How Large Are Teacher Effects?

By

Barbara Nye

Tennessee State University

Spyros Konstantopoulos

Northwestern University

Larry V. Hedges

The University of Chicago

Note: This study was supported by a grant funded jointly by the NSF and ESR, # REC-9987943.

Abstract

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. We also find much larger teacher effect variation in low SES schools than in high SES schools.

How Large Are Teacher Effects?

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 are 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 seem to suggest that teacher effects are negligible, while others suggest that they should be substantial. However these traditions of research have serious limitations.

In this paper 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

The Coleman report (Coleman, et al., 1966) was a watershed event in educational research in the United States. This monumental study 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. The Coleman report initiated a tradition of research on so-called education production functions, in which the relation of specific school resource characteristics (such as teacher experience, teacher education, class size, per pupil expenditures, etc.) with achievement is examined.

It is essential to control for student background characteristics because students are not randomly assigned to schools. Parents choose neighborhoods in which to live (and their associated schools) according to tastes and resources (Tiebout, 1956). Thus student and family backgrounds are confounded with naturally occurring school resource characteristics. Education production function studies attempt to statistically control for the 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, but 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 the problems mentioned above, there is some controversy about the interpretation of the findings of research on education production functions. One influential reviewer of this work 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 examined (Greenwald, Hedges, and 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 (ones 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 variation in (residualized) student achievement gain across classrooms. 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 variation. Typically these studies calculate two regression analyses. One is a regression of student achievement on student background characteristics, including prior achievement. The second regression is of student achievement on the same background variables but also includes a set of teacher-specific dummy variables as predictors. These regression coefficients for the 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 ΔR^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 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; Hanusek, 1971; Hanushek, 1992; Murnane and Phillips, 1981) relied on samples of poor or minority students, two (Goldhaber and Brewer, 1997; Rowan, Correnti, and Miller, 2002) used nationally representative samples of students, and one (Rivkin, et al., 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.

Insert Table 1 Here

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 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 characteristics 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 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 there was no differential assignment of teachers to classrooms to exaggerate or ameliorate existing differences (although this potential problem would also be substantially addressed by the fact that randomization of students assured that there would be no large differences 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, and Konstantopoulos, 2000). The experiment ultimately involved students in 79 elementary schools in 42 school districts in Tennessee. Within each school students were randomly assigned into 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 and teachers 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. 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 random assignment of pupils and teachers to class types from Kindergarten through grade 3.

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 switching of students among class types in Kindergarten and in grades 1 to 3. Finally, there was student attrition between Kindergarten and grades 1, 2, and 3. The effects of these threats to the validity of the experiment have investigated by other researchers who concluded that they did not affect the outcome of the experiment (see Krueger, 1999; Nye, Hedges, & Konstantopoulos, 2000).

The STAR project involves a rather broad range of schools from throughout a rather 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 the results of this study are likely to more generalizable than studies with more circumscribed samples.

Allocation of Teachers to Schools

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 is unclear. However one plausible mechanism that might result in lower income schools having teachers of lower average quality is a "creaming" process. In such a process teacher quality is revealed in their work. 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 variation) in higher income schools and more inconsistent teacher effects (larger between-teacher variation) in lower income schools. We examine within school classroom variation in achievement in low and high SES schools separately to determine the importance of teacher effectiveness in the upper and lower tails of the school SES distribution.

Method

The analyses reported here make use of the SAT reading and mathematics test score collected in Kindergarten through grade 3 as part of Project STAR, and we standardized the test score to have a mean of 0 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 variation 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 variation from between-school variation in the analysis. Finally it is important to separate chance variation from systematic variation by estimating variance components using a statistical model.

The natural method for carrying out such an analysis is the use of a hierarchical linear model (HLM) (see Bryk and Raudenbush, 1992). The appropriate analysis would assign 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.

