inferences about teacher effectiveness, because students enter first grade with widely differing levels of achievement and this model does not control for those differences. These estimates of the proportion of the variance in reading achievement between classrooms are similar to those reported in another ECLS-based study, which estimated that 6.2% of the variance in reading achievement was at the classroom level in kindergarten (Xue & Meisels, 2004).

The achievement gains results are also shown in Table 2. Fall test scores and the assessment time gap were both strongly associated with spring achievement. These two variables accounted for a large proportion of the variation in student achievement status at the end of first grade. Approximately 65% of the total variance in each outcome was explained by these two variables.[22] The percentage of the total variance in achievement gains between classrooms and between schools was higher for reading than for math: 10.7% for reading versus 6.8% for math was between classrooms, and 10.4% for reading versus 6.8% for math was between schools; and 79.0% and 86.4% was at the student level for reading and math achievement gains.[23] The proportion of the variance in achievement gains at the classroom level provides an upper boundary for the degree to which differences in teacher quality across classrooms within schools affect student learning. The actual proportion is likely less than these estimates, however, because some of the variation is due to factors other than teacher quality, such as the differences in the background characteristics of the students. We used a likelihood ratio test (LRT) to determine whether the addition of the prior achievement and assessment gap variables to the model significantly improved the prediction of achievement gains.[24] The LRT for the reading model ($\Delta\chi^2_{df=2} = 3505.15$, $p < .01$) and math model

TABLE 2
Unconditional and Achievement Gains Model Results

	Unconditional (1)		Achievement gains (2)	
	Reading	Math	Reading	Math
	Coefficient estimates			
Intercept	–0.08*	–0.06	–0.03	0.00
Fall test score			0.81**	0.80**
Assessment gap			0.05**	0.09**
	Variance components			
Student level				
Variance	0.737	0.753	0.274	0.306
Percentage of total	72.47	73.97	78.96	86.44
Classroom level				
Variance	0.075	0.080	0.037	0.024
Percentage of total	7.37	7.86	10.66	6.78
School level				
Variance	0.205	0.185	0.036	0.024
Percentage of total	20.16	18.17	10.37	6.78
Total				
Variance	1.017	1.018	0.347	0.354
	Model summary			
Deviance statistic	9453.96	9517.32	5948.81	6157.90
Number of parameters	4	4	6	6
Deviance change[a]	—	—	3505.15**	3359.42**

a. Compared with previous model.
*$p < .05$. **$p < .01$.

($\chi^2_{df = 2} = 3259.42$, $p < .01$) were both highly significant, indicating that adding those two variables explained a significant proportion of the variation for both achievement outcomes.

These findings indicate that there is far more variance in student achievement gains between students within classrooms than between classrooms within schools or between schools. This comes as no surprise, because previous research has consistently indicated that there are vast individual differences in student learning rates and that classroom or school variance components are small by comparison (Bryk & Raudenbush, 1988; Coleman et al., 1966; Hill & Rowe, 1996; Rumberger & Palardy, 2005a; Scheerens & Bosker, 1997). These findings also suggest that there is greater variability in the quality of reading teaching than mathematics teaching in first grade.

The Effects of Student Background Characteristics

Next, we estimated the student background model. The objective was to control for aspects of the students' backgrounds that are associated with achievement gains but are largely outside of the control of their teachers. These are the "student inputs" represented in the conceptual framework (Figure 1). The results, shown in Table 3, indicate that both family SES and ethnicity are associated with achievement gains in both reading and math. However, student background variables account for almost no variation in achievement between classrooms within schools, which suggests that students are not assigned to classrooms within schools according to these factors in the first grade.

121

TABLE 3
Student and Classroom Composition Model Results

	Student (3)		Classroom composition (4)	
	Reading	Math	Reading	Math
Coefficient estimates				
Intercept	0.02	0.05**	0.00	0.03[†]
Fall Test Score	0.80**	0.77**	0.78**	0.75**
Assessment Gap	0.05**	0.10**	0.06**	0.10**
Student background				
Socioeconomic status (SES)	0.05**	0.04**	0.02	0.05**
Asian	–0.10*	–0.16**	–0.08	–0.20**
Black	–0.17**	–0.16**	–0.13**	–0.10*
Hispanic	–0.07[†]	–0.11**	–0.07*	–0.06
Other	–0.04	–0.12**	0.00	–0.11*
Classroom composition				
Mean SES			0.04*	—
Proportion minority			–0.04*	—
Mean math (fall)			—	0.06**
Mean reading (fall)			0.05*	—
Variance in math			—	–0.04**
Variance in reading			–0.06**	—
Variance components				
Student level	0.2731	0.3036	0.2728	0.3022
Classroom level	0.0364	0.0231	0.0298	0.0214
School level	0.0280	0.0190	0.0272	0.0196
Model summary				
Deviance statistic	5900.97	6105.18	5847.53	6079.92
Number of parameters	11	11	15	13
Deviance change[a]	47.84**	52.09**	53.44**	25.26**

a. Compared with previous model.
[†]$p < .10$. *$p < .05$. **$p < .01$.

The Effects of Classroom Composition

The next model was used to control for student compositional characteristics measured at the classroom level. The compositional characteristics with negative effects on achievement gains undermined the classroom learning environment, whereas those with positive associations enhanced the learning environment. Several classroom composition variables were associated with achievement gains. The results are summarized in Table 3, Model 4. Mean SES (effect size = 0.04, $p < .05$), proportion minority (effect size = –0.04, $p < .05$), mean classroom reading achievement in the fall (effect size = 0.05, $p < .05$), and classroom variance in reading achievement measured in the fall (effect size = –0.06, $p < .01$) all had significant associations with reading achievement, controlling for prior achievement and student inputs. The greater the variation in reading achievement and the higher the percentage of minority students in the classroom, the lower the achievement gains in reading tended to be, whereas the higher the mean SES and mean level of reading achievement, the higher the reading gains tended to be. For the math gains outcome, the composition model results were different in that neither mean SES

122

nor proportion minority had significant associations. However, mean classroom math achievement in the fall (effect size $= 0.05$, $p < .05$) and classroom variance in math achievement measured in the fall (effect size $= -0.04$, $p < .01$) had significant associations with gains in math achievement.

Compositional variables accounted for a sizable percentage of the variance in student achievement gains beyond the student model. The LRT for the reading model ($\chi^2_{df = 4} = 53.44$, $p < .01$) and math model ($\chi^2_{df = 2} = 25.26$, $p < .01$) were both significant, indicating that, collectively, the compositional variables improved the prediction of both outcomes. An additional 18.1% of the classroom variance in reading gains and 7.4% of the classroom variation in math gains were explained by the compositional variables. These findings suggest that even after controlling for individual student characteristics, the student body composition in the classroom has a substantial impact on achievement gains. As was noted above for the student model results, these findings suggest that it would be unfair to evaluate the effectiveness of teachers without accounting for differences in the composition of the students in their classrooms.

