Educational Evaluation and Policy Analysis
Fall 2004, Vol. 26, No. 3, pp. 237–257

How Large Are Teacher Effects?

Barbara Nye
Tennessee State University

Spyros Konstantopoulos
Northwestern University

Larry V. Hedges
The University of Chicago

It is widely accepted that teachers differ in their effectiveness, yet the empirical evidence regarding teacher effectiveness is weak. The existing evidence is mainly drawn from econometric studies that use covariates to attempt to control for selection effects that might bias results. We use data from a four-year experiment in which teachers and students were randomly assigned to classes to estimate teacher effects on student achievement. Teacher effects are estimated as between-teacher (but within-school) variance components of achievement status and residualized achievement gains. Our estimates of teacher effects on achievement gains are similar in magnitude to those of previous econometric studies, but we find larger effects on mathematics achievement than on reading achievement. The estimated relation of teacher experience with student achievement gains is substantial, but is statistically significant only for 2nd-grade reading and 3rd-grade mathematics achievement. We also find much larger teacher effect variance in low socioeconomic status (SES) schools than in high SES schools.

Keywords: *experiments, teacher effects, teacher experience*

THE QUESTION of whether teachers differ dramatically in their effectiveness in promoting their students' academic achievement is fundamental to educational research. If differences in teacher effectiveness are large, then identification of more effective teachers and the factors that cause them to be more effective is important both for basic research and for educational reform. If the differences in teacher effectiveness were negligible, then research would be needed to discover whether it is possible to create variations in effectiveness and how to do so. In this case, the immediate prospects for improving teacher effectiveness as the mechanism of educational reform would be less promising.

Folk knowledge suggests that the differences between the effects of teachers on individual students can be dramatic. Attributing academic success to a particular teacher we have had and speaking of them as, "a great [particularly effective] teacher" is commonplace. Yet the research evidence about teacher effects is mixed. Some research traditions (such as that involving education production functions) seem to suggest that teacher effects are negligible, while others suggest that they should be substantial.

This research was supported by two grants from the Interagency Educational Research Initiative.We thank the anonymous reviewers for helpful comments.

However these traditions of research have serious limitations.

In this article we briefly summarize some key results from two major traditions of research on teacher effects and indicate some of their limitations. We then report some new experimental evidence, which is not subject to the same shortcomings as previous research.

Education Production Function Studies

Education production function studies attempt to determine the relation of specific measured teacher or school characteristics (such as teacher experience, teacher education, class size, per pupil expenditures, etc.) with student achievement. However, because parents choose neighborhoods in which to live (and their associated schools) according to tastes and resources (Tiebout, 1956), student and family backgrounds are confounded with naturally occurring school resource characteristics. Education production function studies (e.g., Coleman, et al., 1966), attempt to statistically control for this confounding by using student and family background characteristics as covariates. A particularly important covariate is prior achievement, because it can be seen as summarizing the effects of individual background (including prior educational experiences) and family background up to that time. However, even this covariate may leave important characteristics of the student unmeasured.

Students within schools are often placed into classes or assigned to teachers based on student characteristics (such as achievement), and teachers are not randomly assigned to classes either. While this may not create an analytic problem for estimating relations at the level of the school as a unit of analysis, it creates problems when inferring the relation between characteristics of teachers and student achievement. For example suppose that more experienced teachers are assigned to classes composed of higher achieving students (e.g., as a privilege of seniority) or lower achieving students (e.g., as compensatory strategy of assigning human capital). In either case, the causal direction in the relation between teacher experience and student achievement is not that teacher experience causes achievement but the reverse. This ambiguity of causal direction is a major problem for production function studies of the effects of teacher characteristics on student achievement.

In part because of these problems, there is some controversy about the interpretation of the findings of research on education production functions. For example, the Coleman report (Coleman, et al., 1966) demonstrated that a large proportion of the variance in student achievement was explained by student background factors and that relatively little additional variance was explained by school characteristics. This finding was widely, and incorrectly, interpreted as indicating that schools and teachers made little difference in student achievement.

One influential reviewer of the production function literature finds little reason to believe that measured teacher characteristics such as educational preparation, experience, or salary are related to student achievement (Hanushek, 1986). Others argue that the studies that have been conducted suggest positive effects of some of the resource characteristics such as teacher experience and teacher education (Greenwald, Hedges, & Laine, 1996). But most reviewers of this literature agree that it is difficult to interpret the relation school or teacher characteristics and achievement, even after controlling for student background, because they may be confounded with the influences of unobserved individual, family, school, and neighborhood factors.

It is important to recognize that failure to find that some set of measured teacher characteristics are related to student achievement does not mean that all teachers have the same effectiveness in promoting achievement. It is possible that the wrong characteristics were measured (characteristics that were convenient, but unrelated to achievement) but other (as yet unmeasured) characteristics *would* be related to achievement. Even if researchers attempted to measure the right teacher characteristics, it is possible that the measurement is so poor that the relation was attenuated to the point of being negligible.

Studies of the Variation in Teacher Effects

Another analytic strategy that leads to evidence about teacher effects does not attempt to estimate the relation between specific measured characteristics of teachers, but examines the *variation between classrooms* in achievement controlling for student background. Such analyses usually use prior achievement as a covariate so they can be interpreted as measuring the variance in (residualized) student achievement gain across class-

rooms. The interpretation of these variances is that they represent the variation in achievement gain due to differences in teacher effectiveness. Such analyses assume that between-classroom variation is caused by teacher variation in effectiveness. Consistent with the studies in this tradition we operationalize teacher effects as between-classroom variance in achievement. Typically these studies calculate two regression analyses. One is a regression of student achievement on student background characteristics (including prior achievement), yielding a multiple correlation R_1. The second regression is of student achievement on the same background variables but also includes a set of teacher-specific dummy variables as predictors, yielding a multiple correlation R_2. The regression coefficients for the teacher-specific dummy variables indicate the "effects" of teachers and the difference between the two regressions in variance accounted for (the change in R^2 value or $\Delta R^2 = R_2^2 - R_1^2$) represents the proportion of variance in (residualized) student achievement gain accounted for by teacher effects. The advantage of this design is that it does not require the researcher to identify in advance, and measure adequately, the aspects of teacher behavior or other teacher characteristics that are related to achievement. Of course this design cannot identify the specific characteristics that are responsible for teacher effectiveness.

The findings from 18 analyses from seven studies of variation in teacher effects are given in Table 1. Four of these studies (Armour, 1976; Hanushek, 1971; Hanushek, 1992; Murnane & Phillips, 1981) relied on samples of poor or minority students, two (Goldhaber & Brewer, 1997;

TABLE 1
Summary of Some Previous Studies of the Magnitude of Teacher Effects

Study	Sample	Outcome	Grade	ΔR^2	ΔR
Armour, et al. 1976[1]	LA Blacks	Reading	6	0.14	0.37
Armour, et al. 1976[1]	LA Mexican	Reading	6	0.07	0.26
Goldhaber & Brewer, 1997[2]	NELS	Math	10	0.12	0.35
Hanushek, 1971[3]	White, manual	SAT	3	0.1	0.31
Hanushek, 1971[3]	White, nonmanual	SAT	3	0.09	0.29
Hanushek, 1971[3]	Mexican, manual	SAT	3	0.09	0.3
Hanushek, 1971[4]	White, manual	SAT	2	0.12	0.34
Hanushek, 1971[4]	White, nonmanual	SAT	2	0.13	0.35
Hanushek, 1971[4]	Mexican, manual	SAT	2	0.12	0.35
Hanushek, 1992[5]	Gary, IN	Vocabulary	2–6	0.16	0.4
Hanushek, 1992[5]	Gary, IN	Reading	2–6	0.1	0.32
Murnane & Phillips, 1981[6]	Mid City Blacks	Vocabulary	3	0.1	0.32
Murnane & Phillips, 1981[7]	Mid City Blacks	Vocabulary	4	0.21	0.46
Murnane & Phillips, 1981[8]	Mid City Blacks	Vocabulary	5	0.16	0.4
Murnane & Phillips, 1981[9]	Mid City Blacks	Vocabulary	6	0.21	0.46
Rivkin, et al., 2001[10]	Texas		4–6	>0.01	>0.11
Rowan, Correnti, & Miller, 2002[11]	Prospects	Reading	3–6	.03–.13	0.02–0.36
Rowan, Correnti, & Miller, 2002[11]	Prospects	Math	3–6	.06–.13	0.24–0.36

Notes.