Such a model is a three level hierarchical linear model where the level one model would be a within-classroom model. The level 2 model includes school specific treatment effects but permits the teacher effects (or more precisely, the intercept β_{0j} of the level 1 model) to vary across classes of the same within the same treatment type within schools. This approach permits the estimation of the between-teacher variation 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*. To examine teacher effects on achievement gains, the specific model for achievement test score Y_{ijk} of the i^{th} student in the j^{th} class of the k^{th} school (the level one model) was

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}PRETEST_{ijk} + \beta_{2jk}FEMALE_{ijk} + \beta_{3jk}SES_{ijk}$$
$$+ \beta_{4jk}MINORITY_{ijk} + \varepsilon_{ijk},$$

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, and ε_{ijk} is a student-specific residual. The specific 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

$$\beta_{0jk} = \pi_{00k} + \pi_{01k}SMALL_{jk} + \pi_{02k}AIDE_{jk} + \xi_{0jk} ,$$

where β_{0jk} is the intercept in level 1 model for the j^{th} class of the k^{th} 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 variation 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 $\beta_{mjk} = \pi_{m0k}$ for $m > 0$.

We modeled variation across schools of each of the school-specific regression coefficients as random and therefore free to vary. The specific level 3 model for the a^{th} predictor of the m^{th} level 2 coefficient of the k^{th} school π_{mak} was therefore

$$\pi_{mak} = \gamma_{0am} + \eta_{amk},$$

where η_{amk} is a level 3 residual (random effect). Note that for $m > 0$, $a = 0$ (since $\beta_{mjk} = \pi_{m0k}$ for $m > 0$), but for $m = 0$, a can be 0, 1, or 2 (since the level 2 model for β_{0jk} has three coefficients: π_{00k}, π_{01k}, and π_{02k}). Therefore the object of the statistical analysis is to estimate the seven fixed effects (the γ_{0am}) determining each of the seven π_{mak}'s (and therefore β_{ijk}'s), the between-classes-within-treatment-types-and-schools variance components (the variances of the ξ_{0jk}'s), and the corresponding between-school variance components (the variances of the η_{mj}'s). Note that for simplicity the estimates reported here are from a specification where only the school-specific intercepts were treated as random at the third level.

We conducted separate analyses for each of the two dependent variables (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 gains) 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.

Finally, to investigate whether teacher effects differed in higher and lower SES schools we carried out the same analysis indicated here but restricted to the schools in the upper and lower quartile of student SES, 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 of student SES had an average of 64 to 72 percent of students eligible for free or reduced price lunch, while the schools in the higher quartile of student SES had an average of 28 to 34 percent of students eligible for free or reduced price lunch. We compared the variance components from high and low SES schools using a normal score test.

Results

The results of our variance component estimates from the hierarchical linear model analyses of reading achievement are given in upper panel of Table 2. The estimated between-teacher variance components in reading range from 0.066 to 0.074. Comparing the between-teacher variance components in reading given in Table 2 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.

Insert Table 2 Here

The results of our variance component estimates from the hierarchical linear model analyses for mathematics achievement are given in the lower panel of Table 2. The estimated between-teacher variance components in mathematics achievement range from 0.123 to 0.135. Comparing the between-teacher variance components in reading given in Table 2 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.

It is also worth noting that the variation due to differences among teachers is substantial in comparison to the variation 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*.

It is also interesting that across all grades the variation of the teacher effects in mathematics is much larger than that in reading. In fact in grades 1 to 3 the variation 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 variation 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 Table 3, show that teacher effects on achievement status are similar in magnitude to teacher effects on achievement gains.

Insert Table 3 Here

The results of our separate hierarchical linear model analyses for the schools in the top and bottom SES quartiles are given in Table 4. The table shows that the between-teachers-with-schools-and treatment-type variance component (the teacher effect) is always larger in the low SES schools. The ratios of these teacher effects range from 1.4

to 3.1 in mathematics achievement and 1.6 to 6.1 in reading achievement. Although the power for the tests of differences between variance components is low, four of the six differences are statistically significant at the 0.10 level and one of these is significant at the 0.05 level. It is interesting that the differences between teacher effects in low and high SES schools on achievement status given in Table 5 are even larger (an more statistically reliable) than those on achievement gains.

Insert Tables 4 & 5 Here

Discussion

These results suggest that teacher effects are real and are of a magnitude 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.

Sizeable as these effects may be, we argue that the effects reported here do not constitute an upper bound on teacher effects. Specifically, it is possible that our analyses underestimate teacher effects since 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 the largest when the content covered during instruction is closely aligned with school outcomes such as student achievement measures (see, e.g., Walker and Schaffarzick, 1974; or Brimer et al., 1978).