Approximating the Magnitude of Teacher Effects

The proportion of the variance in achievement gains that is between classrooms for the classroom composition model is 0.090 for reading and 0.062 for math. Given that the model controls for several student background and classroom composition variables, it can be argued that the remaining between-classroom variation is due largely to differences in teacher effectiveness. Hence, the remaining proportion of the variance at the classroom level can be considered an approximation of the proportion of the variance in achievement gains due to differences in teacher effectiveness within schools or an R^2 for teacher effects. Moreover, the R^2 can be used to estimate the size of the teacher effects (Nye et al., 2004). The square root of the R^2 can be interpreted loosely as a measure of the effect of a one standard deviation increase in teacher quality or teacher effectiveness on achievement

gains. Quantified in this manner, teacher effectiveness has a 0.30 effect size for reading gains and a 0.25 effect size for math gains within schools. Although these effect sizes would be classified as small by conventional standards (e.g., Cohen, 1988), they are substantial in comparison to other factors estimated in this article and to intervention effects estimated elsewhere. For example, the teacher quality effect size for math is approximately 5 times greater than the effect of family SES found in this study and more than 2.5 times greater than the effect of a class-size reduction from 25 students to 15 students per classroom (Nye et al., 2004). Moreover, it should be noted that this estimated effect size is for one school year. Assuming that teacher effects are cumulative, having a string of two or three effective or ineffective teachers would result in a moderate to large effect on achievement gains, putting the child substantially ahead of or behind where she would otherwise be. Although these estimates of the magnitude of the teacher effects are slightly lower compared with some previous studies,[25] the minor differences can be attributed to the fact that we controlled for classroom composition effects, whereas the previous research did not. Note that our estimate does not include variation in teacher effectiveness between schools, only within schools. Since there is likely considerable variation in the quality of teachers between schools, our estimates may be considered the lower boundary of the magnitude of the teacher effects. We elaborate on this issue in the Discussion section.

Teacher Background

The next step was to examine the effects of teachers' background on achievement gains. Our conceptual framework (Figure 1) classified teachers' background characteristics (education level, certifications, etc.) as inputs, and teacher attitudes and teaching practices are considered classroom processes. Because inputs precede processes, we entered the teacher background variables into the model before entering the attitude and practice variables. The results of these models are displayed in Table 4.

Of the teacher background variables examined, having full certification (effect size $= 0.09$, $p < .01$)

123

TABLE 4
Teacher Background, Attitudes, and Practices

	Teacher background (5)		Attitudes and instructional practices (6)	
	Reading	Math	Reading	Math
		Coefficient estimates		
Teacher background				
Full certification	0.09**	—	0.11**	—
Teacher attitudes				
Teacher expectations			–0.04*	—
Teacher efficacy				–0.03†
Instructional practices				
Reading frequency			0.03**	—
Phonics			0.03†	—
Silent reading			0.03*	—
Journal writing			–0.03*	—
Writing from diction			0.03†	—
Letter names			–0.02†	
Math time			—	0.02†
Geometric manipulations			—	–0.03*
Calendar			—	0.03**
		Variance components		
Student level	0.2726	0.3022	0.2723	0.3026
Classroom level	0.0291	0.0214	0.0250	0.0195
School level	0.0268	0.0196	0.0238	0.0175
		Model summary		
Deviance statistic	5838.86	6079.92	5792.06	6059.84
Number of parameters	16	13	23	17
Deviance change[a]	8.67**	0	46.80**	20.08**

a. Compared with previous model.
†$p < .10$. *$p < .05$. **$p < .01$.

was the only variable associated with reading achievement gains during first grade, and none was associated with math achievement gains. The LRT for the reading gains outcome with full certification compared with the classroom composition was significant ($\chi^2_{df = 1} = 8.67$, $p < .01$), indicating that adding full certification to the model did improve the prediction of reading achievement gains. This variable accounted for 2.4% of the classroom-level variance.

Teacher Attitudes and Practices

The final set of models examined teacher attitudes and instructional practices. In addition to the new measures of teacher attitudes and practices, these models retained the significant variables from the prior models. Of the seven teacher attitude measures tested, only one, teacher expectations, a principal component composed of four observed variables (see Table A2 for details), was significantly associated with reading achievement gains (effect size = –0.04, $p < .05$). Reading achievement gains were lower, on average, in classrooms led by teachers who held negative expectations, such as student misbehavior and paperwork interfering with teaching, academic standards being too low, and children being incapable of learning. One teacher attitude, teacher efficacy (effect size = –0.03, $p < .10$), which is a principal component measuring whether teachers feel they make a

124

difference and are satisfied with their career, was also associated with math achievement gains.

Measures of instructional practices were categorized into three types. The first consisted of general measures of time spent on instruction for each outcome. The second focused on instructional modalities (e.g., whole class, small group). The third, and by far the most extensive, category was teacher-reported frequency of use of specific instructional practices that are commonly used in first grade (e.g., phonics for reading or geometric manipulations for math). We tested 20 specific measures of instructional practice for reading and 16 for math (see Table A2 for a detailed list of variables). Note that specific practices may have a positive or negative association with achievement gains.

One measure of instruction time and five specific measures of reading instruction frequency had statistically significant associations with reading achievement gains. No measures of instructional modality were significantly associated with reading gains, however. The general measure, reading instruction frequency, had a significant positive association with reading gains (effect size $= 0.03$, $p < .01$), as did the specific instructional measures, frequency of phonics instruction (effect size $= 0.03$, $p < .10$), frequency of silent reading (effect size $= 0.03$, $p < .05$), and frequency of writing from diction (effect size $= 0.03$, $p < .10$), whereas frequency of journal writing (effect size $= -0.03$, $p < .05$) and frequency of letter names (effect size $= -0.02$, $p < .10$) had negative associations. Teacher attitudes and practices for the reading achievement gains outcome model had a highly significant LRT ($\chi^2_{df = 7} = 46.80$, $p < .01$), indicating that together, the attitude and practice variables improved prediction of the outcome. This class of variables accounted for 14.1% of the classroom-level variance in reading achievement gains—nearly 6 times the proportion noted for teacher background.

Three specific measures of math instruction frequency had significant relationships with math achievement gains. Frequency of use of math worksheets (effect size $= 0.02$, $p < .10$) had a positive association with math achievement gains, as did frequency of work on problems with calendar (effect size $= 0.03$, $p < .01$), whereas frequency of use of geometric manipulations had a negative relationship with math

gains (effect size $= -0.03$, $p < .05$). No measures of instructional time or modality were associated with math achievement gains. The teacher attitude and practice model for the math gains outcome model also produced a significant LRT ($\chi^2_{df = 4} = 20.08$, $p < .01$), accounting for 8.9% of the classroom-level variance.

Relative Effects of Teacher Background Versus Attitudes and Practices

The last result examined the additive effect size of the significant variables in the teacher background category compared with the teacher attitudes and instructional practice category. The purpose of this comparison is to judge whether teacher background—the characteristics teachers bring into the classroom—or teacher attitudes and practices—the attitudes and practices they adopt once in the classroom—is most strongly associated with teacher effectiveness, as measured by achievement gains. Just one teacher background measure, full certification, was associated with reading gains, and the effect size was noted as 0.11. One teacher attitude and six teacher practice measures were associated with reading gains, producing an additive effect size of 0.21. For the math outcome, no teacher background variables were associated with achievement gains, whereas one attitudinal measure and three practice measures had an additive effect of 0.11.

These additive effects can be interpreted as the expected increase in achievement gains during first grade for students whose teacher expectations are one standard deviation more positive than average and who employ the positive instructional practices one standard deviation more frequently than average and the negative practices one standard deviation less frequently than average. Although a teacher with those exact characteristics may not exist, this estimate provides an approximation of the expected benefit of having a teacher one standard deviation better than average, on average for all factors. These additive estimates of the effect of attitudes and instructional practices may be considered a lower boundary. This is because there are likely other significant predictors that were not available in the ECLS data set or that were poorly measured and, as a result, yielded nonsignificant

125

associations with achievement gains. These results suggest that *teaching*—the attitudes and instructional practices teachers adopt once in the classroom—collectively have a stronger association with effectiveness than background qualifications.

