The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data

March, 2004

Abstract

Teacher quality is widely believed to be important for education, despite substantial but inconsistent evidence that teachers' credentials matter for student achievement. To accurately measure variation in achievement due to teachers' characteristics—both observable and unobservable—it is essential to identify teacher fixed effects. I use panel data collected from New Jersey school districts to estimate teacher fixed effects while controlling for fixed student characteristics and classroom specific variables. I find large and statistically significant differences among teachers: moving up one standard deviation in the teacher fixed effect distribution raises both reading and math test scores by approximately .1 standard deviations on a nationally standardized scale. In addition, teaching experience has statistically significant positive effects on reading test scores, controlling for fixed teacher quality.

1 Introduction

School administrators, parents, and students themselves widely support the notion that teacher quality is vital to student achievement, despite the lack of evidence linking achievement to observable teacher characteristics. Studies that estimate the relation between achievement and teachers' characteristics, including their credentials, have produced little consistent evidence that students perform better when their teachers have more 'desirable' characteristics (Hanushek, 1986). This is all the more puzzling because of the potential upward bias in such estimates—teachers with better credentials may be more likely to teach in affluent districts with high performing students (Figlio, 1997).

This has led many observers to conclude that, while teacher quality may be important, variation in teacher quality is driven by characteristics that are difficult or impossible to measure. Therefore, researchers have come to focus on using matched student-teacher data to separate student achievement into a series of "fixed effects," and assigning importance to individuals, teachers, schools, and so on. Researchers who have sought to explain wage determination have followed a similar empirical path; they try to separate industry, occupation, establishment, and individual effects using employee-employer matched data (Abowd and Kramarz, 1999). The fixed effects strategy has also been used to examine the role of managers in determining firm behavior (Bertrand and Schoar, 2003).

Despite agreement that the identification of teacher fixed effects is a productive path, this exercise has remained incomplete because of a lack of adequate data. Credible identification of teacher fixed effects requires matched student-teacher data where both student achieve-

ment and teachers are observed in multiple years. This type of data has not been readily available to researchers, in part because school districts do not use panel data for evaluation purposes.[1] In previous studies, researchers have either collected information directly from school districts (Hanushek, 1971, Murnane, 1975, Armor et al., 1975, Park and Hannum, 2002, Uribe et al. 2003) or used data collected by a research institution (Rivkin et al., 2003, Aaronson et al., 2003). All of these studies present evidence that student achievement is affected by the quality of their teachers. Almost all of the empirical difficulties in these studies are related to data quality. For instance, in several of these studies teacher effects cannot be separated from other classroom specific factors because teachers are only observed with one class of students.

In order to provide more accurate estimates of how much teachers affect the achievement of their students, I collected information on test scores and teacher assignment in two contiguous New Jersey school districts. This data links teachers with their students for a period of up to twelve years, contains annual test scores for all students across a number of elementary grades, and covers more than ten elementary schools. Observing students' test scores in multiple years allows me to control for student fixed effects, so that variation in fixed student characteristics does not drive estimated differences in student performance across teachers. Because teachers are observed in multiple classrooms, I am able to measure teacher fixed effects while including direct controls for a number of classroom specific factors, such as peer achievement and class size. In addition, I can identify teacher fixed effects only using variation in student performance within particular schools and years, so that variation

[1] The Tennessee Value Added Assessment System, where districts, schools, and teachers are compared based on test score gains averaged over a number of years, is a noteworthy exception.

in teacher quality will not be confounded with variation in school level educational inputs (e.g., principal quality) or idiosyncratic factors that affect test performance at the school level.

This analysis extends research on teacher quality in two additional ways. First, I use a random effects meta-analysis approach to measure the variance of teacher fixed effects while taking explicit account of estimation error. Since estimation error will bias upward the variance of the distribution of teacher fixed effects, the corrected measure provides a more accurate portrayal of the within-school variation in teacher quality. Second, I measure the relation between student achievement and teaching experience using variation across years for individual teachers. This strategy will not confound the causal effect of teaching experience with non-random selection based on teacher quality or differences in teacher quality across cohorts.

Estimates of teacher fixed effects from linear regressions of test scores consistently indicate that there are large differences in quality among teachers within schools. A one standard deviation increase in teacher quality raises test scores by approximately .1 standard deviations in reading and math on nationally standardized distributions of achievement. I find that teaching experience significantly raises student test scores in reading subject areas. Reading test scores differ by approximately .17 standard deviations on average between beginning teachers and teachers with ten or more years of experience. Moreover, estimated returns to experience are quite different if teacher fixed effects are omitted from my analysis. This suggests that using variation across teachers to identify experience effects may give biased results due to correlation between teacher fixed effects and teaching experience.

Policymakers have demonstrated their faith in the importance of teachers by greatly in-

creasing funding for programs that aim to improve teacher quality in low performing schools.[2] However, the vast majority of these initiatives focus on rewarding teachers who possess credentials that have not been concretely linked to student performance (e.g. certification, schooling, teacher exam scores). My results support the idea that raising teacher quality is an important way to improve achievement, but suggest that policies may benefit from shifting focus from credentials to performance-based indicators of teacher quality.

This paper is organized as follows: in section two, I describe the data I collected for this study; in section three, I present my methodology and empirical findings; section four concludes.

2 Matched Student-Teacher Data

I obtained data on elementary school students and teachers in two contiguous districts from a single county in New Jersey. For purposes of confidentiality, I refer to them as districts 'A' and 'B.' Roughly ten thousand students are present in this data, as well as almost three hundred teachers. Teacher identifiers, taken from student test scores, were matched to teachers' highest degree earned and experience level.[3]

The average socioeconomic status of residents in these school districts is above the state median, but considerably below the most affluent districts.[4] The proportion of students

[2]The most recent example is the 'No Child Left Behind Act,' which appropriated over $4 billion for training and recruitment of teachers in 2002. This is in addition to various other federal and state initiatives targeting teachers, such as forgiving student loans, easing qualifications for home mortgages, and waiving tuition for teachers' children who enroll in state universities.

[3]Information on teachers' education and experience was not available for a small portion of teachers in both districts, and for some teachers only their experience teaching in the district was available. However, for teachers where data is not missing, the vast majority had no previous teaching experience when hired. Roughly one third of teachers in these districts have masters degrees.

[4]School districts in New Jersey are placed into District Factor Groups based on the average socio-economic status of their residents, using a composite index of indicators from the most recent U.S. decennial census.

eligible for free/reduced price lunch in these districts fell near the 33rd percentile in the state during the 2000-2001 school year. Spending per pupil that year was slightly above the state average in district A, and slightly below average in district B. Elementary school populations in these districts grew over this time period, but the racial composition of the students was stable. Students in these districts are predominantly (more than 75%) White, with the remainder made up of relatively equal populations of Black, Hispanic, and Asian students.

Several features of elementary education in these districts are helpful for identifying teacher effects. In both districts there are multiple elementary schools serving each grade, and multiple teachers in each grade within each school. Elementary students remain with a single teacher for most of the school day, receive reading and math instruction from this teacher, and are tested at the end of every school year using nationally standardized exams. I can therefore be confident that a student's current teacher is the person from whom they have received almost all instruction since the last time they were tested. In addition, administrators in these districts claim that students are not placed into classrooms based on ability or achievement. In support of this claim, I show in the appendix that classroom assignment is not systematically related to previous achievement levels and that actual classroom assignment produces a mixing of classmates from year to year that resembles random assignment.

Test score data span the 1989-1990 to 2000-2001 school years in district A and the 1989-1990 to 1999-2000 school years in district B. Students take nationally standardized basic skills exams from Kindergarten through 5th grade in district A and from 2nd through 6th

grade in district B.[5] They take as many as four subject area tests in a given year: Reading Vocabulary, Reading Comprehension, Math Computation and Math Concepts.[6] In both districts, more than half of the students I observe were tested at least three times and over one quarter were tested at least five times. The median number of classrooms observed per teacher is six in district A and three in district B. Approximately one half of the teachers in district A and one third of the teachers in district B are observed with more than five classrooms of students.