The ΔR estimated for Goldhaber and Brewer (1997) may overestimate by as much as 0.003.

Hanushek (1971) used data from a single large California school district.

Murnane and Phillips (1981) used vectors of teacher characteristics and behavior in lieu of school dummies.

1. Covariates were 5th-grade test, sex, SES, ethnicity, health problems, attendance, and additional services received.
2. Covariates were 8th-grade test, sex, race, SES, and family structure.
3. 2nd-grade test, sex, ethnicity, family class background.
4. 1st-grade test, sex, ethnicity, family class background.
5. Pretest, sex, SES, family structure, siblings, sibling position, mother's employment.
6. 2nd-grade test, unnamed child, family, and school characteristics.
7. 3rd-grade test, unnamed child, family, and school characteristics.
8. 4th-grade test, unnamed child, family, and school characteristics.
9. 5th-grade test, unnamed child, family, and school characteristics.
10. Analysis of panel data.
11. Prior test, SES, family background, school composition.

Rowan, Correnti, & Miller, 2002) used nationally representative samples of students, and one (Rivkin, Hanushek, & Kaine, 2001) used a large sample of public school students in Texas. Some characteristics of the studies and the ΔR^2 values for 17 of the analyses range from 0.07 to 0.21, suggesting that from 7% to 21% of the variance in achievement gains is associated with variation in teacher effectiveness. The 18th analysis (Rivkin, et al., 2001) generated a somewhat smaller estimate, but used a slightly different technique than the other studies and the figure given was designed to yield a lower bound on the magnitude of teacher effects. Although the study by Rowan, Correnti, and Miller is included here, they estimated the variance accounted for by teachers directly using a hierarchical linear model analysis.

If we regard the ΔR^2 as the variance accounted for by a perfectly measured index of teacher effectiveness, then the square root of ΔR^2, namely ΔR, can be loosely interpreted as a standardized regression coefficient of student achievement on teacher effectiveness. The ΔR values for each analysis are given in the last column. By most standards, these effects are not negligible. Typical values, such as $\Delta R^2 = 0.10$ correspond to $\Delta R = 0.32$, which says that a one standard deviation increase in teacher effectiveness should increase student achievement gains by about a one third of a standard deviation. By way of comparison, the effect of one year in small classes on residualized gains estimated from the Tennessee class size experiment is about 0.1 (Nye, Hedges, & Konstantopoulos, 2001).

Unfortunately, this design is also subject to some of the same limitations as other production function studies. First, valid interpretation of its results requires that the covariates adequately control for preexisting differences (including unobservable differences that are related to achievement growth) among students assigned to different classrooms. Second, valid interpretation also requires that teachers are not assigned to classrooms on the basis of student characteristics (which may be known to the school but unavailable for use as covariates in the statistical analysis) to exaggerate or attenuate differences between classrooms in achievement gains. For example, schools might assign a particularly effective teacher to students believed to be entering a difficult period as a compensatory resource allocation strategy. Alternatively schools might assign a particularly effective teacher to students believed to have promise for unusually large achievement gains as a reward for accomplishment or a meritocratic resource allocation strategy.

Schools may have many sources of information suitable for identifying students poised for unusually large gains or losses. They include essentially everything known about the child beyond test scores and easily recorded factors such as socioeconomic status (SES), gender, and family structure. For example, an impeding divorce, change of residence, delinquency problems, problems with siblings, unemployment of parents, or adjustment problems in school may all signal potential difficulties in the next school year. Alternatively, improvements in student motivation, compliance, adjustment, or parental involvement may all signal unusually good prospects for the next school year.

Evidence from a Randomized Experiment

The problems in interpretation of both designs discussed above would be eliminated if a study were available that randomly assigned both students and teachers to classes. Random assignment of students would assure that all observable and unobservable differences between students in different classes would be no larger than would be expected by chance. Random assignment of teachers to classes would assure that any differences in teacher skill are uncorrelated with classroom achievement (although this potential problem would also be substantially mitigated by the fact that randomization of students assured that there would be no large differences of student achievement across classrooms.) Fortunately such a study exists: The Tennessee Class Size Experiment.

The Tennessee Class Size Experiment

The Tennessee Class Size Experiment or Project STAR (Student-Teacher Achievement Ratio) is discussed in detail elsewhere (see, e.g., Nye, Hedges, & Konstantopoulos, 2000). The experiment involved students in 79 elementary schools in 42 school districts in Tennessee. Within each school, Kindergarten students were randomly assigned to classrooms in one of three treatment conditions: small classes (with 13 to 17 students), larger classes (with 22 to 26 students) or larger classes with a full-time classroom aide. Teachers

were also randomly assigned to classes of different types. These assignments of students to class type were maintained through the third grade. Some students entered the study in the first grade and subsequent grades, but were randomly assigned to classes at that time. Teachers at each subsequent grade level were randomly assigned to classes as the experimental cohort passed through their grade. Districts had to agree to participate for four years, allow site visitations for verification of class sizes, interviewing, and data collection, including extra student testing. They also had to allow the research staff to randomly assign pupils and teachers to class types and to maintain the assignment of students to class types from Kindergarten through grade 3.

Since the classes within each school are initially equivalent (due to random assignment), any systematic differences in achievement among classes must be due to one of two sources: the treatment or differences in teacher effectiveness. Thus within a school, any systematic variance in achievement between classrooms that had the same treatment must be due to variations in teacher effectiveness. Because there are only a few classrooms in each school, it is necessary pool evidence of between-classroom within-school differences across schools to obtain reasonable measurement precision.

Validity of the Experiment

In the STAR experiment, as in all longitudinal large field studies, the fidelity of implementation was somewhat compromised by three factors. First, there was some overlap between the *actual* sizes of the classes assigned to be large and the *actual* class sizes of those assigned to be small. Second, there was a small amount of switching of students among class types in Kindergarten and later grades. Third, there was student attrition between Kindergarten and grade 3. Researchers investigated these threats to the validity of the experiment and concluded that they did not affect the outcome of the experiment (see Krueger, 1999; Nye, Hedges, & Konstantopoulos, 2000).

To assure the validity of the experiment, it is also crucial that random assignment effectively eliminated preexisting differences between students and teachers assigned to different classrooms. We argue that the fact that the randomization of students and teachers to classes was carried out by the consortium of researchers who carried out the experiment, and not by school personnel, enhances its credibility. However it would be desirable to check whether there were any differences on pre-existing characteristics of teachers or students among the assigned groups. Such analyses cannot of course prove that the groups did not differ on variables that were not measured, they can only make such differences less plausible by confirming that differences were not observed on the variables that *were* measured. We report on some analyses carried out to check the effectiveness of randomization in the next section.

Was the randomization effective?

To check the randomization of students to class assignment, it would be desirable to compare pretest score on student achievement across classes (but within-schools). Unfortunately, no pretest scores were collected. There are however, three student variables that should not have changed as a consequence of assignment: SES (measured as eligibility for free or reduced price lunch), ethnicity (measured as Black, Hispanic, or Asian versus White), and age. Because we are checking an assignment process that occurred within-schools, it is important to carry out these comparisons controlling for school.

Krueger (1999) examined the effectiveness of the randomization into treatment groups (small sized classes, regular sized classes, and regular sized classes with a full time aide). He found that across three variables (SES, minority group status, and age), there were no significant differences between treatment types. Krueger also found that there were no significant differences across treatment types on the teacher characteristics of race, experience, and education (and we replicated his results). However, it is still possible that even though there are no differences between classrooms across treatment types, there might be differences between classrooms that were assigned to the *same* treatment types within schools. Because teacher effects in our analyses are estimated using differences between classroom the mean achievement of classes receiving the same treatment type, it is desirable to check for equality across classrooms within treatment types within schools.

To test the hypothesis that the mean age was the same for every classroom within each treatment group within every school, we regressed

TABLE 2
P *Values for Tests of the Difference Across Classes on Pre-Assignment Student Characteristics*

Student Characteristic	Kindergarten	Grade 1	Grade 2	Grade 3
SES	0.99	0.94	0.98	0.98
Ethnicity	0.99	0.99	0.99	0.99
Age	0.23	0.09	0.25	0.10

age as an outcome on a set of dummy variables for school attended and treatment type within school, then tested whether a set of dummy variables for classroom had any effects. Since SES and minority group status are measured dichotomously (as eligibility for free or reduced price lunch and as Black, Hispanic, or Asian respectively), we used a different procedure to determine whether there were equal proportions of each SES and ethnic group in each class within each school and treatment type. We computed a chi-square test for goodness of fit for the classes within each treatment group, within each school, and pooled these chi-squares across treatment groups and schools to obtain an overall test.