Discussion

Highly Qualified Versus Highly Effective

The impetus for this investigation was the No Child Left Behind Act requirement that schools hire only "highly qualified" teachers. That provision is intended to ensure that all classrooms are led by effective teachers and to close the part of the ethnic and social class achievement gaps caused by differential access to effective teachers. One of our objectives was to examine the extent to which highly qualified teachers, as measured by their background qualifications, were effective in raising student achievement. The results suggest, at least in first grade, that students do make greater gains in reading achievement when taught by a fully certified teacher. However, we found no evidence of a link between teacher certification and math achievement gains. Moreover, two other measures of background qualifications—level of experience and whether the teacher held an advanced degree—were not associated with either the reading or the math outcome.[26]

Because this finding—that full certification and math achievement gains are not statistically associated—has important policy implications, we conducted a post hoc analysis to examine whether fully certified teachers differed from those with less than full certification on key variables. Teachers in these two certification categories used surprisingly similar classroom practices and taught students from similar family backgrounds, but their teaching backgrounds differed in predictable ways. For example, they devoted highly similar amounts of classroom time to math and reading instruction, and the average class size was roughly equal. And although fully certified teachers tended to work with more affluent students, the difference was not statistically significant. However, fully certified teachers were significantly more likely to

have earned an advanced degree (perhaps as part of their certification program) and, on average, had taught at their present school slightly longer (1.1 years, $p = .05$). To examine whether there are systematic differences in the use of effective attitudes and practices for fully certified teachers compared with less than fully certified teachers, we estimated the final teacher attitude and practice model for the math outcome again, this time with full certification included. The magnitude and significance level of the attitudes and practices exhibited almost no change, indicating that full certification did not moderate the effects of attitudes and practices on math achievement gains. In summary, the results of our post hoc analysis suggest that there is surprisingly little difference in the attitudes and practices of certified teachers compared with noncertified teachers.

These findings raise a few other questions. Namely, do the results depend on how the certification variable is coded? And why does certification matter for reading but not for math? The answer to the first question is clearer. In a second post hoc analysis, we reran the final teacher attitude and practice model for the math outcome with a set of indicators of specific certification levels, including uncertified, temporary certification, and advanced certification, with regular certification serving as the reference category. The results of this model provide evidence on whether certain specific levels of certification are associated with teacher effectiveness in math. The mean math achievement gains for students with teachers with these various levels of certification did not differ from students taught by teachers with regular certification. Why full certification matters for reading achievement gains but not math is more ambiguous. Yet as we describe below, first-grade teachers spend far more time on reading instruction. We also speculate that early elementary certification programs emphasize reading instruction methods and language development. These two factors together—greater training background and more time on instruction once in service—may heighten the importance of certification for the reading outcome. Future research is needed to address this issue more definitively.

Another objective was to examine whether two aspects of *teaching*—the attitudes and practices that are adopted once in the classroom—were predictive of effectiveness above and beyond background qualifications. We noted that six teacher attitudes and instructional practices had statistically significant associations with reading achievement gains for a combined effect size of 0.21. Compared with the effect size of full certification (0.11)—the only significant background variable—the additive effect of attitudes and instructional practices was 91% greater. These results are consistent with other findings reported above: The proportion of the classroom-level variance in reading achievement gains accounted for by attitudes and instructional practices is substantially greater than the proportion accounted for by attitudes and practices.

The results for the math achievement gains outcome were similar although systematically smaller. The combined effect of teacher attitudes and instructional practices on math achievement gains was 0.11, whereas no teacher background qualifications were associated with math achievement gains. These findings suggest that although full certification is a useful indicator for screening effective teachers of reading, teacher attitudes and instructional practices are more strongly associated with teacher effectiveness.

Teacher Effects

The results of this study verify what most people have long assumed: Teachers have a substantial impact on student learning. We estimated the size of the teacher effect on reading (effect size = 0.30) and math (effect size = 0.25) learning after controlling for student inputs and classroom composition—two factors that are known to be related to learning but largely beyond the control of teachers. We now convert these teacher effect sizes into another metric to provide an alternative perspective of their magnitudes. Note that the average achievement gain during first grade for the reading outcome is approximately 0.78 standard deviations, whereas for math, it was 0.75.[27] This can be used to approximate the expected learning differential in school years for a child with a teacher one standard deviation better than

average in terms of effectiveness compared with a child with an average teacher. The teacher effect sizes convert to more than a third of a school year $(0.30/0.78 = 0.38)$ of reading achievement gains and one third of a school year $(0.25/0.75 = 0.33)$ for math.[28] The achievement gain discrepancy could easily exceed an entire grade level in a single year if one child has a highly ineffective teacher (two or more standard deviations below average), and the other, a highly effective teacher (two or more standard deviations above average). Similarly, a string of highly effective or ineffective teachers will have an enormous impact on a child's learning trajectory during the course of Grades K through 12. Consequently, although the effect sizes for teacher quality on reading and math achievement gains are small by conventional standards, they are substantially meaningful. Likewise, although the proportion of the variance in student achievement gains at the classroom level is small in comparison to the proportion at the student level, this should not be interpreted as teacher effectiveness being largely irrelevant.

Effect Size Is a Lower Boundary

We consider these teacher effect size estimates to be lower boundaries for reasons related to the statistical models used, selectivity bias—particularly, regarding teachers selecting schools to work in—and the imperfect alignment of achievement test outcomes used in the analysis with first-grade curricula nationwide. The three-level HLM partitions the variance in achievement gains into student, classroom, and school components. The classroom-level variance component is used to estimate the teacher-quality effect sizes. That component measures the variation in mean classroom achievement gains within schools. However, there may also be variation in mean classroom gains between schools that is because of teacher effectiveness. Although the source of the between-school variation in mean achievement is generally considered to be school effects, such as principal leadership, school climate, and so on, differences in the mean effectiveness of teachers between schools may also be a contributing source. This is because teachers are not randomly assigned to schools, and the best teachers may, through

127

greater employment opportunity, tend to be drawn to districts with higher pay or better work conditions. This self-selection or nonrandom assignment of effective teachers to schools results in some degree of underestimating of the size of the teacher effect.

Lack of alignment between the content of the achievement tests and the curriculum delivered in each classroom also undermines the teacher-quality effect size. The achievement tests constructed and administered by NCES as part of the ECLS are designed to match a typical first-grade curriculum. If all teachers were assigned to follow that curriculum, the achievement test gains would provide a highly valid and reliable measure of teachers' effectiveness. However, curricula vary considerably across states and even across districts within states, which results in differences in the degree of alignment of the test to the curricula of individual classrooms. This lack of alignment is a source of modeling error that tends to undermine the teacher effect size. For these reasons, the teacher effect sizes reported here may be considered lower boundaries.