Students' scores are reported on a Normal Curve Equivalent (NCE) scale.[7] NCE scores range from 1 to 99 (with a mean of 50 and standard deviation of 21) and are standardized by grade level.[8] Figure 1 compares the pooled distribution of NCE scores in these districts to the nationally standardized distribution. Students in these districts score 10-15 NCE points higher on average than the nationwide mean in all subjects. The variance in test score performance within these districts is considerable, though less than the national distribution, and relatively few students score below 30 NCE points.[9]

[5]Students in cohorts who are tested just once will not help to identify teacher fixed effects or other covariates unless they repeat grades. This is because student fixed effects are included in my regression analysis. I therefore drop cohorts who were in the final grade tested during the first year of my sample (e.g., students in 6th grade in 1990 in district A) and in the first grade tested during the final year of my sample (e.g., students in 2nd grade in 2000 in district B).

[6]Both districts administered the Comprehensive Test of Basic Skills (CTBS) at the start of these time period, but switched at some point—District A to the TerraNova CTBS (a revised version of CTBS) and District B to the Metropolitan Achievement Test (MAT). The subtest names are identical across all of these tests, and it is therefore unlikely that the changes reflect a radical shift in the type of material tested or taught to students.

[7]Test makers assert that each NCE point represents an equal increase in test performance, allowing scores to be added, subtracted, or averaged in a more meaningful way than national percentiles. Using national percentiles in my analysis does not noticeably alter the results.

[8]Using scores that are standardized at a particular grade level may be problematic if the distribution of student achievement changes as students grow older. For example, if a change of one NCE point at the 6th grade level represents a much smaller difference in learning than one NCE point at the 1st grade level, then we might want to regard a given amount of variation in 1st grade student performance as representing larger variation in teacher quality than the same variation among 6th graders. I do not attempt to reconcile this possibility in my analysis.

[9]A small but non-trivial percentage (about 3-6%) of scores in each district are at the maximum possible

Analyzing the districts separately reveals no marked differences in results or conclusions, and for simplicity I combine them in the results presented below. Because the number of tests administered varies somewhat over grades and years, and because teacher quality may vary by subject, I examine each subject area separately, and then consider to what extent my results differ across them.

3 Measuring the Importance of Teachers

Equation 1 provides a linear specification of the test score of student i in year t.

$$(1) \quad A_{it} = \alpha_i + \gamma X_{it} + \sum_j (\theta^{(j)} + f(Exp_t^{(j)}) + \eta C_t^{(j)}) D_{it}^{(j)} + \sum_s \pi_{st} S_{it}^{(s)} + \varepsilon_{it}$$

The test score (A_{it}) is a function of the student's fixed characteristics (α_i), time-varying characteristics (X_{it}), a teacher fixed effect ($\theta^{(j)}$), teaching experience ($Exp_t^{(j)}$), observable classroom characteristics ($C_t^{(j)}$), a school-year effect (π_{st}), and all other factors that affect test scores (ε_{it}), including measurement error.[10] $D_{it}^{(j)}$ and $S_{it}^{(s)}$ are indicator variables for

for the test taken, raising the possibility that censoring of 'true' achievement might affect the results of my analysis. I checked for this by performing the main part of my analysis with censored-normal regressions, and the results were not qualitatively different to those presented below. Also, enrolled students who are absent on the day of the test, or change districts earlier in the year, are not observed in the testing data. To see whether the probability of being tested was related to achievement, I use enrollment information available in district B since the 1995-1996 school year. I find no significant relationship between students' previous test scores and their probability of being tested in the following year, both in linear probability and probit regressions.

[10] Learning is a cumulative process, and current inputs, such as teacher quality, may affect both current and future student achievement. In equation 1 there is no explicit relationship between current test scores and past inputs except those that span across years, like α_i. Correlation between the quality of current and past inputs, conditional on the other control variables, will bias my estimates. However, classroom assignment appears similar to random assignment in these districts, so this source of bias is unlikely to affect my results. A simple way to incorporate persistence, used in a number of other studies, is to model teacher effects on test score gains, as opposed to levels. However, this type of model restricts changes in test scores to be perfectly persistent over time, which, if not true, would lead to the same source of bias. In addition, test scores gains can be more volatile, since the idiosyncratic factors that affect test score levels will affect gains to twice the extent.

whether the student had teacher j and school s, respectively, during year t. Implicitly, this model restricts effects to be independent across ages, and assumes no correlation between current inputs and future test scores—zero persistence—except for inputs that span across years, like α_i.[11]

Two issues of collinearity create difficulties in the estimation of equation 1. Experience and year are collinear within teachers (except for a few who leave and return) and grade and year are collinear within students (except for a few who repeat grades).[12] Because of these issues, consistent estimation of teacher fixed effects and experience effects can be achieved only under some identifying assumptions.

I assume that additional experience does not affect student test scores after a certain point ($\overline{Exp}$). Under this assumption, year effects can be separately identified from students whose teachers have experience above the cutoff ($\overline{Exp}$). This assumption is summarized by equation 2, where $D_{Exp_t^{(j)}<\overline{Exp}}$ is an indicator variable for whether teacher j has less than $\overline{Exp}$ years of experience.

$$(2) \quad f(Exp_t^{(j)}) = \widetilde{f}(Exp_t^{(j)}) D_{Exp_t^{(j)}<\overline{Exp}} + \widetilde{f}\left(\overline{Exp}\right)\left(1 - D_{Exp_t^{(j)}<\overline{Exp}}\right)$$

This restriction is supported by previous research, which suggests that the marginal effect of experience declines quickly, and any gains from experience are made in the first few years

[11] Correlation between the quality of current and past inputs, conditional on the other control variables, will bias my estimates. Because classroom assignment appears similar to random assignment in these districts (see appendix), this source of bias is likely to be unimportant. A simple way to incorporate persistence, used in a number of other studies, is to model teacher effects on test score gains, as opposed to levels. However, this type of model restricts changes in test scores to be perfectly persistent over time, which, if not true, would lead to the same source of bias. In addition, test scores gains can be more volatile, since the idiosyncratic factors that affect test score levels will affect gains to twice the extent.

[12] In these districts, I find 9% of teachers had discontinuous careers, and less than 1% of students repeated grades.

of teaching (Rivkin et al., 2001). Moreover, the plausibility of this assumption can be examined by viewing the estimated marginal experience effects at $\overline{Exp}$.[13]

Grade and year are collinear within students (except for the few who repeat grades), so grade and year effects cannot be included simultaneously. I prefer to control for school-year variation because test scores are already normalized by grade level and because there may be considerable idiosyncratic year-to-year variation in school average test scores (Kane and Staiger, 2001). Substituting school-grade effects for school-year effects does not change the results except to increase the estimated impact of teaching experience. If one estimates the model under the assumption that student fixed characteristics are uncorrelated with teacher assignment, so that student fixed effects can be omitted, all interactions between school, grade, and year can be included. This change produces larger estimated impacts of teacher fixed effects and teaching experience than those presented below.