Table 2 gives the p values for the analyses comparing classes within treatment types within schools. None of the 12 p values is less than 0.05 and only one was less than 0.10. If the randomization of teachers and students was successful and the 12 tests of equivalence were independent (which they were not), we might have expected 5% or 0.6 of the 12 p values to be less than 0.05, but none were. We might have expected 1.2 of the p values to be less than 0.10, and 1 was. These results are therefore consistent with what would be expected if randomization was successful.

How generalizable are these findings?

The STAR project involves a rather broad range of schools from throughout a diverse state. It includes both large urban districts and small rural ones, and a range of wealth ranging from some of the wealthiest school districts in the country to some of the poorest. Thus results obtained from the entire Project STAR sample are likely to more generalizable than studies with more circumscribed samples. However the main analyses in this article depended on a subset of the schools in the project STAR sample (those with four or more classrooms) for the information used to estimate the between-teacher-within-treatment variance component (what we call the teacher effect variance). This is because only schools with four or more classrooms in the same grade have at least two classrooms assigned to the same treatment condition. We carry out analyses on both teacher effects on student achievement (achievement status) and student achievement controlling for previous achievement (achievement gains). Table 3 shows that the subset of schools that provided information for the analyses of teacher effects on achievement gains ranged from 71% (in grade 1) to 78% (in grades 2 and 3) of the schools in the STAR sample. The subset of schools that

TABLE 3
Numbers of schools in Analysis of Teacher Effects Having Various Numbers of Classrooms

	Number of classes per school					
Grade	3	(%)	4 or more	(%)	5 or more	(%)
			Achievement Gain			
First grade	22	29.0	54	71.0	28	36.8
Second grade	16	21.6	58	78.4	28	37.8
Third grade	16	21.3	59	78.7	29	38.7
			Achievement Status			
K	29	36.7	50	63.3	24	30.3
First grade	22	28.9	54	71.1	29	38.2
Second grade	13	17.6	61	82.4	28	37.8
Third grade	15	20.0	60	80.0	30	40.0

Note. K = Kindergarten.

provided information for the analyses of achievement status is approximately the same. Table 3 also provides information about the number of schools in the STAR study with five or more classrooms in the same grade. This subset of schools provides information for some of the analyses considered later to determine whether teacher experience or teacher education explain school effects. The subset of schools with five or more classes is a much smaller proportion of the whole STAR sample (approximately 37% of the total in grades 1, 2, and 3).

Table 4 provides some demographic information on the SES composition, racial composition, and geographic location of the schools that provided information for the teacher effects analyses (schools with four or more classes), compared with the whole sample. Comparing the values in columns 2 and 3 and those in columns 5 and 6 demonstrates that this sample is very similar to the whole sample. None of these differences is larger than 5%. Table 4 also provides information about the demographic composition of the subset of schools in the STAR study with five or more classrooms in the same grade (schools that provided information for the analyses of the effects of teacher experience and teacher education). Comparing the values in columns 2 and 4

TABLE 4

Comparison of Whole STAR Sample and Two Sub-samples on Important Characteristics

	Achievement Status			Achievement Gains		
Grade	Whole sample	4 or more classes	5 or more classes	Whole sample	4 or more classes	5 or more classes
		Kindergarten				
Percent minority	32.7%	34.5%	42.2%			
Percent low SES	48.1%	47.7%	54.4%			
Average mathematics achievement (SAT)	485.7	485.8	483.8			
Average reading achievement (SAT)	436.8	436.9	437.4			
Percent of urban schools	31.0%	31.4%	41.9%			
Percent of suburban schools	21.7%	23.2%	12.7%			
Percent of rural schools	47.3%	45.4%	45.4%			
Days absent from school	10.3	9.9	10.0			
Percent of teachers with graduate degree	35.3%	34.7%	31.2%			
Percent of teachers with > 3 years of experience	80.6%	78.8%	77.4%			
N	5766	4239	2409			
		First				
Percent minority	33.9%	32.2%	42.3%	30.3%	28.3%	37.3%
Percent low SES	49.2%	45.9%	51.6%	44.9%	41.4%	46.4%
Average mathematics achievement (SAT)	531.0	531.8	530.1	535.2	536.4	535.0
Average reading achievement (SAT)	520.9	522.7	519.9	527.4	529.7	528.1
Percent of urban schools	29.9%	26.1%	39.6%	28.7%	25.4%	36.9%
Percent of suburban schools	23.9%	27.3%	19.6%	19.5%	22.4%	15.7%
Percent of rural schools	46.3%	46.6%	40.9%	51.8%	52.2%	47.3%
Days absent from school	7.5	7.5	7.5	7.3	7.3	7.3
Percent of teachers with graduate degree	34.7%	34.2%	28.6%	35.0%	34.5%	28.3%
Percent of teachers with > 3 years of experience	79.5%	80.0%	80.7%	80.9%	81.8%	83.6%
N	6377	5118	3206	4045	3193	2003

(*continued on next page*)

TABLE 4 (*Continued*)

Grade	Achievement Status			Achievement Gains		
	Whole sample	4 or more classes	5 or more classes	Whole sample	4 or more classes	5 or more classes
		Second				
Percent minority	34.2%	32.9%	42.4%	30.6%	28.1%	38.9%
Percent low SES	46.7%	44.5%	47.8%	42.4%	39.7%	44.7%
Average mathematics achievement (SAT)	581.1	581.8	579.5	584.7	586.1	583.4
Average reading achievement (SAT)	584.4	585.6	581.9	588.8	590.6	586.7
Percent of urban schools	27.9%	24.6%	31.5%	25.3%	21.0%	29.3%
Percent of suburban schools	25.4%	28.2%	28.8%	23.7%	25.8%	27.0%
Percent of rural schools	46.8%	47.2%	39.7%	51.0%	53.2%	43.7%
Percent of teachers with graduate degree	36.9%	36.4%	30.3%	36.2%	35.5%	28.5%
Percent of teachers with > 3 years of experience	86.0%	87.5%	86.0%	85.7%	87.1%	86.4%
N	5968	5186	2805	4525	3828	2120
		Third				
Percent minority	33.6%	32.3%	40.8%	31.6%	30.1%	39.1%
Percent low SES	47.1%	44.7%	49.1%	43.3%	40.8%	46.2%
Average mathematics achievement (SAT)	617.9	617.9	614.5	620.6	620.6	616.9
Average reading achievement (SAT)	615.6	615.9	613.5	618.0	618.5	615.7
Percent of urban schools	27.8%	23.2%	33.9%	26.6%	22.3%	32.5%
Percent of suburban schools	25.1%	29.5%	24.2%	23.5%	27.8%	23.2%
Percent of rural schools	47.0%	47.3%	49.4%	49.9%	50.0%	44.3%
Days absent from school	6.7	6.7	6.6	6.5	6.5	6.4
Percent of teachers with graduate degree	44.3%	43.6%	42.4%	44.3%	43.9%	42.6%
Percent of teachers with 3 years of experience	87.9%	89.3%	90.9%	87.4%	89.0%	90.7%
N	5903	5032	2809	4627	3924	2220

and those in columns 5 and 7 of Table 4 shows that there are some moderate differences (two are nearly as large as 10%) between the sub-sample of schools with 5 or more classrooms and the whole STAR sample.

How Large Are Teacher Effects?

The analyses reported here make use of the Stanford Achievement Test (SAT) reading and mathematics test scores collected from Kindergarten through grade 3 as part of Project STAR. We standardized the test score to have a mean of zero and a standard deviation of unity in each grade to simplify interpretations. Since the classes within each school are initially equivalent (due to random assignment), any systematic differences in achievement among classes must be due to one of two sources: the treatment or differences in teacher effectiveness. Thus within a school, any systematic variance in achievement between classrooms that had the same treatment must therefore be due to variations in teacher effectiveness. Because there are only a few classrooms in each school, we pool evidence of between-classroom within-school differences across schools. Since both students and teachers will vary systematically between schools, it is important to separate between-classroom, within-school variance from between-school variance in the analysis. Finally it is important to separate chance variance from systematic variance by estimating variance components using a statistical model.