Classroom Composition Is Associated With Effectiveness

The findings of this study verified that student background (e.g., SES, ethnicity) and the composition of the classroom are strongly predictive of achievement gains. Reading achievement gains are particularly susceptible to the ethnic and socioeconomic composition of the students. These findings are consistent with a growing body of research showing that the characteristics of students in students' schools and classrooms affects their learning above and beyond the effects of their own background characteristics (Burns & Mason, 2002; Hanushek et al., 2003; Hoxby, 2000; Kahlenberg, 2001; Lee et al., 2006; Palardy, 2008; Robertson & Symons, 2003; Rumberger & Palardy, 2005a). To the degree that teacher effectiveness is measured by mean classroom achievement gains, it too is influenced by the characteristics of the students in the classroom. These results underscore the importance of controlling for student characteristics when assessing teacher effectiveness or studying teacher effects, particularly given that schools are becoming increasingly segregated along racial,

ethnic, and social class lines (Frankenberg, Lee, & Orfield, 2003).[29] Failing to do so will tend to underestimate the effectiveness of teachers serving educationally disadvantaged students. A biased system of this nature is not only unfair but also counterproductive, because it will fail to provide valid evidence of teacher effectiveness.

Policy Implications and Recommendations

The findings of this study suggest that the highly-qualified-teacher provision of the NCLB act is insufficient for ensuring that classrooms are led by highly *effective* teachers in first grade. Although full certification had the largest effect size on reading gains of any teacher measure examined in this study, it is not associated with gains in math achievement.[30] Moreover, other measures of background qualifications, such as attaining an advanced degree or 5 or more years of classroom teaching experience, were not predictive of achievement gains, suggesting that background qualifications in general are not predictive of teacher effectiveness.[31]

Not only is the highly-qualified-teacher provision insufficient for ensuring that classrooms are led by effective teachers, it may also legislate the removal of some effective and needed teachers from classrooms. Approximately 28% of the teachers in this study had less than full certification and could be removed from the classroom under NCLB.[32] Given the mixed evidence supporting full certification, removing teachers from the classroom on the basis of that criterion alone may be counterproductive in geographic areas (e.g., urban schools) and instructional specialties (e.g., special education or bilingual education) where there are teacher shortages.

The findings of this study may appear to have negative implications about the efficacy of the teacher certification process or, more generally, teacher preparation programs. For example, it would seem that if the teacher preparation process were critical training for effective teaching, the certification indicator would have a large and robust association on achievement gains. Yet such a conclusion is not merited for a few reasons. One is that most teachers who enter the classroom with less than full certification have completed some coursework toward certification

and may be close to completing the requirements. So the gap in training between teachers who are fully certified and those who are less than fully certified is considerably smaller than the gap between those who are fully certified and those with no teacher training at all. Another issue is that teacher certification requirements vary across states. Teachers who are not fully certified in one state may meet the requirements in some other states. For these reasons, full certification is an insensitive indicator for assessing level of preparation among those already employed as classroom teachers, and perhaps its lack of robust association with effectiveness should not come as a surprise. Consequently, based on the findings of this study, it would be wrong to dismiss the importance of teacher training and background qualifications for effective teaching. It may be, for example, that specific coursework or specific aspects of the directed teaching experience are critical preparation for effective teaching, whereas the rest of the certification program contributes little.[33] So although the results of this study suggest that screening teachers on the basis of background qualifications legislated by the NCLB Act will do little to ensure that classrooms are led by effective teachers of math, more research is needed to understand how teacher training may be contributing to this.

Ensuring that classrooms are led by highly effective teachers will require going beyond the screening of teachers based on background qualifications to implementing policies aimed at improving *teaching* effectiveness. Presently, considerable legislation and educational policies are directed toward criteria that qualify teachers to enter the profession and far less toward performance once in the classroom. Yet given that the results of the present study suggest that the background qualifications teachers enter the classroom with are less important than what they do once they get there, we recommend that more policy attention be directed toward efforts to improve effectiveness once teachers are in service. It seems that any serious effort to improve ineffective instruction will need an assessment component to identify teachers who need help as well as a structured in-service training program designed to improve their performance. To this end, we recommend a two-stage approach, with the first stage involving

the assessment of teachers' effectiveness and the second stage involving the improvement of instructional, attitudinal, and behavioral deficits through in-service training and mentoring.

It seems clear that teachers need to be assessed and evaluated regularly to document levels of effectiveness, to be held accountable for their performance, and to pursue informed improvement efforts. Unfortunately, research on in-service teacher evaluations shows that school-based teacher evaluations tend to be uncritical, to be based on sparse evidence, and to be of limited usefulness for improving teacher effectiveness (Loup, Garland, Ellett, & Rugutt, 1996). These authors found that approximately 99% of the tenured public school teachers working in large districts receive satisfactory or better evaluations, which are typically based on a single biennial visit by one administrator. Such assessment protocols and outcomes leave little confidence that ineffective in-service teachers are even being identified, let alone that there are concerted efforts to improve their effectiveness. For these reasons, a first step in improving teacher effectiveness is instituting regular and objective assessments designed to identify ineffective teachers.

Value-added models (VAMs) have received considerable attention in recent years as a reliable technique for estimating the effectiveness of individual teachers based on the achievement gains of their students (McCaffrey et al., 2003; Sanders & Horn, 1995). These models are particularly suitable for identifying the effectiveness of teachers in raising achievement when the curriculum is uniform across classrooms. These models may be used to annually evaluate teacher effectiveness and particularly to identify underperforming teachers. However, VAMs are less suitable for providing the type of fine-grained evaluative feedback necessary to promote positive change (Ball & Rowan, 2004). For this reason, improving ineffective teaching will likely require more than value-added assessments but also qualitative reviews of teaching performance once ineffective teachers have been identified as well as structured in-service training and mentoring to ineffective teachers.

The research literature suggests that in-service professional development programs with certain characteristics are successful in improving teacher effectiveness. The most effective

129

programs are sustained, concentrate on improving instruction, and provide active learning opportunities in an interactive environment between teachers that allows immediate and regular feedback (Darling-Hammond & McLaughlin, 1995; Desimone, Porter, Garet, Yoon, & Birman, 2002; Garet, Porter, Desimone, Birman, & Yoon, 2001). Programs with these characteristics can help detect specific ineffective practices, attitudes, and behaviors as well as help develop the skills and attitudes needed to remedy them.

Summary

By advocating a holistic conceptual framework of the effects of teachers and teaching on achievement gains and modeling that framework using a large and rich nationally representative sample of first graders—all using a sophisticated model that partitions student achievement gains into student, classroom, and school components—this study provides new evidence on teacher effectiveness with important policy implications related to NCLB. Rather than the qualifications teachers bring into the classroom, it is aspects of their teaching—practices, attitudes, and beliefs—that are most relevant to their effectiveness in first grade. These findings suggest that educational policy designed to ensure that classrooms are staffed with effective teachers should include assessment of teacher effectiveness once teachers are in service as well as systematic efforts to improve instruction. Moreover, this study highlights the need for continued research on teacher effectiveness, particularly in terms of how to best assess and improve the effectiveness of in-service teachers.

Notes

[1]In this study, *teacher effectiveness* is defined in terms of mean student achievement gains or learning.

[2]When describing the outcome variables in this study, we use the following terms interchangeably: *achievement gains*, *learning*, and *achievement controlling for prior achievement*, although strictly speaking, our model estimates the latter.

[3]For a detailed discussion of these measures, see Tschannen-Moran, Hoy, and Hoy (1998).

[4]All of the studies controlled for the prior achievement of students.

[5]It is also important to note that this estimate is not a comparison of the effect of having a quality teacher with not having a teacher at all but rather the effect of having a teacher one standard deviation above average (compared with average) during a single school year. Also, because the studies they reviewed did not employ experimental designs that, when property implemented, effectively control for all of the differences in the background characteristics of students and other factors that affect student learning and differ across classrooms, their summary conclusion may overstate the magnitude of the teacher effects.