The importance of fixed teacher quality can be measured by the variation in teacher fixed effects. For example, one might measure the expected rise in test score for moving up one standard deviation in the distribution of teacher fixed effects. However, the standard deviation of the estimated fixed effects will overstate the true variation in teacher quality because of sampling error. In order to correct for this bias, I assume that teacher fixed effects ($\theta^{(j)}$) are independently drawn from a normal distribution with some variance σ_θ^2. The set of J true teacher fixed effects (θ) can therefore be considered a mean zero vector

[13] For example, if $f(Exp_{jt})$ is estimated as a quadratic, then $f(Exp_{jt}) = aExp_{jt} + bExp_{jt}$, and one can test whether $a + 2b\overline{Exp} = 0$.

with common variance, as shown by equation 3.

$$(3)\ \ \theta^{(j)} \overset{i.i.d.}{\sim} N(0, \sigma_\theta^2) \Rightarrow \theta \sim N(0, \sigma_\theta^2 I_J)$$

A set of consistent estimates of teacher fixed effects ($\hat{\theta}$) is a normally distributed random vector whose expected value is the set of true teacher effects with some variance. This notion is expressed by equation 4.

$$(4)\ \ \hat{\theta} \sim N(\theta, \hat{V})$$

Given the assumed distribution of teacher fixed effects, these estimates can be re-written as a mean zero vector with variance equal to the sum of the true fixed effects variance (σ_θ^2) and sampling error (equation 5). The true variance of teacher fixed effects can be estimated via maximum likelihood, where ($\hat{V}$) is estimated by the part of the variance-covariance matrix pertaining to the teacher fixed effects estimates.[14]

$$(5)\ \ \widehat{\theta} \sim N(0, \hat{V} + \sigma_\theta^2 I_J)$$

3.1 Test Score Regression Estimates

Table 1 shows estimates of equation 1; each column contains results for one of the four subject areas. $f\left(Exp_{jt}\right)$ is a cubic and $\overline{Exp}$ is set at ten years of experience.[15] Because errors are

[14] This approach is parallel to a random effects meta-analysis, where $\theta^{(j)}$ is an estimated treatment effect from one of many studies, $\hat{V}$ is the estimated variance of these estimates, and σ_θ^2 (the parameter of interest) is the variance of treatment effects across studies.

[15] The cutoff restriction is implemented by recoding experience as follows:

heteroskedastic and possibly serially correlated within students over time, standard errors are clustered at the student level.[16]

The time-varying student controls (X_{it}) I am able to include are indicator variables for being retained or repeating a grade. Students perform lower than their own average in years when they are subsequently held back. They perform higher than their average when repeating a grade, taking the same test for a second time.

The classroom controls ($C_t^{(j)}$) are class size, being in a split-level classroom, and being in the lower half of a split-level classroom.[17] Class size has a statistically insignificant effect on student test scores in all four subject areas. Students in split-level classrooms, both above and below the split, do not perform significantly differently than they do in regular classrooms.[18]

In other regressions (not reported), I check for non-linear effects of class size by including its square and cube. I also try interacting class size with a dummy for being Black or Hispanic, since Krueger (1999) and Rivkin et al. (2001) find that minorities may be more sensitive to class size effects. I do not find statistically significant effects of class size in any of

---

$Exp_{jt} = \begin{cases} Exp_{jt} \text{ if } Exp_{jt} \leq \overline{Exp} \\ \overline{Exp} \text{ if } Exp_{jt} > \overline{Exp} \end{cases}$

Results are similar with other cutoff levels, but this specification is preferred because teachers with more than ten years of experience teach about half of the students in the school-year cells in my data. Results are also similar with other polynomial specifications of $f(Exp_{jt})$, but while the cubic term appears to be important in at least one subject area, quartic or higher order terms do not.

[16] Measurement error in test scores is heteroskedastic by construction. Since tests are geared toward measuring achievement at a particular grade and time, e.g. spring of 3rd grade, the test is less accurate for students who find the test very difficult or very easy.

[17] Split-level classrooms refer to classes where students of adjacent grades are placed in the same classroom. This arrangement was used in district B, albeit infrequently, to help balance class sizes.

[18] The insignificance of classroom characteristics in these regressions may be viewed as somewhat surprising, given the recent literature on these issues and evidence from some studies of teacher effects (Hanushek 1972, Summers and Wolfe 1979). However, these estimates should not be interpreted as causal, since I am not making an effort to credibly identify the effects of these variables from exogenous variation; I am including them as controls so that I can be certain that differences in teacher fixed effects are not driven by differences in these factors.

these specifications. I also include classroom level controls for the average past performance of students' classmates and other peer characteristics as proxies for peer quality.[19] These had no discernible effect on test scores. I do not include these measures here because the use of past achievement as a control variable forces me to drop a substantial fraction of observations from my analysis—an entire grade and year.

The insignificance of classroom characteristics in these regressions may be viewed as somewhat surprising, given the recent literature on these issues and evidence from some studies of teacher effects (Hanushek 1971, Summers and Wolfe 1979). However, these estimates should not be interpreted as causal, since I am not making an effort to credibly identify the effects of these variables from exogenous variation; I include them as controls so that I can be certain that differences in teacher fixed effects are not driven by differences in these factors.

At the bottom of table 1, I report the results of F-tests of the joint statistical significance of teacher fixed effects and the significance of experience. Teacher fixed effects are highly significant predictors of achievement in all four subject areas, with p-values below .001.[20] Experience is a significant predictor of test scores in Vocabulary, Reading Comprehension, and Math Computation, but not Math Concepts.

The raw standard deviation and the estimated underlying standard deviation of teacher fixed effects (σ_θ) are shown in table 2. These are expressed in standard deviations on the

[19] In particular, I also tried including the variance of classmates' test scores, and the number (or proportion) of a students' classmates who: 1) had previous scores one standard deviation above/below the mean, 2) were classified students 3) were enrolled in ESL, 4) were held back or repeating a grade, 5) were female, 6) were Black or Hispanic. I also tried various combinations and interactions of these factors.

[20] In order to be sure that outlying observations on transient teachers do not drive these results, I repeat these tests using only teachers observed in at least three years. P-values for this more selective test are lower in all subject areas. P-values for tests on teacher effects in the regressions that included classmates' previous test scores are all also below .001, both for all teachers and teachers observed at least three times.

national distribution of test scores. For all four subjects, the adjusted standard deviation is considerably lower than the raw standard deviation; for reading and math test scores, the adjusted measures are, respectively, about one half and one third the size of the raw measures. However, the adjusted measures still imply that teacher quality has a large impact on student outcomes. Moving one standard deviation up the distribution of teacher fixed effects is expected to raise both reading and math test scores by about .1 standard deviations on the national scale.[21]

Variation in teacher fixed effects is given in terms of nationally standardized exam scores and is thus easily interpretable. However, it is difficult to know how the distribution of teacher quality in these districts compares to the distribution of quality among broader groups of teachers, for example, statewide or nationwide. Nevertheless, salaries, geographic amenities, and other factors that affect districts' abilities to attract teachers vary to a much greater degree at the state or national level. This suggests that variation in quality within groups of teachers at broader geographical levels may be considerably larger, and that my estimates of the importance of teachers may be conservative. The controls for school-year effects may also lead me to underestimate the magnitude of variation in teacher quality, since any variation in average teacher quality across school-year cells is taken up by these controls.[22]

To better interpret experience effects, I plot point estimates and 95% confidence intervals for the function $\tilde{f}(Exp_t^{(j)})$ in figures 2 and 3. These results provide substantial evidence

[21] Transient teachers do not drive these results either; repeating these calculations using only teachers observed in at least three years gives similar magnitudes.

[22] F-tests of the joint significance of school-year effects show them to be important predictors of test scores in all four subject areas.

that teaching experience improves reading test scores. Ten years of teaching experience is expected to raise both Vocabulary and Reading Comprehension test scores, respectively, by about .15 and .18 standard deviations (figure 2). However, the path of these gains is quite different between the two subject areas. In line with the identifying assumption, the function for Vocabulary scores exhibits positive and declining marginal returns, and gains approach zero as experience approaches the cutoff point. Marginal returns to experience exhibit much slower declines for Reading Comprehension, and suggest that my identification assumption may be violated in this case.[23] However, if returns to experience were positive after the cutoff, as it appears they might be, the experience function I estimate would be biased downward, because estimated school-year effects would be biased to rise over time. Thus, these results may provide a conservative estimate of the impact of teaching experience on Reading Comprehension test scores.