The analysis used a hierarchical linear model (HLM) to specify the fixed effects and variance components of interest (see Bryk & Raudenbush, 1992). The analysis assigns a component of variance to differences between classes receiving the same type of treatment. While there are relatively few classrooms within the same school receiving the same treatment, there are enough to carry out this analysis when all three treatment types are distinguished. Thus it is possible to estimate the between-classroom but within-school-and-treatment variance component, which is our measure of the variance in teacher effects.

We estimated a three-level hierarchical linear model where the level one model is a within-classroom model. The level-2 model includes school-specific treatment effects but permits the teacher effects (or more precisely, the intercept β_{0jk} of the level-1 model) to vary across classes of the same treatment type within schools. This approach permits the estimation of the between-teacher variance within schools, net of the small class and instructional aide effects.

We carried out two sets of hierarchical linear model analyses, one to examine teacher effects on achievement *gains* and the other to examine teacher effects on achievement *status*. In each case we estimated an unconditional model with no covariates at any level and a full model. To examine teacher effects on achievement gains, the full model for achievement test score Y_{ijk} of the ith student in the jth class of the kth school (the level-1 model) was

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}\text{PRETEST}_{ijk} + \beta_{2jk}\text{FEMALE}_{ijk} + \beta_{3jk}\text{SES}_{ijk} + \beta_{4jk}\text{MINORITY}_{ijk} + \varepsilon_{ijk},$$

where PRETEST_{ijk} is the achievement test in the pervious year corresponding to that measured for Y, FEMALE_{ijk} is a dummy variable for gender, SES_{ijk} is a dummy variable for free or reduced price lunch eligibility, MINORITY_{ijk} is a dummy variable for minority group membership (indicating that the student was Black, Hispanic, or Asian), and e_{ijk} is a student-specific residual. The model used to examine teacher effects on achievement status was identical except that PRETEST was omitted from the level-1 model.

The specific model for variation of coefficients between classes within schools (the level-2 model) was

$$\beta_{0jk} = \pi_{00k} + \pi_{01k}\text{SMALL}_{jk} + \pi_{02k}\text{AIDE}_{jk} + \xi_{0jk},$$

where β_{0jk} is the intercept in level-1 model for the jth class of the kth school, π_{00k} is a school-specific intercept for school k, SMALL_{jk} is an indicator for small class size, π_{01k} is a school-specific slope for SMALL in school k, AIDE_{jk} is an indicator for having a full time classroom aide (among regular sized classes), π_{02k} is a school-specific slope for AIDE in school k, and ξ_{0jk} is classroom-specific random effect. Thus the variance of the ξ_{0jk} provides a description of the variance of average achievement gains across classes net of the effects of student gender, SES, minority group status, and treatment assignment. All other coefficients were constrained to be constant within schools, that is $\beta_{1jk} = \pi_{10k}$, $\beta_{2jk} = \pi_{20k}$, $\beta_{3jk} = \pi_{30k}$, and $\beta_{4jk} = \pi_{40k}$.

We modeled variation across schools of each of the school-specific regression coefficients as random and therefore free to vary. The level-3 model for the intercept at level-2 coefficient of the kth school were therefore

$$\pi_{00k} = \gamma_{000} + \eta_{00k},$$
$$\pi_{01k} = \gamma_{010},$$
$$\pi_{02k} = \gamma_{020},$$

where m = 0, . . . , 2, the γ_{0m0} are fixed effects and η_{00k} is a school-specific random effect. Similarly, the level-3 models for the other level-2 coefficients are

$$\pi_{10k} = \gamma_{100},$$
$$\pi_{20k} = \gamma_{200},$$
$$\pi_{30k} = \gamma_{300},$$
$$\pi_{40k} = \gamma_{400},$$

where the γ_{m00} are fixed effects and m = 1, . . . , 4. Therefore the object of the statistical analysis is to estimate the seven fixed effects (γ_{000}, γ_{010}, γ_{020}, γ_{100}, γ_{200}, γ_{300}, and γ_{400}) determining each of the seven π_{mak}'s (and therefore β_{ijk}'s), the between-classes-within-treatment-types-and-schools variance components (the variance of ξ_{0jk}), and the corresponding between-school variance components (the variance of η_{00k}). Note that for simplicity the estimates reported here are from a specification where only the classroom-specific

and school-specific intercepts were treated as random at the second and third level respectively, but other analyses using additional random effects led to quite similar results.

We conducted separate analyses for each of the two dependent variables, the SAT mathematics and reading test scores, for each of the three (in the case of teacher effects on achievement gains) or four (in the case of teacher effects on achievement status) grade levels. Note that although we had data on achievement status in Kindergarten, there was no pretest available at that grade level so no analysis of gains in Kindergarten was possible. Therefore the analysis described here for achievement gains was repeated six times and that for achievement status was repeated eight times.

Results

The results of our variance component estimates from the hierarchical linear model analyses are given in Table 5. Results for mathematics achievement are given on the left and those for reading achievement are given on the right. The top two panels in the body of the table give the results of analyses for achievement gains, while the bottom two panels of the table give the results for analyses of achievement status. In each case (that is for reading or mathematics achievement gains or status), a set of unconditional analyses for each grade are reported first, followed by a set of analyses based on the full model (with student characteristics as covariates at the student level and treatment type at the classroom level).

The results of our variance component estimates from the hierarchical linear model analyses for mathematics achievement gains are given on the top left side of Table 5. The estimated between-teacher variance components in mathematics achievement for the full model range from 0.123 to 0.135. Comparing the between-teacher variance components in mathematics achievement given in Table 5 with the ΔR^2 values in Table 1, we see that they are quite close to the median of the ΔR^2 estimates, which is 0.11.

TABLE 5
Variance Decomposition by Grade

	Mathematics			Reading		
Grade	Within-classroom	Between-classroom	Between-school	Within-classroom	Between-classroom	Between-school
			Achievement Gains			
Unconditional Model						
First grade	0.696	0.148*	0.198*	0.746	0.092*	0.209*
Second grade	0.729	0.139*	0.178*	0.740	0.096*	0.163*
Third grade	0.739	0.123*	0.168*	0.793	0.090*	0.121*
Full Model						
First grade	0.397	0.128*	0.090*	0.439	0.066*	0.097*
Second grade	0.323	0.135*	0.044*	0.312	0.068*	0.026*
Third grade	0.312	0.123*	0.048*	0.317	0.074*	0.019*
			Achievement Status			
Unconditional Model						
Kindergarten	0.709	0.126*	0.165*	0.724	0.114*	0.166*
First grade	0.698	0.131*	0.177*	0.734	0.084*	0.184*
Second grade	0.736	0.125*	0.169*	0.751	0.098*	0.152*
Third grade	0.736	0.115*	0.155*	0.800	0.088*	0.115*
Full Model						
Kindergarten	0.663	0.113*	0.155*	0.675	0.100*	0.142*
First grade	0.647	0.110*	0.106*	0.677	0.065*	0.096*
Second grade	0.673	0.108*	0.096*	0.696	0.078*	0.063*
Third grade	0.700	0.104*	0.082*	0.748	0.075*	0.041*

Note. The unconditional models include only intercepts at each level.
The full models include student-level covariates and treatment type at the classroom level.
Because of slight differences in the sample that was standardized to have unit variance, the variances do not sum to unity.
*$p < .05$.

The results of our variance component estimates from the hierarchical linear model analyses for reading achievement gains are given on the top right side of Table 5. The estimated between-teacher variance components for the full model in reading range from 0.066 to 0.074. Comparing the between-teacher variance components in reading given in Table 4 with the ΔR^2 values in Table 1, we see that they are within the range of previous estimates, but somewhat smaller than the median. Thus the results of this experiment are consistent with previous non-experimental estimates of the magnitude of teacher effects on student achievement.

It is also worth noting that the variance due to differences among teachers is substantial in comparison to the variance between schools. In reading, the between teacher variance component is over twice as large as between-school variance component at grade 2 and over three times as large at grade 3. In mathematics, the pattern is similar. This suggests that naturally occurring teacher effects are typically larger than naturally occurring school effects. Thus (at least in these data), which teacher a student happens to get within a school matters more than which school the student happens to attend. This finding is provocative because it suggests that policies that emphasize school choice fail to attend to teacher effects that may have a larger impact on student achievement gains.