[6]Although variation in student achievement gains is arguably the result of student, classroom, and school factors, studies of teacher effectiveness that have employed multilevel models typically limited their analysis to two levels, ignoring the third. Nye and her colleagues (Nye, Konstantopoulos, & Hedges, 2004) modeled all three levels, which arguably produces more precise estimates, but those estimates are incompatible for comparison with other studies.

[7]We use the following pairs of terms interchangeably: (a) *student level* and *within-classroom level*, (b) *classroom level* and *between-classroom level*, and (c) *school level* and *between-school level*.

[8]This result is consistent with a review of international studies that found teacher effects were about 1.5 times as large as school effects (Luyten, 2003).

[9]After first grade, students were only tested biennially.

[10]The National Center for Educational Statistics (NCES) developed a sample weight for the first-grade longitudinal student sample (C3C4cw0), which when applied transforms it into a nationally representative sample of first graders. This weight variable has been applied to the student data for all descriptive statistics and hierarchical linear modeling (HLM) analyses presented in this study. NCES did not develop a teacher weight for the first-grade longitudinal sample, however.

[11]Our sample selection procedure did not reduce the average within-classroom sample size, because some of the omitted students were the only members of their classroom and hence their classroom was also omitted. The average classroom size of our sample was 3.98, slightly higher than the full first-grade longitudinal sample of 3.26.

[12]We include the school level for the purpose of partitioning the school-level variation in achievement gains from the classroom-level variation. However, we do not include school-level variables in this study because our focus is on teacher effects. Note that omitting school-level variables does not affect classroom-level estimates, so omitting them does not confound our results.

[13]Our data sample had an average of four children per classroom. However, this did not present a serious

130

limitation for our HLM analysis, because our focus was primarily on estimating between classroom coefficients (i.e., fixed classroom or teacher effects) rather than on random slopes within classroom, which require larger average within-classroom samples for reliable estimation. Our sample of 877 classrooms was sufficiently large for estimating fixed teacher effects.

[14]Typically, growth models require a minimum of three repeated measurements on the outcome. However, in the multivariate context, a growth is possible with only two time points. See Thum (2003) for a full treatment of that model.

[15]The "residual gain score" approach has been criticized in favor of a difference score outcome for measuring change when only two repeated measurements are available (e.g., see Rogosa & Willett, 1985, 1988). It is worth noting that the multilevel difference score outcome model that includes initial achievement as a student-level covariate is equivalent to the multilevel residual gain score model used in this study. That is, all variance components and coefficients will be identical for these two models with the exception of the magnitude of the intercept coefficient. This fact makes it rather straightforward to compare these two models and empirically examine the effect of adding the prior achievement control to the difference score outcome model. We did this for the math outcome and noted that adding prior achievement changed the compositional effects but no other coefficients in our final teacher practice model. Mean SES and percentage minority both went from significant to nonsignificant when prior achievement was added to the student-level model, and mean prior achievement went from significantly negative to significantly positive. These findings suggest that adding prior achievement moderates the effects of classroom composition. This is what we suspected, particularly for mean prior achievement, because without the student-level control, the classroom-level effects capture both the within- and between-classroom sources of variation associated with prior achievement. In summary, the choice of a difference score outcome model or a residual gain score model made very little difference to the estimates of teacher effects in the present study.

[16]Ideally, the fall tests would be administered at the very beginning of the school year. In reality, the fall tests were administered an average of 1.4 months after the beginning of school and ranged from 0.5 to 4.5 months. The time between tests also varied widely, with a mean of 6.8 months and minimum and maximum values of 4.8 to 9.0 months.

[17]The unconditional model in this study is presented for the purpose of facilitating comparisons with other studies. However, because our focus is on achievement gains or learning, Model 2 is the base model for estimating the proportion of the variance in

achievement gains at each level and for estimating variance explained by subsequent models.

[18]Note that the No Child Left Behind (NCLB) highly-qualified-teacher provision requires that teachers have at least a bachelor's degree. However, of the 877 teachers in this data set, only 3 failed to meet that standard—an insufficient number for modeling the effect of having a bachelor's degree. For this reason, we use an alternative measure of education level, whether the teacher obtained a master's degree or higher, to examine the association between educational attainment of effectiveness. Thirty-three percent of the teachers in our sample earned at least a master's degree.

[19]Models 1 to 5 were initially estimated with all predictors included. However, in an effort to strike a balance between parsimony and completeness, variables that were nonsignificant at the liberal p value of .10 were removed from the model and the model was estimated again. To guard against omitting important variables, the new coefficients were examined for substantial changes in magnitude or significance and a likelihood ratio test (LRT) was conducted comparing the old and new models. None of these safeguards provided evidence that an important variable was omitted. Model 6 was built slightly differently because of the number of teacher attitude and practice measures. Variables were added in theoretically cohesive and sequential sets (e.g., attitudes, instructional time, and instructional modality) and reduced as described above.

[20]Random assignment of students and teachers to classrooms and schools is generally not feasible.

[21]One aspect of the Early Childhood Longitudinal Study that likely minimizes this type of response bias is that the surveys were handled confidentially through the mail.

[22]The proportion of the variance explained is computed by comparing the variance estimates from a base model (the unconditional model here) with the variance of the present model. This describes the proportion of the variance in the base model that is accounted for by the set of predictors in the present model. See Raudenbush and Bryk (2002) for details on this computation.

[23]The percentage of the variation between classrooms for the reading achievement gains outcome is similar to a recent study using data from Tennessee (Nye et al., 2004) that estimated 8.8%. However, they found a far higher percentage of the variation in math gains between classrooms (14.2%) compared with the present study.

[24]In general terms, the LRT is a test of the contribution of a parameter or set of parameters to the model fit (Cohen, Cohen, West, & Aiken, 2003, p. 505). This is a useful test in the present study because new variables added to the model tend be intercorrelated with each other and with variables already in the model, in which case—even though the new variables may be significant

131

predictors of the outcome—their addition may not improve the overall model fit.

[25]See Nye et al. (2004) for a summary of the literature on the magnitude of teacher effects.

[26]These results are similar to those from two other recent studies using the same data set that found no association between teacher qualifications and student achievement gains in kindergarten (Guarino, Hamilton, Lockwood, & Rathbun, 2006; Xue & Meisels, 2004).

[27]This estimate is from the fall achievement test score slope in the classroom composition models shown in Table 3.

[28]The results suggest that teacher effectiveness has a larger effect on reading than on math achievement gains. This conflicts with the literature, which suggests that children learn math mostly at school whereas reading skills are typically acquired both in school and outside of school (Nye et al., 2004). However, we found that teachers spent an average of 68% more time on reading instruction compared with math instruction (approximate daily average of 89 min per day for reading vs. 53 min per day for math). Moreover, the average daily minutes spent on reading instruction varied about 71% more across classrooms than the average daily minutes spent on math instruction (416 for reading vs. 243 for math). The far greater variation in the amount of time spent on reading instruction across classrooms likely contributed to the larger teacher effect for reading observed in this study.

[29]This does not suggest that the effects of classroom composition are completely beyond the influence of teachers. However, the results of this study suggest that teacher effects have only a weak moderating role on classroom composition. The teacher effects model reduced the significance level of both percentage minority and mean prior achievement from highly significant ($p = .02$) to marginally significant ($p = .10$). No other compositional effects had more than miniscule changes.