Evidence of gains from experience for the two math subjects is much weaker (figure 3). While the first few years of teaching experience appear to raise scores significantly in Math Computation (about .1 standard deviations), subsequent years of experience appear to lower test scores, though standard errors are too large to conclude anything definitive about these trends. There is not a statistically significant relationship between teaching experience and Math Concepts scores, though point estimates suggest positive returns that come in the first few years of teaching.

Estimates of experience effects should not be affected by any correlation between teachers' fixed effects and their propensity to remain teaching in these districts. However, if teachers

[23]The hypothesis that gains are zero near the cutoff cannot be rejected and the cubic term is negative, but the functional form of $f(Exp_{jt})$ appears fairly linear.

who stay were selected based on their gains from experience, this identification strategy would lead to biased estimates of the expected experience effects for all teachers. While the direction of this potential bias is unclear, these estimates should be interpreted as the expected gains from experience for teachers who stay in these districts.[24]

3.2 Correlation of Teacher Quality Across Subjects

It is quite possible that a teacher is better at teaching one subject than another, and this variation in skill might be important for policy decisions. For example, if the quality of teachers' mathematics instruction was inversely related to the quality of reading instruction, then exchanging teachers between students would have an ambiguous effect on student outcomes, and having teachers specialize in teaching one subject might be more efficient. I briefly examine this question by looking at the pairwise correlations between teachers' fixed effects across subjects, shown in table 3. There are positive correlations between all tests, although correlations between Vocabulary and other subject areas are considerably smaller (.16 to .32) than among the other three subject areas (.46 to .67). There is little indication that teachers who are better at mathematics instruction are worse at reading instruction or vice versa.[25]

[24] Teachers who improve greatly may be more likely to remain if they have gained more firm-specific or occupation-specific human capital, if the district administration is more likely to reappoint them, or if their probability of eventually being offered tenure has increased. On the other hand, if teachers tend to leave after a particularly bad year, and the cause of that poor performance is not persistent, then there may be a negative correlation between expected gains and the probability of staying.

[25] It is also possible that some teachers are better at teaching certain types of students than others. If this were true, then there might be efficiency gains through active matching of students and teachers. In contrast, if the 'good' teachers are equally good for everyone, then the matching of students and teachers probably has more to do with equity than efficiency.

To examine this issue, I estimate quantile regressions at the 25th, 50th, and 75th quantiles. These regressions are of the same form as that used to estimate equation 3, but do not include student fixed effects. (Including student fixed effects requires too much computational power. Even without student fixed effects, the estimated variance-covariance matrix of the estimators must be obtained via bootstrapping, and this can take weeks.) I find teacher fixed effects are significant predictors of test scores in all of these regressions.

Sampling error may bias measures of the correlation of teacher fixed effects across subjects, but the direction of the bias is unclear *a priori*. Errors that are common across subjects will lead to upward bias, and errors that are independent across subjects will lead to downward bias. If the true correlation between subjects is the same for all teachers in this sample, I can gain some insight into the direction of bias by recalculating the correlations using only teachers observed with at least three classrooms, since sampling error is smaller for this subsample. Pairwise correlations among this group of teachers are between .05 and .1 higher in all subject areas, indicating that sampling error is likely to have biased down the correlations shown in table 3.[26]

3.3 Variance Decomposition

To give an idea of the potential scope of teachers' impact on the overall distribution of scores, I estimate upper and lower bounds on the proportion of test score variance accounted for by teacher fixed effects and experience effects. This also serves to demonstrate the potential scope of policies targeted at improving teacher quality. However, my data come from only two districts (and they are quite similar in many respects), so it would be naïve to draw conclusions from these results about how variation in teacher quality across districts might explain variation in achievement.

The upper bound estimate of the variance accounted for by teachers is the adjusted R^2 from a linear regression of test scores on teacher fixed effects and experience effects.

They are also positively correlated: correlation coefficients between the 25th and 75th quantiles for the same subject area range from .50 to .79.

[26] To truly correct for sampling error in these calculations, one would simultaneously estimate teacher effects on all four subject areas in a multiple equation regression framework, and locate the corresponding error variance estimates in the variance-covariance matrix. Since the direction of bias is likely to be downward, and these findings are only an extension to the main results above, I do not pursue this strategy.

The lower bound estimate is the increase in the adjusted R^2 when teacher fixed effects and experience effects are added to a regression specification that contains dummies for students who are retained or repeat grades, student fixed effects, and school-year effects.[27] For comparison, I also estimate lower and upper bounds in this same way for the school-year effects and the student level effects (i.e., fixed effects and the controls for being retained and repeating a grade). Table 4 shows these results. Across subject areas, the upper bound estimates range from 5.0-6.4% for teacher effects, 2.7-6.1% for school-year effects, and 59-68% for student fixed effects. The lower bound estimates range from 1.1-2.8% for teacher effects, .4% to 2.3% for school-year effects, and 57-64% for student effects.

The lower bound estimates of test score variance accounted for by teacher effects may seem small. However, when thinking about the role of policies, one should keep in mind that explaining the total variance in test scores with policy-relevant factors is probably impossible. Idiosyncratic factors and natural variation in cognitive ability among students are surely beyond policymakers' control. Moreover, policymakers often avoid intervention in the home, and household factors may play a large role in determining test score outcomes.

A better characterization may be to calculate the proportion of "policy-relevant" test score variance accounted for by teachers. An estimate of policy-relevant variance can be found by taking the fraction of test score variance due to measurement error–say .10–and the lower bound estimate of the fraction of test score variance attributed to student-level variables–.57 to .64–and subtracting their sum from 1.[28] Using the estimates in table 4,

[27] I omit classroom characteristics from this part of my analysis because they do not have significant predictive power for test scores.

[28] A tenth of variance due to measurement error is a standard and perhaps conservative estimate. Standardized test makers publish reliability coefficients, which estimate the correlation of test-retest scores for the same student, and these usually are about .9 or slightly below. One minus this reliability coefficient is equivalent to the percentage of variance due to idiosyncratic factors, or what we call measurement error for

I find differences among teachers explain proportions of policy-relevant test score variance ranging from lower bounds of 4-9% to upper bounds of 16-23%.

4 Conclusion

The empirical evidence in this paper suggests that raising teacher quality may be a key instrument in improving student outcomes. However, in an environment where many observable teacher characteristics are not related to teacher quality, policies that focus on recruiting and retaining teachers with particular credentials may be less effective than policies that reward teachers based on performance.

As measures of effective teaching, test scores are widely available, objective, and (though they may not capture all facets of what students learn in school) they are widely recognized as important indicators of achievement by educators, policymakers, and the public. A number of states have begun rewarding teachers with non-trivial bonuses based on the average test performance of students in their schools, but few areas (Cincinnati, Denver) have pursued programs that link individual teacher salaries to their own students' achievement. Recent studies of pay-for-performance incentives for teachers in Israel (Lavy 2002a, Lavy 2002b) indicate that both group- and individual-based incentives have positive effects on students' test scores, and that individual-based incentives may be more cost-effective.

Teacher evaluations may also present a simple and potentially important indicator of teacher quality. There is already substantial evidence that principals' opinions of teacher

simplicity. On the other hand, it is probably the case that some of the variance in test scores stemming from cognitive ability and household factors can be affected by education-based policy initiatives. For example, special education programs may increase the average test score performance of students with learning disabilities. Measuring the degree to which this is possible is clearly an extremely difficult exercise, and certainly beyond the scope of this paper.

effectiveness are highly correlated with student test scores (Murnane 1975, Armor et al. 1976), and while evaluations introduce an element of subjectivity, they may also reflect valuable aspects of teaching other than improving test performance.