It is also interesting that across all grades, the variance of the teacher effects in mathematics is much larger than that in reading. In fact, in grades 1 to 3 the variance in mathematics is nearly twice as large. This may be because mathematics is mostly learned in school and thus may be more directly influenced by teachers, or that there is more variation in how (or how well or how much) teachers teach mathematics. Reading, on the other hand, is more likely to be learned (in part) outside of school and thus the influence of school and teacher on reading is smaller, or there is less variation in how (or how well or how much) reading is taught in school.

We have also estimated the between-teacher-with-treatment-type variance components without controlling for pretest scores. This analysis estimates the variance of teacher effects on achievement status (not gains). The analysis of teacher effects on achievement status is not directly comparable to those on achievement gains. However these analyses, reported in the bottom two panels of Table 5, show that teacher effects on achievement status are similar in magnitude to teacher effects on achievement gains. However, the between-school variance components for achievement status are larger than between-school variance components for achievement gains. Thus when looking at achievement status, teacher effects are closer in magnitude to school effects than when looking at achievement gains. This indicates that for students with equal previous achievement scores teacher effects are much larger than school effects.

Can Teacher Effects be Explained by Variation in Actual Class Size?

Although the experiment had target class sizes of 13 to 17 for small classes and 22 to 26 for larger classes, there was some variation in the actual class sizes of the treatment groups and even a modest amount of overlap between the actual class sizes of the treatment groups (see Nye, Hedges, & Konstantopoulos, 2000). Given that the experiment found effects of class size, it is therefore possible that some of the teacher effects might be due to variation in the actual class sizes within treatment groups.

One way to test this hypothesis is to control for actual class size in the level-2 model used in data analysis. This approach however has the disadvantage that while target class is randomly assigned, actual class size is not and may be a result of non-random factors that may also be related to outcome. That is, any relation between actual class size and achievement may not be a causal effect.

One way to overcome this problem is to use the treatment assignment as an instrumental variable for actual class size (see, e.g., Angrist, Imbens, & Rubin, 1996). Such an analysis yields estimates of the causal effects of actual class size, and therefore estimates of teacher effects after controlling for actual class size. We carried out this analysis by using treatment assignment to predict actual class sizes and then used that predicted class size in the HLM analysis.

The teacher effect (between-teacher-within-school-and-treatment type) variance components from the instrumental variables analysis were essentially identical those obtained from the analyses reported in Table 5. None of the corresponding variance components differed by more than

2%. Therefore it appears that variation in actual class sizes cannot explain variance across teachers in student achievement or achievement gains. In other words, variation in actual class sizes within treatment groups cannot explain teacher effects in this experiment.

Can Teacher Effects be Explained by Variations in Teacher Experience or Education?

We have established that teachers do differ in effectiveness, but it would be useful to determine the characteristics of teachers that are more or less effective. Two of the characteristics that have been investigated in production function models are teacher experience and teacher education. One way to test the hypothesis that teacher experience or teacher education is related to student achievement is to control for teacher experience or education, estimate the effect of teacher experience, and determine whether the variance of teacher effects decreases. Although it would have been desirable to examine the joint effects of these two variables, the sample of schools with sufficient numbers of teachers was insufficient to do so. We did, however, examine each of these two variables separately in a subsequent analysis.

Teacher experience is often measured by the number of years of service. We hypothesized however that there would be a nonlinear effect of teacher experience, with teachers becoming more skilled after the first few years of their career, which is consistent with empirical findings of Murnane and Phillips (1981). Thus we coded teacher experience dichotomously. To determine whether variations in teacher experience accounted for teacher effects, we introduced teacher experience in the level-2 model by replacing the level-2 model for the intercept by

$$\beta_{0jk} = \pi_{00k} + \pi_{01k}\text{SMALL}_{jk} + \pi_{02k}\text{AIDE}_{jk} + \pi_{03k}\text{EXPERIENCE}_{jk} + \xi_{0jk},$$

where EXPERIENCE_{jk} is a dummy variable taking the value 0 for teachers with three years of experience or less and the value 1 for teachers with more than three years of experience. The experience effect was constrained to be constant across schools at level 3, so that $\pi_{03k} = \gamma_{030}$, otherwise the level-3 model was the same as in the previous analyses.

We also hypothesized that teacher education would also have nonlinear effects. Greenwald, Hedges, and Laine's (1996) review of production function studies found stronger relations between teacher education and student achievement in studies that coded teacher education dichotomously as Master's degree or higher (or not). To determine whether variations in teacher education explained teacher effects, we introduced the dummy variable $\text{TEACHER EDUCATION}_{ij}$ (taking the value of 1 for teachers with graduate or advanced degrees and 0 otherwise) into the level-2 model exactly as we did for EXPERIENCE, and carried out the same analysis.

Results

The estimated effects of teacher education and teacher experience are summarized in Table 6. The effects on mathematics achievement are given on the left side of the table and the effects on reading achievement are given on the right

TABLE 6
3-Level HLM Fixed Effects Estimates for Teacher Experience and Education

	Mathematics		Reading	
Grade	Experience	Education	Experience	Education
		Achievement Gains		
First grade	−0.023	0.028	0.074	−0.013
Second grade	0.089	−0.010	0.149*	0.006
Third grade	0.189*	0.093*	0.058	0.045
		Achievement Status		
Kindergarten	0.045	−0.022	0.015	−0.0003
First grade	0.022	0.041	0.069	0.004
Second grade	0.014	−0.045	0.142*	−0.042
Third grade	0.079	0.091	0.032	0.058

Note. *$p < .05$, **$p < .10$.

side of the table, while effects on achievement gains are given in the top panel of the table and the effects on achievement status are given in the bottom panel of the table. Neither teacher experience, nor teacher education explained much variance in teacher effects (never more than 5%). The estimated effect of teacher experience on achievement gains was positive in every case except grade 1 mathematics achievement, where it was negative, but near zero. The magnitudes of the estimated positive effects were not negligible, ranging from 0.06 to 0.19 standard deviations or from about one-half to slightly less than two-times the small class effect on achievement gains found in previous analyses of these data (see Nye, Hedges, & Konstantopoulos, 2001). However, the effects of teacher experience on achievement gains were statistically significant only for grade-2 reading and grade-3 mathematics achievement.

The estimated effects of teacher education on achievement gains were generally smaller than those of teacher experience, and were negligible in grades 1 and 2. The estimated effects of teacher education at grade 3 were somewhat larger (0.06 and 0.09 standard deviations in reading and mathematics, respectively), but were statistically significant only for grade-3 mathematics achievement gains.

The effects of teacher experience on achievement status were generally smaller than the corresponding effects on achievement gains, and were statistically significant only for grade-2 reading achievement and close to being statistically significant for grade-3 mathematics. None of the effects of teacher education on achievement status was statistically significant.

Do Teacher Effects Vary by School SES?

It is clear that teachers are not randomly allocated to schools. Research on teacher allocation to schools has documented that schools with high proportions of low income or minority students often have difficulty recruiting and retaining high-quality teachers (Darling-Hammond, 1995). Two recent studies provide evidence that low-income students are more likely to be exposed to less effective teachers. Krei (1998) argued that low-income urban students are more likely to be exposed to less effective teachers than other students. Langford, Loeb, and Wyckoff (2002) also concluded that low-achieving, minority, and low-income students in urban settings attend schools with less competent teachers.

The origins and exact nature of the differences in teacher quality between lower and higher income schools are unclear. However one plausible mechanism that might result in lower income schools having teachers of lower average quality is a "creaming" process. Suppose teacher quality is imperfectly correlated with pre-service characteristics, but is revealed (or developed) after individuals begin to work as teachers. Once teacher quality becomes observable, higher income schools lure high quality teachers away from lower income schools using incentives of better pay or working conditions. If this were true, one would expect more consistent teacher effects (lower between-teacher variance) in higher income schools. Teacher effects would be more inconsistent (there would be larger between-teacher variance) in lower income schools since lower income schools would experience influx and then loss of high quality teachers, while maintaining a cadre of low quality teachers. We examine within school between classroom variance in achievement in low-and high-SES schools separately to determine the variance of teacher effectiveness in the upper and lower tails of the school SES distribution.

To investigate whether teacher effects differed in higher and lower SES schools we carried out the first analysis described but restricted the sample to the schools in the upper and lower quartiles of the school SES distribution, respectively. Here school SES was defined as the proportion of the sample in the school that was eligible for free or reduced price lunch. The schools in the lower quartile had an average of 64% to 72% of students eligible for free or reduced price lunch, while the schools in the higher quartile had an average of 28% to 34% of students eligible for free or reduced price lunch. We compared the variance components from high- and low-SES schools using a normal score statistic (the difference between the estimates divided by the standard error of the difference).