[30]We note that full certification generally requires a bachelor's degree. Therefore, this indicator alone generally meets the NCLB highly-qualified-teacher provision at the primary school level.

[31]We report the results of 5 or more years of experience, but a sensitivity analysis of the effect of experience suggests that its noneffect is rather robust, as tests of 3 and 8 years of experience were also nonsignificant.

[32]Note that the sample used in this study was collected in 1999–2000, before NCLB was enacted.

[33]Note, however, that a recent study found that the number of reading methods courses taken by the teacher was unrelated to reading achievement gains in kindergarten (Guarino et al., 2006).

References

Ball, D. L., & Rowan, B. (2004). Measuring instruction. *Elementary School Journal*, *105*, 3–10.

Betts, J. R., & Shkolnik, J. L. (1999). The behavioral effect of variations in class size: The case of math teachers. *Educational Evaluation and Policy Analysis*, *21*, 193–216.

Betts, J. R., Zau, A. C., & Rice, L. A. (2003). *Determinants of student achievement: New evidence from San Diego.* San Francisco: Public Policy Institute of California.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillan.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, *97*, 65–108.

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

Burstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T. H., Mirocha, J., & Guiton, G. (1995). *Validating national curriculum indicators.* Santa Monica, CA: RAND.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum.

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

Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, *8*, 305–336.

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

Darling-Hammond, L., & McLaughlin, M. W. (1995). Policies that support professional development in an era of reform. *Phi Delta Kappan*, *76*, 597–604.

Appendix A

TABLE A1
Variable Descriptions

Variable name	M	SD	Min.	Max.	Description (ECLS variable label)
				Student variables (Level 1, *N* = 3,496)	
Achievement test scores					
Reading (fall)	38.12	12.55	12.57	86.36	Fall first-grade reading item response theory scale score (c3rrscal)
Reading (spring)	54.98	14.27	16.02	88.95	Spring first-grade reading IRT scale score (c4rrscal)
Math (fall)	32.74	9.58	8.10	59.84	Fall first-grade reading IRT scale score (c3mrscal)
Math (spring)	43.26	9.44	10.38	60.50	Spring first-grade reading IRT scale score (c4mrscal)
Assessment gap	6.84	4.83	9.01	0.65	Months between fall and spring assessments
Demographic characteristics					
Socioeconomic status (SES)	−0.08	0.77	−2.17	2.88	Composite of family income, parents' educational, and occupational prestige (w1sesl)
Gender	0.49	0.50	0.00	1.00	Recoded to 0 = *male*, 1 = *female* (gender)
Asian	0.02	0.16	0.00	1.00	(race = 5)
Black	0.18	0.37	0.00	1.00	(race = 2)
Hispanic	0.15	0.39	0.00	1.00	(race = 3 or 4)
Other	0.05	0.22	0.00	1.00	(race = 6, 7, or 8)
				Teacher/Classroom variables (Level 2, *N* = 877)	
Classroom composition					
Mean SES	−0.03	0.58	−1.99	2.01	Classroom mean of (w1sesl)
Percentage minority	2.99	1.43	1.00	5.00	Percentage of minorities in class (a4pmin)
Mean math	31.92	7.87	8.82	59.84	Classroom mean on (c3mrscal)
Mean reading	37.47	10.18	12.57	85.24	Classroom mean on (c3rrscal)
Variance math	6.94	3.19	0.03	20.04	Classroom variance on (c3mrscal)
Variance reading	9.09	5.30	0.01	35.77	Classroom variance on (c3rrscal)
Teacher background					
Advanced degree	0.33	0.47	0.00	1.00	Teacher has advanced degree (b4hghstd = 4 or 5)
Experienced teacher	0.69	0.46	0.00	1.00	More than 5 years teaching experience at school (b1yrsc > 5)
Full certification	0.72	0.45	0.00	1.00	Regular or advanced certification (b4typcer = 4 or 5)

(continued)

133

Appendix A

TABLE A1 (continued)

Variable name	M	SD	Min.	Max.	Description (ECLS variable label)
					Teacher/Classroom variables (Level 2, $N = 877$)
Teacher attitudes					
Daily homework	3.81	1.03	1.00	5.00	Homework should be given daily (b4hmwk1)
Parents supportive	3.74	0.77	1.00	5.00	Parents support school staff (b4psupp)
Curriculum control	4.30	0.81	1.00	5.00	How much teachers control curriculum (b4cntrlc)
Teacher community	-0.02	1.04	-5.26	1.69	Teacher community principal component
Teacher efficacy	-0.02	1.05	-4.87	1.09	Teacher efficacy principal component
Teacher expectations	0.04	1.04	-2.24	4.12	Teacher expectations principal component
					Teacher instructional practices (times spent, modalities, and frequency of specific practices)
Instructional time					
Math frequency	4.91	0.33	1.00	5.00	How often mathematics (a40ftmth)
Math time	2.27	0.52	1.00	4.00	Time for mathematics (a4txmth)
Math homework	1.92	0.75	0.00	4.00	Time on math homework (a4timemt)
Reading frequency	4.98	0.16	3.00	5.00	How often reading and language art (a40ftrdl)
Reading time	3.47	0.68	1.00	4.00	Time for reading and language arts (a4txrdla)
Reading homework	1.92	0.75	0.00	4.00	Time on reading homework (a4timer)
Instructional modalities					
Whole class	4.11	0.77	1.00	5.00	Time spent on whole-class instruction (a4whlcls)
Small group	3.35	0.78	1.00	5.00	Time spent on teacher-directed small group activities (a4smlgrp)
Individual activities	2.69	0.73	1.00	5.00	Time spent on teacher-directed individual activities (a4indvdl)
Child selected	2.41	0.60	1.00	5.00	Time spent on child-selected activities (a4chclds)
Mixed reading groups	4.23	1.42	1.00	6.00	Freq mixed-level reading groups (a4mxdgrp)
Peer reading tutoring	3.82	1.39	1.00	6.00	Freq peer tutoring in reading (a4prtutr)
Small-group reading	3.73	1.24	1.00	6.00	Freq reading projects in small group (a4smgrpj)
Mixed math groups	4.13	1.42	1.00	6.00	Freq mixed-ability group math work (a4mxmath)
Peer math tutoring	3.60	1.40	1.00	6.00	Freq peer tutoring in math (a4peer)

(continued)

Appendix A

TABLE A1   (continued)