However, efforts to improve the quality of public school teachers face some difficult hurdles, the most daunting of which is the growing shortage of teachers. Hussar (1998) estimated the demand for newly hired teachers between 1998 and 2008 at 2.4 million—a staggering figure, given that there were only about 2.8 million teachers in the U.S. during the 1999-2000 school year.[29] Underlying this prediction is the fact that the fraction of teachers nearing retirement age has been growing steadily over the past two decades and continues to do so. In 1978, 25.7% of elementary and secondary public school teachers were over the age of 45; by 1998 that figure was 47.8%.

There is also evidence that union wage compression and improved labor market opportunities for highly skilled females have led to a decline in the supply of highly skilled teachers over the last several decades (Corcoran et al., 2002, Hoxby and Leigh, 2003). Indeed, the average income of female teachers relative to college-educated women in other professions has declined substantially over this time period.[30] Although recent evidence indicates women who were once full-time teachers usually do not leave the education profession for a job that pays more money (Scafadi et al. 2002), there may be many women (and men) who would

[29] Notably, this prediction does not take into account possible reductions in class size, which would considerably increase the need for new teachers. Even if lowering class size has a significant beneficial effect on student achievement, it will certainly cause a temporary drop in average experience levels, and may lower long run teacher quality if new teachers are of lower quality than current teachers. Moreover, the impact of class size reduction may vary by district, since wealthy districts may fill their increased demand for new teachers with the highest quality teachers from poorer areas. Jepsen and Rivkin (2002) provide evidence that this type of shifting in teacher quality took place after class size reduction legislation was enacted in California.

[30] See Hanushek and Rivkin (1997).

make excellent teachers, but choose not to teach for monetary reasons.

Given this set of circumstances, it is clear that much research is still needed on how high quality teachers may be identified, recruited, and retained. Seeking out and compensating teachers solely on the basis of education and experience (above the first few years) is unlikely to yield large increases in teacher quality, though currently this is common practice. Finding alternative sources of information on teacher quality may be crucial to the creation of effective policies to raise student achievement.

References

[1] Aaronson, Daniel, Lisa Barrow and William Sander, "Teachers and Student Achievement in the Chicago Public High Schools," Federal Reserve Bank of Chicago Working Paper, February, 2003.

[2] Abowd, John M and Francis Kramarz, "The Analysis of Labor Markets Using Matched Employer-Employee Data" Handbook of Labor Economics. Volume 3B. Ashenfelter, Orley Card, David, eds., Handbooks in Economics, vol. 5. Amsterdam; New York and Oxford: Elsevier Science, North-Holland. p 2629-2710. 1999.

[3] Armor, David, et al., "Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools," RAND Publication, August, 1976.

[4] Bertrand, M. and A. Schoar, "Managing with Style: The Effect of Managers on Firm Policies", Working Paper, University of Chicago and MIT, April, 2002.

[5] Corcoran, Sean, William Evans and Robert Schwab, "Changing Labor Market Opportunities for Women and the Quality of Teachers 1957-1992," Working Paper, August, 2002.

[6] Figlio, David N., "Teacher Salaries and Teacher Quality," Economics Letters, August, 1997.

[7] Hanushek, Eric A., "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data," American Economic Review, May, 1971.

[8] Hanushek, Eric A., "The Economics of Schooling: Production and Efficiency in Public Schools," Journal of Economic Literature, September, 1986.

[9] Hanushek, Eric A. and Steven G. Rivkin, "Understanding the Twentieth-Century Growth in U.S. School Spending," Journal of Human Resources, Winter, 1997.

[10] Hoxby, Caroline and Andrew Leigh, "Pulled Away or Pushed Out? Explaining the Decline of Teacher Quality in the United States," Working Paper, December, 2003.

[11] Hussar, William J., "Predicting the Need for Newly Hired Teachers in the United States to 2008-09," Research and Development Report, National Center for Educational Statistics, August, 1999.

[12] Kane, Thomas and Douglas Staiger, "Improving School Accountability Measures,"

NBER Working Paper 8156, March, 2001.
[13] Krueger, Alan, "Experimental Estimates of Education Production Functions," Quarterly Journal of Economics, May, 1999.
[14] Lavy, Victor, "Paying for Performance: The Effect of Teachers' Financial Incentives on Students' Scholastic Outcomes," Working Paper, July, 2002.
[15] Lavy, Victor, "Evaluating the Effect of Teacher Performance Incentives on Students' Achievements," Journal of Political Economy, December, 2002.
[16] Murnane, Richard, The Impact of School Resources on the Learning of Inner City Children. Cambridge, MA: Ballinger, 1975.
[17] Park, Albert and Emily Hannum, "Do Teacher Affect Learning in Developing Countries?: Evidence from Matched Student-Teacher Data from China," Working Paper, April, 2002.
[18] Rivkin, Steven G., Eric A. Hanushek and John F. Kain, "Teachers, Schools, and Academic Achievement," Working Paper, April, 2001.
[19] Scafidi, Benjamin, David Sjoquist and Todd R. Stinebrickner, "Where Do Teachers Go?," Working Paper, October, 2002.
[20] Summers, Anita and Barbara Wolfe, "Do Schools Make a Difference?," American Economic Review, September, 1977.
[21] Uribe, Claudia, Richard J. Murnane and John B. Willett, "Why Do Students Learn More in Some Classrooms Than in Others? Evidence from Bogota," Working Paper, November 4, 2003.

A Tests for Systematic Classroom Assignment

To test for systematic differences in the groups of students assigned to particular teachers (i.e., tracking), I test if current classrooms are significant predictors of past test scores. To do so, I calculate the residuals from a regression of past test scores on school-year-grade dummies, regress these residuals on classroom dummies, and test the significance of variation in past test scores across classrooms using a joint F-test on these dummy variables. I only look at variation within school-year-grade cells because administrators can only change the classroom to which they assign students, not the school, year or grade. Table A.1 shows, by district, the F-statistics and p-values for these tests in each of the four subject areas. All of the p-values are close to one, substantiating administrators' claims that there was no systematic classroom assignment based on ability/achievement.

I also examine how students are mixed from year to year as they progress to higher grades, i.e., if administrators tend to keep the same groups of students together for successive years. This type of systematic classroom assignment would not be captured by differences in past achievement across classrooms. I examine this issue through calculation of dissimilarity indices, commonly used to measure spatial segregation (e.g. of racial groups in neighborhoods within a city). One can see the intuition for using this measure by asking: are students in a particular school-grade-year cell 'segregated' across current classrooms by their previous classroom? If one considers a school-grade-year cell like a city, a classroom like a neighborhood, and a student's previous classroom like a racial group, the issues are clearly parallel.

To indicate what dissimilarity indices would look like with random assignment, I generate data where students from four 'classrooms' of 20 students each are randomly placed into four new 'classrooms' of 20 students each—this is fairly representative of the school-year-grade cells in my data. Dissimilarity indices from this monte carlo exercise are located predominantly between .1 and .3. Figure A.1 shows, by district, the actual proportion of school-grade-year dissimilarity indices falling between zero and .1, .1 and .2, etc. A large majority of cells have indices between .1 and .3, giving strong evidence that the mixing of classmates from year to year in these districts is similar to random assignment.[31]

[31] Though indices decrease with the number of students in each classroom and increase with the number of classrooms, but large changes in the parameters I use (e.g., 100 students per classroom or 20 classrooms per school-grade cell) are needed to radically change the results. Also, a tiny fraction of school-grade-year cells in district B have indices above .6. This is driven by the small number of classrooms in district B that are 'split-level', i.e. they have students from adjacent grades placed in the same classroom. It is obvious when looking at the data that many of the students placed in the lower grade of a split-level classroom remain with that teacher the following year if that teacher is assigned a split-level classroom.