Results

The variance decomposition results of our separate hierarchical linear model analyses of achievement gains for the schools in the top and bottom SES quartiles are given in Table 7. Results for low-SES schools are given on the left side of the table and results for high-SES schools are

TABLE 7
Variance Decomposition by Grade by Low- and High-SES Schools: Achievement Gains

Grade	Low-SES Schools			High-SES Schools		
	Within-classroom	Between-classroom	Between-school	Within-classroom	Between-classroom	Between-school
			Mathematics Achievement Gains			
Unconditional Model						
First grade	0.546	0.186*	0.113*	0.723	0.137*	0.112*
Second grade	0.613	0.252*	0.175*	0.765	0.083*	0.157*
Third grade	0.618	0.199*	0.067*	0.806	0.075*	0.089*
Full Model						
First grade	0.387	0.139*	0.120*	0.387	0.099*	0.024*
Second grade	0.312	0.159*[a]	0.0003	0.323	0.096*[a]	0.065*
Third grade	0.275	0.179*[a]	0.024	0.334	0.103*[a]	0.025*
			Reading Achievement Gains			
Unconditional Model						
First grade	0.534	0.123*	0.099*	0.774	0.064*	0.139*
Second grade	0.569	0.158*	0.191*	0.810	0.049*	0.059*
Third grade	0.651	0.199*	0.033*	0.833	0.007*	0.067*
Full Model						
First grade	0.386	0.098*[a]	0.099*	0.440	0.049*[a]	0.036*
Second grade	0.244	0.079*	0.019*	0.322	0.049*	0.013*
Third grade	0.297	0.140*[b]	0.004	0.342	0.038*[b]	0.013*

Note. $*p < .05$.
[a]The p value of the z statistic is $< .10$.
[b]The p value of the z statistic is $< .05$.

given on the right side of the table. The top panel of the table gives the results for analyses of mathematics achievement gains while the bottom panel of the table gives the results for reading achievement gains. Within each panel the results for both the unconditional (no covariates at any level) and the full models area given. The table shows that the between-classroom-within-schools-and treatment-type variance component (the teacher effect) is always larger in the low-SES schools. In addition, the proportion of the total variance accounted for by the teacher effect is higher in low-SES schools. The ratios of the teacher effect variances in low-SES schools to those in high-SES schools range from 1.4 to 1.7 in mathematics achievement and 1.6 to 3.7 in reading achievement (full models). Although the tests of differences between variance components in low-versus high-SES schools have low power, four of the six differences between teacher-effect variances in low-versus high-SES schools are statistically significant at the 0.10 level and one of these is significant at the 0.05 level. There is no clear pattern of differences between low- and high-SES schools in between-school variance in achievement gains. However the pattern of larger between-teacher variance (teacher effects) than between-school variance (school effects) found in the entire sample of schools also seems to hold in both high-and low-SES schools considered separately. This pattern is more pronounced in low-SES schools.

The variance decomposition results of our separate hierarchical linear model analyses of achievement status for the schools in the top and bottom SES quartiles are given in Table 8, which has the same structure as Table 7. The differences between teacher effect variance in low-versus high-SES schools on achievement status given in Table 8 are somewhat larger (and more statistically reliable) than those on achievement gains given in Table 7. All of the differences are statistically significant at the 0.10 level and all but one of these differences is significant at the 0.05 level. There is no clear pattern of differences between low-and high-SES schools in between-school variance in achievement status. However the pattern of larger between teacher variance (teacher effects) than between-school variance (school effects) on achievement status that was found in the entire sample of schools seems to

TABLE 8

Variance Decomposition by Grade by Low-and High-SES Schools: Achievement Status

Grade	Low-SES Schools			High-SES Schools		
	Within-classroom	Between-classroom	Between-school	Within-classroom	Between-classroom	Between-school
			Mathematics Achievement			
Unconditional Model						
Kindergarten	0.651	0.176*	0.296*	0.750	0.048*	0.107*
First grade	0.565	0.167*	0.110*	0.728	0.104*	0.113*
Second grade	0.595	0.191*	0.161*	0.761	0.081*	0.138*
Third grade	0.599	0.183*	0.064*	0.796	0.061*	0.085*
Full Model						
Kindergarten	0.639	0.157*[b]	0.285*	0.704	0.051*[b]	0.095*
First grade	0.549	0.146*[a]	0.094*	0.680	0.077*[a]	0.115*
Second grade	0.578	0.159*[b]	0.116*	0.735	0.064*[b]	0.145*
Third grade	0.576	0.165*[b]	0.062*	0.763	0.057*[b]	0.077*
			Reading Achievement			
Unconditional Model						
Kindergarten	0.464	0.224*	0.151*	0.903	0.033*	0.138*
First grade	0.521	0.111*	0.089*	0.774	0.042*	0.146*
Second grade	0.561	0.156*	0.129*	0.848	0.045*	0.056*
Third grade	0.672	0.171*	0.042*	0.839	0.015*	0.063*
Full Model						
Kindergarten	0.451	0.209*[b]	0.151*	0.861	0.034*[b]	0.119*
First grade	0.489	0.100*[b]	0.046*	0.726	0.037*[b]	0.132*
Second grade	0.534	0.109*[b]	0.111*	0.804	0.036*[b]	0.062*
Third grade	0.644	0.154*[b]	0.025	0.792	0.012*[b]	0.055*

Note. $*p < 0.05$.

[a]The p value of the z statistic is <.10

[b]The p value of the z statistic is <. 05

hold in low-SES schools but not in high-SES schools when each is considered separately.

Do Teacher Effects Vary by Student SES?

In the previous analysis we found that there was more variance in teacher effects in low-SES schools than in high-SES schools. One interpretation of this finding is that teachers (each of whom has their own level of effectiveness) are sorted into schools so that there is greater variance in the effectiveness of teachers in low-SES schools than in high-SES schools. In this interpretation, high-SES schools obtain (or create) teachers of more uniform effectiveness, while low-SES schools do not or cannot achieve this uniformity in teacher effects.

There is another possibility. Perhaps the teachers are not equally effective for all students, so that a teacher's effectiveness depends on the kinds of students they teach. It is certainly plausible (and widely believed) that some teachers are more effective with some kinds of students than with others. In particular, it may require more knowledge or skill to promote achievement in low-SES students, while promoting achievement in high-SES students is relatively easy. If this is the case, there could be greater variability of teacher effects on low-SES students than on high-SES students. Then there would be greater variance in net teacher effects (the average of the teacher effects for all of his or her students) in low-SES schools than in high-SES schools because low-SES schools have more low-SES students.

To investigate the possibility that there is more variance in teacher effects for low-SES students than for high-SES students, we introduced a random effect in the level-2 model for the coefficient of SES by replacing the level-2 equation for β_{3jk} by

$$\beta_{3jk} = \pi_{30k} + \xi_{3jk},$$

where ξ_{3jk} is a random effect indicating the difference in achievement of low-SES students in

the jth class of the kth school (controlling for the teacher effect for high-SES children and all of the other student-level covariates). Thus ξ_{3jk} is the degree to which the teacher effect in the jth class of the kth school differs for low-SES students and high-SES students. The SES effect was also allowed to vary randomly across schools at level 3, so that $\pi_{30k} = \gamma_{300} + \eta_{30k}$.

Results

The between-classrooms within treatment type and between-school variance components for the SES effect (the variance of the ξ_{3jk} and η_{30k}, respectively) are given in column 2 (for mathematics achievement) and column 5 (for reading achievement) of Table 9. The total between-teacher within treatment type and between-school variance components with SES random are given in columns 4 (for mathematics achievement) and 7 (for reading achievement). The total between-teacher within treatment type and between-school variance components from the analysis with SES effects fixed from Table 6 are also given (in columns 3 and 6) for reference.

Teacher-level SES effects on achievement gains

The top panel of Table 9 gives the variance components for the SES effect and between-classroom within treatment type (total teacher effect) for achievement gains. The teacher effect variance components are virtually identical whether the SES effect is fixed or random. Thus it does not appear that variations in teacher effectiveness as a function of student SES can explain differences in the variance of teacher effectiveness across schools.