If teacher effects are normally distributed, these findings suggest that the difference in achievement gains between having a 25^{th} percentile teacher (a not so effective teacher) and a 75^{th} percentile teacher (an effective teacher) is over a third of a standard deviation (0.34) in reading and almost half a standard deviation (0.48) in mathematics. Similarly, the difference in achievement gains between having a 50^{th} percentile teacher (an average teacher) and a 90^{th} 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 larger for reading. For example, the difference in achievement status in kindergarten between having a 50^{th} percentile teacher and a 90^{th} percentile teacher is about 0.40 standard deviations in reading and 0.43 standard deviations in mathematics. These are certainly large enough effects to have policy significance.

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 the estimate above is an upper bound on what might be accomplished via such interventions—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, if they were possible, 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 would have a proportionately smaller effect. For example, the difference in achievement gains between having a teacher at the 25^{th} percentile versus the 75^{th} percentile on a measure correlated $\rho = 0.5$ with teacher effectiveness would be only half as large as the figures cited above.

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 much 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 paper, is potentially compromised by the fact that teacher effects may be cumulative. 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. However if teacher have effects that are not entirely captured in end of year test scores, our analysis should overestimate the variance of teacher effects and the overestimate should be greater the higher the grade level.

This is because the estimates of teacher effects in the first grade include the effect of the first grade teacher plus whatever effect of the Kindergarten teacher on grade one achievement is not captured by the end-of-Kindergarten test score. Similarly, the estimate of second grade teacher effect includes 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^{KO1} be the portion of the Kindergarten teacher effect on grade 1 achievement that is not included in the Kindergarten test score (the part of that effect observed only at grade 1). Let ξ^{1} be the true first grade teacher effect and let ξ^{1E} be the first grade teacher effect that is estimated. Therefore

$$\xi^{1E} = \xi^{KO1} + \xi^{1}$$

Since teachers are assigned at random both of the pairs (ξ^{K1}, ξ^{1}) and (ξ^{KO1}, ξ^{1}) are independent. The variance of the estimated teacher effects is therefore

$$\mathrm{Var}(\xi^{1E}) = \mathrm{Var}(\xi^{1}) + \mathrm{Var}(\xi^{KO1}).$$

Defining ξ^2 as the actual effect of the second grade teacher on second grade tests score, ξ^{KO2} as the portion of the Kindergarten teacher effect on second grade test score not included in either Kindergarten or first grade test score, ξ^{1O2} as the portion of first grade teacher effect on second grade test score not included in first grade test score, we have

$$\xi^{2E} = \xi^{KO2} + \xi^{102} + \xi^2$$

and thus the estimated variance of second grade teacher effects is

$$\mathrm{Var}(\xi^{2E}) = \mathrm{Var}(\xi^{KO2}) + \mathrm{Var}(\xi^{102}) + \mathrm{Var}(\xi^2).$$

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

$$\mathrm{Var}(\xi^{3E}) = \mathrm{Var}(\xi^{KO3}) + \mathrm{Var}(\xi^{103}) + \mathrm{Var}(\xi^{2O3}) + \mathrm{Var}(\xi^3).$$

The seriousness of the problem depends on how large the lagged effects of teachers (ξ^{KO1}, ξ^{1O2}, ξ^{2O3}, ξ^{KO2}, 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 and Sanders, 2000). However their analyses suggest that these effects are independent of gains in student achievement.

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, and Konstantopoulos, 2001). Moreover the costs of such class size reduction are very high. Recently, Kruger (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 1/10 of a standard deviation. The costs of improving teacher effectiveness by one standard deviation might well be lower than reducing class size by this much, hence leading to a more cost effective intervention. However, even if the cost of such intervention is comparable to reducing class size according to our findings the positive effect is more than 3 times as large.

While the present analysis supports the finding that teacher effects are large enough to be important, it does not provide a mechanism to identify, which teachers are more effective than others or suggest intervention that can improve teacher effectiveness. Research is needed to determine how effective teachers can be identified and what the costs of producing more effective teachers might be.

References

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: U.S. 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.

Hedges, L. V., Laine, R. D., & Greenwald, R. (1994). Does money matter?: A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher*, *23(3)*, 5-14.

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.