Figure 1: Distribution of School Districts' Test Scores Relative to National Distribution

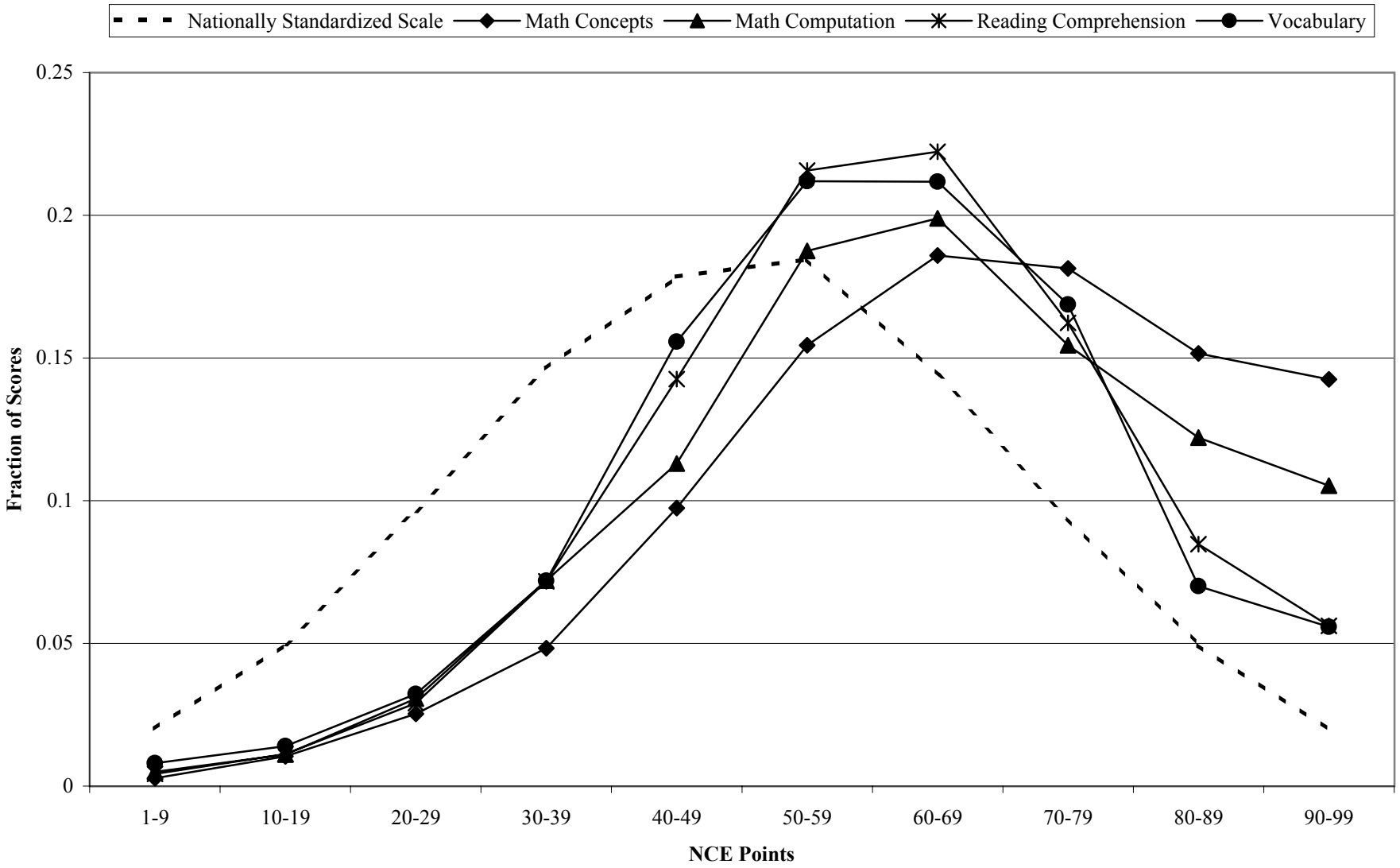
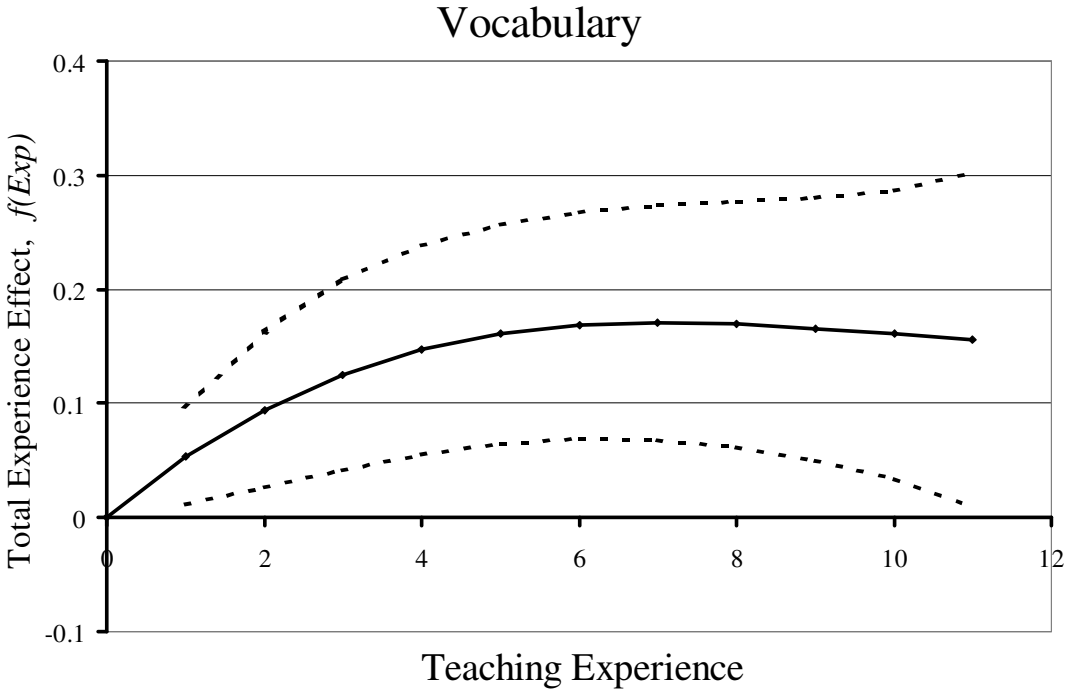
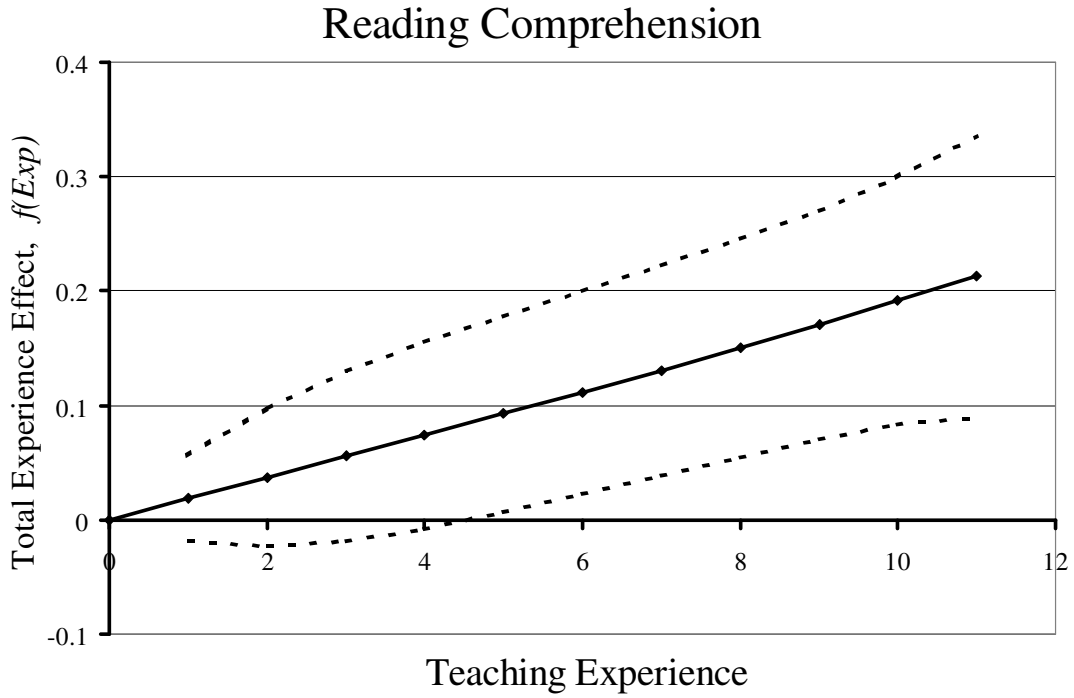