Variable name	M	SD	Min.	Max.	Description (ECLS variable label)
Teacher instructional practices (times spent, modalities, and frequency of specific practices)					
Freq of reading-specific instructional practices					
Name letters	4.19	1.88	1.00	6.00	Freq learning letter names (a4lernlt)
Write letters	4.99	1.41	1.00	6.00	Freq practice writing letters (a4praclt)
New vocabulary	5.53	0.68	1.00	6.00	Freq new vocabulary (a4newvoc)
Dictate stories	3.28	1.45	1.00	6.00	Freq dictate stories (a4dictat)
Phonics	5.75	0.56	1.00	6.00	Freq work on phonics (a4phonic)
Follow print	5.06	1.12	1.00	6.00	Freq follow story in print (a4seepri)
Listen to story	4.81	1.51	1.00	6.00	Freq listen to story, do not see print (a4noprnt)
Retell stories	4.41	1.01	1.00	6.00	Freq retell stories (a4retell)
Read aloud	5.55	0.67	1.00	6.00	Freq read aloud (a4readld)
Silent reading	5.62	0.74	1.00	6.00	Freq read silently (a4silent)
Write from diction	4.35	1.12	1.00	6.00	Freq write from dictation (a4wrtwrd)
Invented spelling	5.42	0.85	1.00	6.00	Freq write with invented spellings (a4invent)
Chose books	5.61	0.68	1.00	6.00	Freq chose books to read (a4chsbk)
Controlled vocabulary	5.34	0.90	1.00	6.00	Freq read controlled vocabulary (a4cntvoc)
Phonetic reading	5.11	0.99	1.00	6.00	Freq phonetic reading (a4phonet)
Pattern text	5.04	0.96	1.00	6.00	Freq read patterned text (a4pattrn)
Literature text	4.92	1.06	1.00	6.00	Freq literature-based text (a4litera)
Write stories	4.35	1.11	1.00	6.00	Freq write stories/report (a4compos)
Book project	4.10	1.06	1.00	6.00	Freq project related to book (a4doproj)
Journal writing	4.49	1.41	1.00	6.00	Freq write in journal (a4jrnl)
Freq of mathematics-specific instructional practices					
Count aloud	4.97	1.08	1.00	6.00	Freq count out loud (a4outlou)
Geometric manipulation	3.62	1.06	1.00	6.00	Freq geometric manipulative (a4geomet)
Counting manipulation	4.61	1.01	1.00	6.00	Freq counting manipulative (a4manips)
Calculator use	1.93	0.98	1.00	6.00	Freq use calculator (a4calcul)

(*continued*)

135

TABLE A1 (continued)

Appendix A

Variable name	M	SD	Min.	Max.	Description (ECLS variable label)
					Teacher instructional practices (times spent, modalities, and frequency of specific practices)
Math games	4.17	1.00	1.00	6.00	Freq math-related games (a4mthgme)
Music for math	1.87	1.05	1.00	6.00	Freq music to learn math (a4musmth)
Drama for math	1.98	1.05	1.00	6.00	Freq creative movement or drama to learn math (a4crtive)
Measuring	3.05	1.03	1.00	6.00	Freq use measuring instruments (a4rulers)
Explain solution	4.86	1.02	1.00	6.00	Freq explain/solve math (a4expmth)
Calendars	5.60	0.88	1.00	6.00	Freq calendar-related activities (a4calend)
Worksheets	5.02	1.00	1.00	6.00	Freq do math worksheets (a4mthsht)
Textbooks	4.16	1.89	1.00	6.00	Freq use math textbooks (a4mthtxt)
Partners	4.04	1.15	1.00	6.00	Freq solve math with partner (a4prtnrs)
Real life	4.39	1.04	1.00	6.00	Freq solve real-life math (a4realli)
Several solutions	3.75	1.33	1.00	6.00	Freq work on problems with several solutions (a4sevsol)
Drilling from worksheets	4.53	1.17	1.00	6.00	Freq of use of worksheets to drilling (a4drill)

Note. ECLS = Early Childhood Longitudinal Study. All continuous variables were standardized to a mean of 0 and standard deviation of 1 before conducting the analyses. The descriptive statistics shown are prior to standardization. Freq = frequency.

TABLE A2

Principal Component Measurement Models for Teacher Ratings of School Climate

Component title and ECLS label	Item description	Item loading
Community		
B4ACCPT	Staff see me as colleague	0.80
B4CNTNL	Staff learn/seek new ideas	0.81
B4SCHPL	How much teachers affect policy	0.60
Percentage of variance explained		54.6
Efficacy		
B4ENJOY	Teacher enjoys present teaching job	0.84
B4MKDIF	Teacher makes a difference in children's lives	0.76
B4TEACH	Teacher would choose teaching again	0.80
Percentage of variance explained		64.1
Expectations		
B4MISBV[a]	School-level misbehavior (noise, fighting, etc.) affects teaching	0.68
B4NTCAP	Children incapable of learning	0.66
B4PAPRW	Paperwork interferes with teaching	0.63
B4STNDLO	Academic standards too low	0.66
Percentage of variance explained		43.1

Note. All variables are coded 1 = *strongly disagree* to 5= *strongly agree* except b4schpl, which is coded 1 = *no influence* to 5 = *great deal of influence*.
a. B4MISBV is a teacher measure for which more than 80% of the variance is within schools and less than 20% between schools. This suggests that it measures primarily teacher attitudes rather than level of misbehavior at the school.

Appendix B
Statistical Models

At Level 1, the student model, a math or reading achievement outcome measured in the spring of the first grade was used. The subscript ics describes the nested structure of the data. Students (individuals), i, are nested in classrooms, c, which are nested in schools, s. The model controls for fall achievement, the assessment time gap, student socioeconomic status (SES), gender, and ethnicity (as measured by a set of dummy coded variables), resulting in a residual, e_{ics}, which can be interpreted as a "residual gain score." That is, the residual for each child describes the student's spring achievement relative to the expected spring achievement for a child who earned an average score on the fall test, controlling for differences in the assessment gap, SES, and ethnicity.

Level 1: Student Model

$$(\text{Spring Achievement})_{ics} = \pi_{0cs} + \pi_{1cs}$$
$$(\text{Fall Achievement})_{ics} + \pi_{2cs}$$
$$(\text{Assessment Gap})_{ics} + \pi_{3cs}(\text{SES})_{ics} +$$
$$\pi_{4cs}(\text{Gender})_{ics} + \pi_{5cs}(\text{Asian})_{ics} + \pi_{6cs}$$
$$(\text{Black})_{ics} + \pi_{7cs}(\text{Hispanic})_{ics} +$$
$$\pi_{8cs}(\text{Other})_{ics} + e_{ics,}$$

The intercept (π_{0cs}) is the only Level 1 coefficient set to randomly vary across classrooms. It is used as the outcome at Level 2, the teacher model, which is the primary interest of this study. The other Level 2 predictors are fixed because we are not interested in how they vary across classrooms but, rather, only wish to control for differences across classrooms. The variation in the "adjusted" intercepts allows us to study the teacher effects, adjusting for differences in the backgrounds of the students in each classroom.

137

Level 2: Teacher Model

$$\pi_{0cs} = \beta_{00s} + \beta_{01s} X_{1cs} + \cdots +$$
$$\beta_{0ps} X_{pcs} + r_{0cs}$$
$$\pi_{1cs} = \beta_{10s}$$
$$\pi_{2cs} = \beta_{20s}$$
$$\pi_{3cs} = \beta_{30s}$$
$$\pi_{4cs} = \beta_{40s}$$
$$\pi_{5cs} = \beta_{50s}$$
$$\pi_{6cs} = \beta_{60s}$$
$$\pi_{7cs} = \beta_{70s}$$
$$\pi_{8cs} = \beta_{80s}$$

$$\beta_{40s} = \gamma_{400}$$
$$\beta_{50s} = \gamma_{500}$$
$$\beta_{60s} = \gamma_{600}$$
$$\beta_{70s} = \gamma_{700}$$
$$\beta_{80s} = \gamma_{800}$$

In the intercept outcome (π_{0cs}) model at Level 2, variables were added in three successive sets. The first set was classroom context; the second, teacher background; and the third was teacher attitudes and practices. At each step, variables were added one at a time, and only variables with a *p* value of .10 were retained. The teacher variables are represented in the Level 2 model equation by X_{1cs}, which is the first variable, through X_{pcs}, which is the final variable. Continuous teacher variables were grand mean centered, and the indicators were left uncentered. This centering scheme results in the interpretation of the intercept of this model (β_{00s}) as the expected achievement gain for classrooms with average levels of the continuous variables and zero on the dummy coded variables.