Teacher-level SES effects on achievement status

The bottom panel of Table 9 gives the variance components for the SES effect and between-

TABLE 9
Variance Components Indicating Variation of SES Effect Between Classrooms (Within Treatment Types within Schools) and Between Schools

	Mathematics			Reading		
Grade	SES Effect	Teacher or school effect (SES fixed)	Teacher or school effect (SES random)	SES Effect	Teacher or school effect (SES fixed)	Teacher or school effect (SES random)
			Achievement Gains			
First grade						
Between-classrooms	0.020*	0.128*	0.127*	0.004*	0.066*	0.066*
Between-schools	0.015*	0.090*	0.089*	0.019*	0.097*	0.095*
Second grade						
Between-classrooms	0.008*	0.135*	0.134*	0.0005*	0.068*	0.069*
Between-schools	0.0002	0.044*	0.044*	0.009*	0.026*	0.027*
Third grade						
Between-classrooms	0.018*	0.123*	0.121*	0.005*	0.074*	0.075*
Between-schools	0.00006	0.048*	0.048*	0.003	0.019*	0.020*
			Achievement Status			
Kindergarten						
Between-classrooms	0.009*	0.113*	0.113*	0.005*	0.100*	0.104*
Between-schools	0.017	0.155*	0.154*	0.021	0.142*	0.137*
First grade						
Between-classrooms	0.028*	0.110*	0.107*	0.034*	0.065*	0.062*
Between-schools	0.042*	0.106*	0.108*	0.052*	0.096*	0.098*
Second grade						
Between-classrooms	0.001*	0.108*	0.109*	0.002*	0.078*	0.078*
Between-schools	0.003	0.096*	0.097*	0.010*	0.063*	0.068*
Third grade						
Between-classrooms	0.039*	0.104*	0.099*	0.015*	0.075*	0.068*
Between-schools	0.009	0.082*	0.082*	0.010	0.041*	0.042*

Note. $*p < .05$

classroom within treatment type (total teacher effect) for achievement status. As in the case of achievement gains, the teacher effect variance components are virtually identical whether the SES effect is fixed or random. This suggests that, while teacher effects vary by SES, this variation cannot explain apparent variation in mean classroom achievement.

School-level SES effects

For both achievement gains and achievement status, school effects are virtually identical whether the school-level SES effect is fixed or random. Thus it appears that school-level variation in SES effects has little impact on between school variation in either achievement gains or achievement status.

Discussion

These results suggest that teacher effects are real and are of a magnitude that is consistent with that estimated by previous studies. However we would argue that, because of random assignment of teachers and students to classrooms in this experiment, our results provide stronger evidence about teacher effects. The results of this study support the idea that there are substantial differences among teachers in the ability to produce achievement gains in their students.

If teacher effects are normally distributed, these findings would suggest that the difference in achievement gains between having a 25th percentile teacher (a not so effective teacher) and a 75th percentile teacher (an effective teacher) is over one third of a standard deviation (0.35) in reading and almost half a standard deviation (0.48) in mathematics.[1] Similarly, the difference in achievement gains between having a 50th percentile teacher (an average teacher) and a 90th percentile teacher (a very effective teacher) is about one third of a standard deviation (0.33) in reading and somewhat smaller than half a standard deviation (0.46) in mathematics. In Kindergarten the effects are comparable, but somewhat larger for reading. For example, the difference in achievement status in kindergarten between having a 50th percentile teacher and a 90th percentile teacher is about 0.40 standard deviations in reading and 0.43 standard deviations in mathematics. These effects are certainly large enough effects to have policy significance.

Sizeable as these effects may be, we argue that the effects reported here do not necessarily constitute an upper bound on teacher effects for at least two reasons. First, our design only permits the measurement of within-school variance of teacher effects. Presumably there is also a non-zero between-school component of variance of teacher effects, and because the total variance of teacher effects is the sum of within-school and between-school components, the total must be larger than either of the parts. Second, our analyses may underestimate teacher effects because it is not clear that the outcome measures in Project STAR were strongly aligned with the intention of instruction. The effects of school inputs such as teacher effectiveness are expected to be largest when the content covered during instruction is closely aligned with school outcomes such as student achievement measures (see, e.g., Walker & Schaffarzick, 1974; or Brimer et al., 1978).

This suggests that interventions to improve the effectiveness of teachers or identify effective teachers might be promising strategies for improving student achievement. However it is important to recognize that while our estimates may not be an upper bound on teacher effects themselves, one must be cautious in estimating the effect that an intervention, which is based on manipulating teacher effects might have. Consider the intervention of replacing a teacher *estimated to be* at the 25th percentile with a teacher *estimated to be* at the 75th percentile. Our calculations probably overstate the effect of such an intervention because it estimates the potential effects of interventions if a *perfect* predictor of teacher effectiveness was available. No such perfect predictors are available. Even direct empirical estimates of teacher effects, for example from value added models, would have substantial statistical estimation error, and would therefore be imperfectly correlated with true teacher effectiveness. An intervention using an imperfect correlate of teacher effectiveness as a proxy for true teacher effectiveness would have a proportionately smaller effect. For example, the difference in achievement gains between having a teacher at the 25th percentile versus the 75th percentile on a measure correlated $\rho = 0.5$ with teacher effectiveness would be only half as large as the figures cited.

The finding that teacher effects are larger than school effects has interesting implications

for improving student achievement. Many policies attempt to improve achievement by substituting one school for another (e.g., school choice) or changing the schools themselves (e.g., whole school reform). The rationale for these policies is based on the fact that there is variation in school effects. If teacher effects are larger than school effects, then policies focusing teacher effects as a larger source of variation in achievement may be more promising than policies focusing on school effects. By analogy, policies derived from a teacher effects rationale might include substituting one teacher for another (e.g., teacher choice, or teacher accountability) or changing teachers (e.g., teacher development).

The finding that teacher effects are much larger in low-SES schools suggests that the distribution of teacher effectiveness is much more uneven in low-SES schools than in high-SES schools. To put it another way, in low-SES schools, it matters more *which* teacher a child receives than it does in high-SES schools. The larger variance in teacher effectiveness in low-SES schools suggests, however, that interventions to replace less effective teachers with more effective teachers (or turning one into the other) may be more promising in low-SES schools than in high-SES schools.

The fact that teacher effects were smaller in the higher SES schools than the low-SES schools is interesting for another reason. It might be argued that the higher SES schools would have afforded greater resources to teachers and greater autonomy in deploying those resources, a situation that could accentuate the differences in teacher skill. This hypothesis seems not to be confirmed. Similarly, it might be imagined that teacher effects would be larger in the small classes than in regular sized classes, because teachers in the small classes have greater opportunity to interact with individual children. Other analyses not reported in detail here show that this is not the case. Specifically, we conducted sensitivity analysis employing the same specifications, but restricting our sample to the control group (regular size classes) to eliminate possible influence of treatment effects. The results were comparable to those reported here using all students in all types of classes.

The interpretation of these effects, and those of the other studies discussed in this article, could be compromised if teacher effects were cumulative and some of those effects were unobserved at the end of the year. For example, first-grade students experience both the effects of their Kindergarten teacher and their first-grade teacher. In the present study students are equated (on both observable and unobservable factors) at the time of assignment by randomization. However, teacher effects occur after randomization and therefore teacher effects on test scores accumulate over time. If the effects of teachers are entirely captured by end-of-year test scores, then our analysis, which is based on residualized gains should yield pure estimates of teacher effects on achievement gains. On the other hand if effects of previous teachers are only observed later, then our analyses may overestimate the variance of teacher effects (see Appendix). However if there were substantial teacher effects that were not immediately observable, we might expect teacher effect variances to grow from year to year. The fact that the estimated teacher effect variances in Table 5 do not appear to increase systematically with grade level, suggests that this may not be a substantial concern.

One could argue that the best evidence for teacher effects is given by the variance components in Kindergarten, since only in that grade were all students randomly assigned to classes with no other possible lagged teacher effects from previous years. It is unfortunate that previous achievement was not available in Kindergarten, and hence was not included in our specifications. However, in principle randomization of students within schools should make adjustment for previous achievement unnecessary.