Figure 2: The Effect of Teacher Experience on Reading Achievement, Controlling for Fixed Teacher Quality

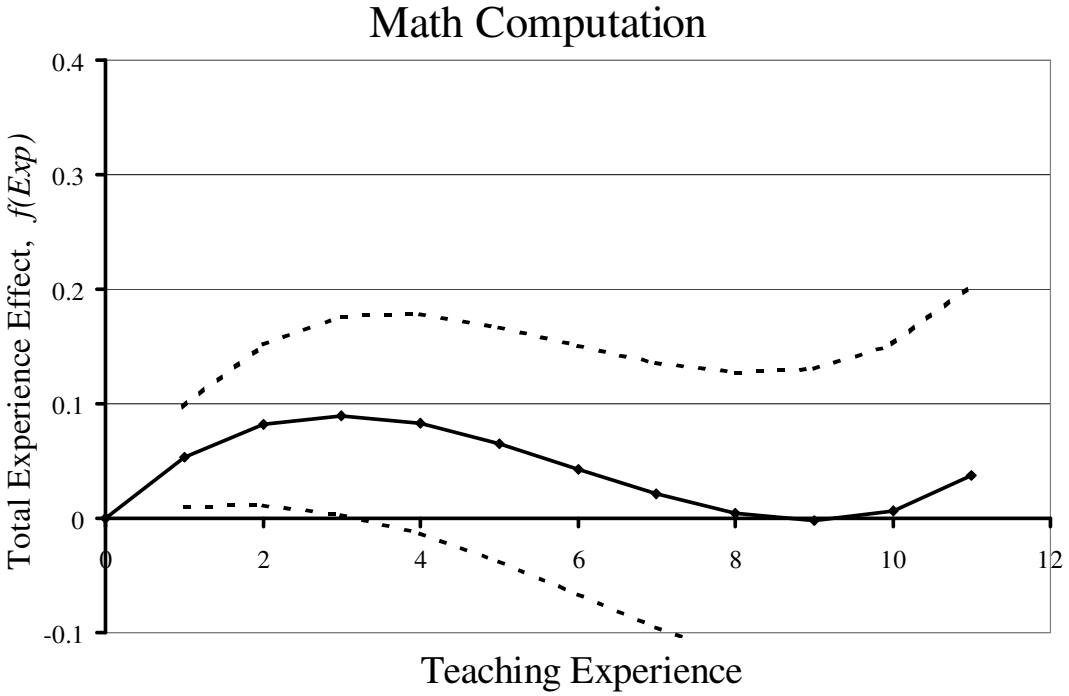


Note: Dotted lines are bounds of the 95% confidence interval.

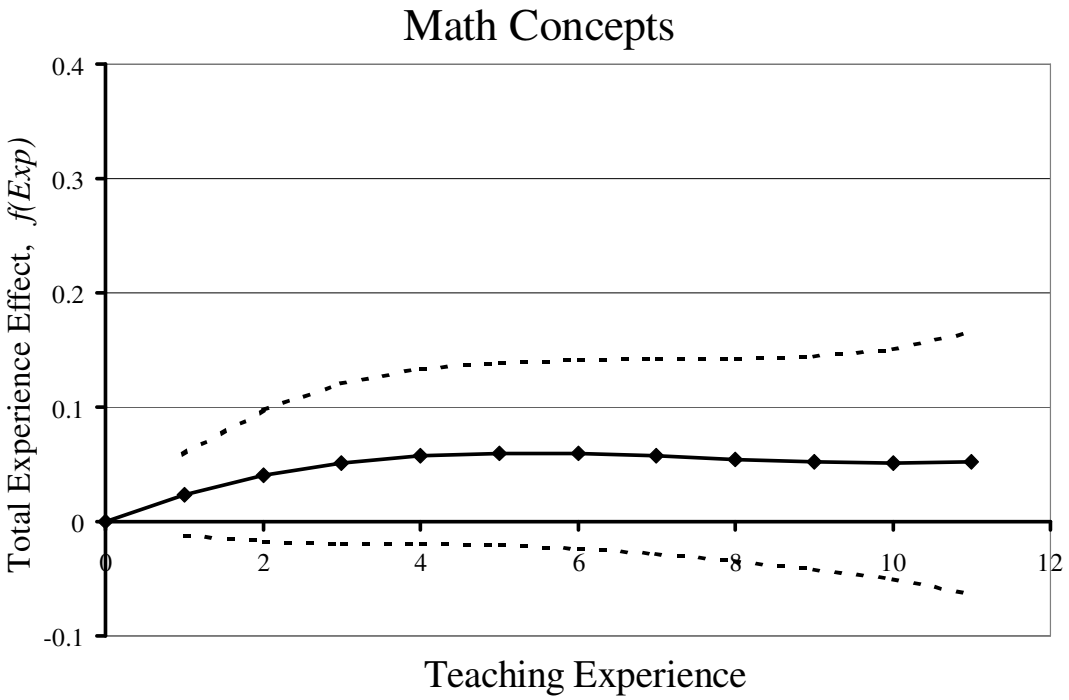


Note: Dotted lines are bounds of the 95% confidence interval.

Figure 3: The Effect of Teacher Experience on Math Achievement, Controlling for Fixed Teacher Quality



Note: Dotted lines are bounds of the 95% confidence interval.



Note: Dotted lines are bounds of the 95% confidence interval.

Table 1: Student Test Score Regression Results

	Reading Vocabulary	Reading Comprehension	Math Computation	Math Concepts
Held Back	-7.383	-9.470	-17.204	-16.129
	(2.692)**	(2.033)**	(4.601)**	(2.020)**
Repeating Grade	10.144	11.404	9.935	9.190
	(2.807)**	(2.727)**	(2.820)**	(2.307)**
Class Size	0.048	-0.087	0.095	0.085
	(0.068)	(0.062)	(0.077)	(0.058)
Split-level Classroom	0.877	0.232	-0.704	0.003
	(0.599)	(0.524)	(0.604)	(0.578)
Below Split in Split-level Classroom	-0.283	-1.402	0.344	-0.873
	(0.665)	(0.598)*	(0.687)	(0.672)
Experience	1.250	0.399	1.431	0.571
	(0.538)*	(0.475)	(0.558)*	(0.453)
Experience^2	-0.138	-0.004	-0.321	-0.082
	(0.107)	(0.094)	(0.115)**	(0.089)
Experience^3	0.005	0.000	0.018	0.004
	(0.006)	(0.005)	(0.007)**	(0.005)
Constant	54.803	53.504	50.497	54.930
	(4.456)**	(4.571)**	(6.097)**	(4.354)**
Observations	23921	27610	24705	30316
R-squared	0.80	0.79	0.78	0.81
F-test, H_0: $f(exp)=0$	**4.01**	**4.77**	**2.78**	**0.81**
p-value	**(0.01)****	**(<0.01)****	**(0.04)***	**(0.49)**
F-test, H_0: $\{\theta\}=0$	**4.43**	**2.75**	**3.72**	**5.30**
p-value	**(<0.01)****	**(<0.01)****	**(<0.01)****	**(<0.01)****