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	0.10	0.32
Hanushek, 1971 [3]	White,nonmanual	SAT	3	0.09	0.30
Hanushek, 1971 [3]	Mexican, manual	SAT	3	0.09	0.30
Hanushek, 1971 [4]	White, manual	SAT	2	0.12	0.35
Hanushek, 1971 [4]	White,nonmanual	SAT	2	0.13	0.36
Hanushek, 1971 [4]	Mexican, manual	SAT	2	0.12	0.35
Hanushek, 1992 [5]	Gary, IN	Vocabulary	2-6	0.16	0.40
Hanushek, 1992 [5]	Gary, IN	Reading	2-6	0.10	0.32
Murnane & Phillips,1981 [6]	Mid City Blacks	Vocabulary	3	0.10	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	0.40
Murnane & Phillips,1981 [9]	Mid City Blacks	Vocabulary	6	0.21	0.46
Rivkin, et al., 2001 [10]	Texas		4-6	> 0.01	0.10
Rowan, Correnti, & Miller, 2002 [11]	Prospects	Reading	3-6	0.03-0.13	0.17-0.36
Rowan, Correnti, & Miller, 2002 [11]	Prospects	Math	3-6	0.06-0.13	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 used vectors of teacher characteristics and behavior in lieu of school dummies.

1. Covariates 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

Table 2. Three Level HLM Estimates for Reading and Mathematics Achievement Controlling for Previous Achievement: Grades 1-3

	Grade 1		Grade 2		Grade 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Reading						
Between Teachers Within Schools	0.066	0.008	0.068	0.008	0.074	0.009
Between Schools	0.100	0.012	0.026	0.005	0.019	0.004
Mathematics						
Between Teachers Within Schools	0.128	0.015	0.135	0.016	0.123	0.014
Between Schools	0.090	0.011	0.044	0.008	0.048	0.007

Table 3. Three Level HLM Estimates for Reading and Mathematics Achievement Status: Grades K-3

	Grade K		Grade 1		Grade 2		Grade 3	
	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
Reading								
Between Teachers Within Schools	0.100	0.012	0.065	0.008	0.078	0.009	0.075	0.009
Between Schools	0.142	0.017	0.096	0.011	0.063	0.007	0.041	0.006
Mathematics								
Between Teachers Within Schools	0.113	0.013	0.110	0.013	0.108	0.013	0.104	0.012
Between Schools	0.155	0.018	0.106	0.012	0.096	0.011	0.082	0.010

Table 4
Three Level HLM Estimates of Variance Components for Achievement Gains by School SES

	Mathematics School SES			Reading School SES		
Variance Component	Bottom 25%	Top 25%	Z-Test	Bottom 25%	Top 25%	Z-Test
Grade 1						
Between Teachers Within Schools	0.139**	0.099**	1.07	0.098**	0.049**	1.81*
Between Schools	0.120**	0.024**		0.099**	0.036**	
Grade 2						
Between Teachers Within Schools	0.159**	0.096**	1.66*	0.079**	0.049**	1.41
Between Schools	0.0003	0.065**		0.019**	0.013**	
Grade 3						
Between Teachers Within Schools	0.179**	0.103**	1.86*	0.140**	0.038**	3.41**
Between Schools	0.024	0.025		0.004	0.013**	

** $p < 0.05$, * $p < 0.1$

Table 5
Three Level HLM Estimates of Variance Components for Achievement Status by School SES

Variance Component	Mathematics School SES			Reading School SES		
	Bottom 25%	Top 25%	Z-Test	Bottom 25%	Top 25%	Z-Test
Grade K						
Between Teachers Within Schools	0.157**	0.051**	2.78**	0.209**	0.034**	3.94**
Between Schools	0.285**	0.095**		0.151**	0.119**	
Grade 1						
Between Teachers Within Schools	0.146**	0.077**	1.91*	0.100**	0.037**	2.46**
Between Schools	0.094**	0.115**		0.046**	0.132**	
Grade 2						
Between Teachers Within Schools	0.159**	0.064**	2.52**	0.109**	0.036**	2.63**
Between Schools	0.116**	0.145**		0.111**	0.062**	
Grade 3						
Between Teachers Within Schools	0.165**	0.057**	2.81**	0.154**	0.012**	4.15**
Between Schools	0.062**	0.077**		0.025*	0.055**	

** $p < 0.05$, * $p < 0.1$