The Level 3 model is the school-level model. Similar to the teacher model, only the intercept model has a random effect associated with it. The school model was not developed in this study. That is, no predictors were added to explain variance in the school-level model. The reason for this is that this study focuses on teacher effects. The school-level model in this study is included for the purpose of separating the variance in achievement gains due to school effects from that due to classroom effects in an effort to obtain a better estimate of the variance due to classroom effects and to obtain better estimates of the teacher quality coefficients.

Level 3: School Model

$$\beta_{00s} = \gamma_{000} + u_{00s}$$
$$\beta_{10s} = \gamma_{100}$$
$$\beta_{20s} = \gamma_{200}$$
$$\beta_{30s} = \gamma_{300}$$

References (continued)

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*, 81–112.

Dunkin, M. J., & Biddle, B. J. (1974). *The study of teaching.* New York: Holt, Rinehart & Winston.

Ehrenberg, R. G., & Brewer, D. J. (1994). Do school and teacher characteristics matter? Evidence from high school and beyond. *Economics of Education Review*, *13*, 1–17.

Ehrenberg, R. G., & Brewer, D. J. (1995). Did teachers' verbal ability and race matter in the 1960s? Coleman revisited. *Economics of Education Review*, *14*, 1–21.

Frankenberg, E., Lee, C., & Orfield, G. (2003). *A multiracial society with segregated schools: Are we losing the dream?* Cambridge, MA: Harvard University, Civil Rights Project.

Garet, M., Porter, A., Desimone, L., Birman, B., & Yoon, K. (2001). What makes professional development effective? Analysis of a national sample of teachers. *American Education Research Journal*, *38*, 915–945.

Goddard, R. D., Hoy, W. K., & Hoy, A. W. (2000). Collective teacher efficacy: Its meaning, measure, and impact on student achievement. *American Educational Research Journal*, *37*, 479–507.

Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effects of school resources on student achievement. *Review of Educational Research*, *66*(3), 361–396.

Guarino, C. M., Hamilton, L. S., Lockwood, J. R., & Rathbun, A. H. (2006). *Teacher qualifications, instructional practices, and reading and mathematics gains of kindergartners* (NCES 2006-031). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

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

Hanushek, E. A. (1989). The impact of differential expenditures on school performance. *Educational Researcher*, *18*, 45–62.

Hanushek, E. A. (1997). Assessing the effects of school resources on student performance: An

update. *Educational Evaluation and Policy Analysis*, *19*, 141–164.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, *18*, 527–544.

Hill, P. W., & Rowe, K. J. (1996). Multilevel modeling in school effectiveness research. *School Effectiveness and School Improvement*, *7*(1), 1–34.

Hoxby, C. (2000). *Peer effects in the classroom: Learning from gender and race variation* (NBER Working Paper W7867). Cambridge, MA: National Bureau of Economic Research.

Hoy, A. W., & Spero, R. B. (2005). Changes in teacher efficacy during the early years of teaching: A comparisons of four measures. *Teaching and Teacher Education*, *21*, 343–356.

Kahlenberg, R. D. (2001). *All together now: Creating middle-class schools through public school choice*. Washington, DC: Brookings Institution.

Lee, V. E., Burkam, D. T., Ready, D. D., Honigman, J., & Meisels, S. J. (2006). Full-day versus half-day kindergarten: In which program do students learn more? *American Journal of Education*, *112*, 163–208.

Lee, V. E., Dedrick, R. F., & Smith, J. B. (1991). The effect of social organization of schools on teachers' efficacy and satisfaction. *Sociology of Education*, *64*, 190–208.

Lee, V. E., Smith, J. B., & Croninger, R. G. (1997). How high school organization influences the equitable distribution of learning in mathematics and science. *Sociology of Education*, *70*, 128–150.

Loup, K. S., Garland, J. S., Ellett, C. D., & Rugutt, J. K. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest school districts. *Journal of Personnel Evaluation in Education*, *10*, 203–226.

Luyten, H. (2003). The size of school effects compared to teacher effects: An overview of the research literature. *School Effectiveness and School Improvement*, *14*, 31–51.

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

National Center for Education Statistics. (2002). *User's manual for the ECLS-K first grade public-use data files and electronic codebook* (NCES 2002-135). Washington, DC: Author.

Newmann, F. M., Rutter, R. A., & Smith, M. S. (1989). Organizational factors that affects school sense of efficacy, community, and expectations. *Sociology of Education*, *62*, 221–238.

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

Palardy, G. J. (2008). Differential school effects among low, middle, and high social class composition schools: A multilevel, multiple group latent growth curve analysis. *School Effectiveness and School Improvement*, *19*, 21-49.

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

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1992). Contextual effects on the self-perceived efficacy of high school teachers. *Sociology of Education*, *65*, 150–167.

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

Robertson, D., & Symons, J. (2003). Do peer groups matter? Peer group versus schooling effects on academic attainment. *Econometrica*, *70*, 31–53.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, *50*, 203–228.

Rumberger, R. W., & Palardy, G. J. (2005a). Does segregation still matter? The impact of social composition on academic achievement in high school. *Teachers College Record*, *107*, 1999–2045.

Rumberger, R. W., & Palardy, G. J. (2005b). Test scores, dropout rates, and transfer rates as alternative measures of school performance. *American Education Research Journal*, *42*, 1–42.

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

Rowan, B., Raudenbush, S. W., & Kang, S. J. (1991). Organizational design in high schools: A multilevel analysis. *American Journal of Education*, *99*, 238–266.

Sanders, W. L., & Horn, S. P. (1995). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*, 299–311.

Schacter, J., & Thum, Y. M. (2004). Paying for high- and low-quality teaching. *Economics of Education Review*, *23*, 411–430.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness research*. Oxford, UK: Pergamon.

Stasz, C., & Stecher, B. M. (2000). Teaching mathematics and language arts in reduced size and

non-reduced size classrooms. *Educational Evaluation and Policy Analysis*, *22*, 313–329.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: Free Press.

Stipek, D. (2004). Teaching practices in kindergarten and first grade: Different strokes for different folks. *Early Childhood Research Quarterly*, *19*, 548–568.

Thum, Y. M. (2003). Measuring progress towards a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods and Research*, *32*, 153–207.

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research*, *68*, 202–248.

U.S. Department of Education. (2002). *No Child Left Behind: A desktop reference.* Washington, DC: Office of Elementary and Secondary Education.

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

Wenglinsky, H. (2002, February 13). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives*, *10*(12). Retrieved October 23, 2005, from http://epaa.asu .edu/epaa/v10n12/

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

Xue, Y., & Meisels, S. J. (2004). Early literacy instruction and learning in kindergarten: Evidence from the early childhood longitudinal study—Kindergarten class of 1998–1999. *American Educational Research Journal*, *41*, 191–229.

Authors

GREGORY J. PALARDY is a faculty member in the Graduate School of Education at the University of California, Riverside. His research addresses methodological issues related to modeling teacher and school effectiveness as well as substantive and policy issues regarding how families, teachers, and schools affect educational outcomes and opportunity.

RUSSELL W. RUMBERGER is a professor of education at the Givertz Graduate School of Education at the University of California, Santa Barbara. He also serves as director of the University of California Linguistic Minority Research Institute. His research interests focus on education and work; the schooling of disadvantaged students, particularly, school dropouts and linguistic minority students; school effectiveness; and education policy.