Test scores are expressed on a Normal Curve Equivalent scale; one standard deviation on this scale is 21 points. All regressions include teacher and student fixed effects, a cubic in experience, and school-year effects. Standard errors (in parentheses) are clustered by pupil. * significant at 5%; ** significant at 1%

Table 2: Variation of Teacher Fixed Effects

	Raw S.D.	Adjusted S.D.	*# Teachers*
Reading Vocabulary	0.21	0.11 (0.04)	*224*
Reading Comprehension	0.20	0.08 (0.03)	*252*
Math Computation	0.28	0.11 (0.02)	*263*
Math Concepts	0.30	0.10 (0.04)	*297*

Note: Teacher fixed effects are estimated in regressions that include controls for being held back or repeating a grade, class size, being in a split-level classroom and being in the lower half of a split-level classroom, student fixed effects, school-year effects, and experience effects. Adjusted measures are based on maximum likelihood estimates of the underlying variance of the teacher fixed effect distribution. See explanation in text. Standard errors in parentheses.

Table 3: Correlation of Teacher Fixed Effect Estimates Across Subject Area Tests

	Reading Vocabulary	Reading Comprehension	Math Computation	Math Concepts
Reading Vocabulary	1.00			
Reading Comprehension	0.27	1.00		
Math Computation	0.16	0.46	1.00	
Math Concepts	0.32	0.58	0.67	1.00

Note: These are the pairwise correlations of teacher fixed effects across subjects. The teacher fixed effects used to calculate these correlations are estimated in regressions of test scores that include controls for students who are retained or repeat a grade, class size, being in a split-level classroom and being in the lower half of a split-level classroom, student fixed effects, a cubic in experience, and school-year effects.

Table 4: Test Score Variance Decomposition

Teacher Fixed Effects and Experience	Lower Bound R-sq[2]	Upper Bound R-sq[1]	*Base R-sq*[2]
Reading Vocabulary	0.018	0.050	*0.690*
Reading Comprehension	0.011	0.051	*0.691*
Mathematics Computation	0.028	0.052	*0.619*
Mathematics Concepts	0.025	0.064	*0.700*
School-Year Effects			
Reading Vocabulary	0.009	0.034	*0.699*
Reading Comprehension	0.004	0.039	*0.698*
Mathematics Computation	0.015	0.027	*0.632*
Mathematics Concepts	0.023	0.061	*0.703*
Student-Level Effects			
Reading Vocabulary	0.643	0.676	*0.065*
Reading Comprehension	0.641	0.683	*0.061*
Mathematics Computation	0.575	0.595	*0.073*
Mathematics Concepts	0.624	0.658	*0.102*

Notes: Upper bound estimates are the adjusted R^2 from a regression of test scores on just the factor in question: school year effects, teacher dummy variables and a cubic in experience, or student fixed effects and controls for students who are retained or repeat a grade. Lower bound estimates are the increase in adjusted R^2 from adding one of the sets of factors to a regression of test scores that included the other two sets of factors as controls. The adjusted R^2 from this latter regression is the Base R^2, shown in the third column.

Figure A.1: Dissimilarity Indices by School-Grade-Year Cell (Segregation by Previous Classroom)

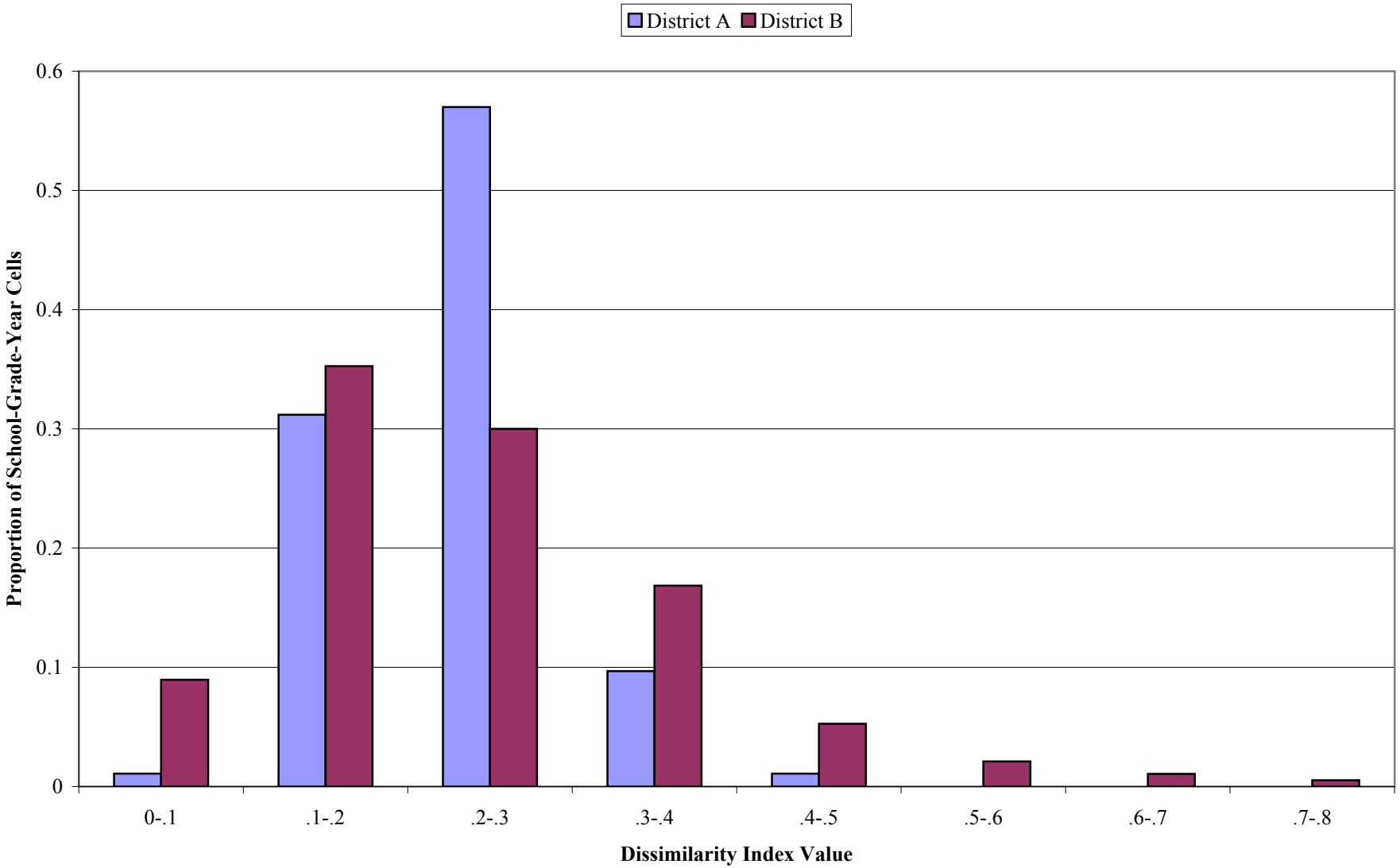


Table A.1: Statistical Tests for Tracking by District and Test

	District A		District B	
	F-statistic	P-value	F-statistic	P-value
Reading Vocabulary	0.74	1.00	0.88	0.97
Reading Comprehension	0.77	1.00	0.91	0.90
Math Computation	0.77	1.00	0.90	0.95
Math Concepts	0.74	1.00	0.94	0.85

Notes: F-tests are on the joint significance of classroom dummies to predict past test scores within school-year-grade cells.