It is tempting to compare the potential of interventions based on teacher effectiveness with other potential strategies for educational improvement such as class-size reduction. The effect of a one standard deviation change in teacher effectiveness is larger than, for example, that of reducing class size from 25 to 15 (Nye, Hedges, & Konstantopoulos, 2001). Moreover the costs of such class size reduction are very high. Recently, Krueger (2003) conducted a cost benefit analysis of Project STAR and concluded that the minimum cost effective gain from class-size reduction of the magnitude undertaken in Project STAR would be one tenth of a standard deviation. It is unclear what the costs of improv-

ing teacher effectiveness by one standard deviation might be. However, if the cost of such an intervention is comparable to that of reducing class size, our findings suggest that the positive effect is at least as large and therefore such an intervention to improve teacher quality would be cost effective.

While the present analysis supports the finding that teacher effects are large enough to be important, it was less successful in identifying teacher characteristics that could be used to predict which teachers are more effective. While the estimated effects of teacher experience on student achievement gains were positive and non-negligible in five out of six cases examined, only two of these teacher experience effects (in second-grade reading and third-grade mathematics) were statistically significant. Teacher-education effects tended to be smaller and were statistically significant only in third-grade mathematics. However, it is important to recognize that this experiment was not designed to detect the effects of teacher characteristics, and consequently it is not a particularly sensitive instrument to detect such effects. This point is illustrated by the fact that estimated effects of teacher experience that were large enough to be substantively meaningful could not be identified as statistically significant. Nevertheless, this study does provide the first evidence from a large-scale randomized experiment that teacher experience is (at least sometimes) related to student achievement gains.

Note

[1]The teacher effect variance on reading achievement gains rounds to $\sigma^2 = 0.07$ in every grade, which implies a teacher effect standard deviation of $\sigma = 0.26$. The difference between the 25th and 75th percentiles of the standard normal distribution is 1.34 standard deviations, so the difference in class mean reading achievement between a 25th and 75th percentile teacher is $(1.34)(0.26) = 0.35$. The other calculations are analogous.

Appendix

If teachers have effects that are not entirely captured in end-of-year test scores, our analysis could overestimate the variance of teacher effects. This is because the estimates of teacher effects in the first grade include the effect of the first-grade teacher plus whatever effect the Kindergarten teacher had on grade one achievement is not captured by the end-of-Kindergarten test score. Similarly, the estimate of the second-grade teacher effect includes the second-grade teacher effect plus whatever component of first-grade teacher effect on second-grade test scores is not included in the end-of-first grade test score and whatever component of Kindergarten-teacher effect is not captured by the Kindergarten-or first-grade test scores. Finally, the estimate of third-grade teacher effect includes third-grade teacher effect plus whatever portion of second-grade teacher effect is not included in the end-of-second-grade test score, plus whatever component of first-grade teacher effect is not included in either the first-or second-grade test scores, plus whatever component of Kindergarten-teacher effect is not captured in by the Kindergarten-, first-, or second-grade test score.

Symbolically, let ξ^K be the Kindergarten teacher effect for a particular class on Kindergarten-test score, ξ^{K01} be the portion of the Kindergarten teacher effect on first-grade achievement that is not included in the Kindergarten test score (the part of that effect observed only at first grade). Let ξ^1 be the true first-grade teacher effect and let ξ^{1E} be the first-grade teacher effect that is estimated. Therefore

$$\xi^{1E} = \xi^{K01} + \xi^1$$

Since teachers are assigned at random both of the pairs (ξ^{K1}, ξ^1) and (ξ^{K01}, ξ^1) are independent. The variance of the estimated teacher effects is therefore

$$\mathrm{Var}(\xi^{1E}) = \mathrm{Var}(\xi^1) + \mathrm{Var}(\xi^{K01}).$$

Defining ξ^2 as the actual effect of the second-grade teacher on second-grade tests score, ξ^{K02} as the portion of the Kindergarten teacher effect on second-grade test score not included in either Kindergarten-or first-grade test score, ξ^{102} as the portion of first-grade teacher effect on second-grade test score not included in first-grade test score, we have

$$\xi^{2E} = \xi^{K02} + \xi^{102} + \xi^2$$

and thus the estimated variance of second-grade teacher effects is

$$\mathrm{Var}(\xi^{2E}) = \mathrm{Var}(\xi^{K02}) + \mathrm{Var}(\xi^{102}) + \mathrm{Var}(\xi^2).$$

In a similar way and using analogous notation, the estimated variance of third grade teacher effects is

$$\mathrm{Var}(\xi^{3E}) = \mathrm{Var}(\xi^{K03}) + \mathrm{Var}(\xi^{103}) + \mathrm{Var}(\xi^{203}) + \mathrm{Var}(\xi^{3}).$$

Whether this problem is serious or not depends on how large the lagged effects of teachers (ξ^{K01}, ξ^{102}, ξ^{203}, ξ^{K02}, etc.) are in comparison to the effect of the current teacher (ξ^{1}, ξ^{2}, or ξ^{3}). Logic suggests that these effects should be considerably smaller than the effects of the current teacher and may be negligible. Some researchers using the Tennessee Value Added Assessment System contend that effects of previous teachers on achievement status (not gains) can be observed for up to five years (see Sanders, 1998; Topping & Sanders, 2000). However their analyses suggest that these effects are independent of gains in student achievement.

References

Angrist, J., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91,* 444–472.

Armour, D. T. (1976). *Analysis of the school preferred reading program in selected Los Angeles minority schools.* R-2007-LAUSD. Santa Monica, CA: Rand Corporation.

Brimer, A., Madaus, F. G., Chapman, B., Kallaghan, T., & Wood, R. (1978). *Sources of difference in school achievement.* Windsor, Berks. SL4 1DF: NFER Publishing Company.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Thousand Oaks, CA: Sage Publications.

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

Darling-Hammond, L. (1995). Inequality and access to knowledge. In J. A. Banks (Ed.), *The handbook of research on multicultural education.* New York: Macmillan.

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *The Journal of Human Resources, 32,* 505–523.

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

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement; estimation using micro data. *American Economic Review, 61,* 280–288.

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

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: estimation using micro data. *American Economic Review, 61,* 280–288.

Hanushek, E. A. (1992). The tradeoff between child quantity and quality: Some empirical evidence. *Journal of Political Economy, 100,* 84–117.

Krei, M. S. (1998). Intensifying the barriers: The problem of inequitable teacher allocation in low-income urban schools. *Urban Education, 33,* 71–94.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics, 114,* 497–532.

Krueger, A. B. (2003). Economic considerations and class size. *Economic Journal, 113,* 34–63.

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

Murnane, R. J., & Phillips, B. R. (1981). What do effective teachers of inner-city children have in common? *Social Science Research, 10,* 83–100.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The effects of small classes on achievement: The results of the Tennessee class size experiment. *American Educational Research Journal, 37,* 123–151.

Nye, B., Hedges, L. V., and Konstantopoulos, S. (2001). Are the effects of small classes cumulative? Evidence from a Tennessee experiment. *The Journal of Educational Research, 94,* 336–345.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2001). *Teachers, schools, and academic achievement.* NBER Report.

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

Sanders, W. L. (1998). Value added assessment. *The School Administrator, 55*(11), 24–32.

Tiebout, C. M. (1956). A pure theory of local expenditures. *Journal of Political Economy, 64,* 416–424.

Topping, K. J., & Sanders, W. L. (2000). Teacher effectiveness and computer assessment of reading: Relating value added and learning information system data. *School Effectiveness and School Improvement, 11,* 305–337.

Walker, D. F., & Schaffarzick, J. (1974). Comparing Curricula. *Review of Educational Research, 44,* 83–111.

Authors

BARBARA NYE is Director, Center of Excellence for Research and Policy on Basic Skills, Tennessee State University, 3301 10th Ave., Box 141, Nashville, TN 37203; bnye@coe.tsuniv.edu. Her areas of specialization are research on class size, teacher education, science education and policy issues.

SPYROS KONSTANTOPOULOS is an Assistant Professor of Human Development and Social Policy, and Learning Sciences at the School of Education and Social Policy at Northwestern University, 2120 Campus Drive, Evanston, IL 60208; spyros@northwestern.edu. He specializes in the application of statistical methods in social science research, particularly research on the effects of class size, and school and teacher effect research.

LARRY V. HEDGES is Stella M. Rowley Professor of Education Psychology, Sociology, The University of Chicago, 5835 Kimbark Ave., Chicago, IL 60637; lhedges@uchicago.edu. His areas of specialization are the application of statistical methods in education research and educational policy analysis, particularly research on the effects of class size and educational technology, school effects research, and research on motivation.

Manuscript received April 14, 2003
Revision received March 15, 2004
Accepted July 19, 2004