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

By JONAH E. ROCKOFF*

School administrators, parents, and students themselves widely support the notion that teacher quality is vital to student achievement, despite inconsistent evidence linking achievement to observable teacher characteristics (Erik A. Hanushek, 1986). 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. Researchers have therefore 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.

Credible identification of teacher fixed effects requires matched student–teacher data wherein both student achievement and teachers are observed in multiple years. This type of data is not readily available to researchers, in large 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; Richard Murnane, 1975; David Armor et al., 1976; Albert Park and Emily Hannum, 2002; Claudia Uribe et al., 2003) or used data collected by a research institution (Daniel Aaronson et al., 2003; Steven G. Rivkin et al.,

2003). Almost all of the empirical difficulties in these studies are related to data quality. For instance, teacher effects cannot be separated from other classroom-specific factors in several of these studies because teachers were only observed with one class of students.

I use a rich set of panel data on student test scores and teacher assignment to estimate more accurately how much teachers affect student achievement. Panel data on students' test scores allow me to focus on differences in the performance of the same student with different teachers, and thus to distinguish variation in teacher quality from variation in students' cognitive abilities and other characteristics. Observing the same teacher with multiple classrooms allows me to differentiate teacher quality from factors such as class size. In addition, by focusing on variation in student achievement within particular schools and years, I separate variation in teacher quality from variation in school-level educational inputs (e.g., principal quality) and time-varying 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 nonrandom selection based on teacher quality or differences in teacher quality across cohorts.

My empirical results indicate large differences in quality among teachers within schools. A one-standard-deviation increase in teacher quality raises test scores by approximately 0.1 standard deviations in reading and math on

[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.

nationally standardized distributions of achievement. I also find evidence that teaching experience significantly raises student test scores, particularly in reading subject areas. Reading test scores differ by approximately 0.17 standard deviations on average between beginning teachers and teachers with ten or more years of experience.

I. Matched Student–Teacher Panel Data

I obtained data on elementary-school students and teachers from two contiguous districts located in a single New Jersey county. I will refer to them as districts A and B.[2] 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 the entire school day, receive reading and math instruction from this teacher, and are tested in basic reading and math at the end of each school year using nationally standardized exams.[3] Analyzing the districts separately revealed no marked differences in results, and I combine them here for simplicity and estimation power.

Test-score data span the school years from 1989–1990 to 2000–2001 in district A and from 1989–1990 to 1999–2000 in district B. Students are tested from Kindergarten through 5th grade in district A, and from 2nd through 6th grade in district B. They take as many as

four subject-area tests per year: Reading Vocabulary, Reading Comprehension, Math Computation, and Math Concepts.[4]

Roughly 10,000 students are present in these data, as well as almost 300 teachers. 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.

II. Student Test Scores and Teacher Quality

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

$$(1) \quad A_{it} = \alpha_i + \gamma X_{it} + \sum_j (\theta^{(j)} + f(\text{Exp}_t^{(j)})$$

$$+ \eta C_t^{(j)})D_{it}^{(j)} + \sum_s \pi_{st} S_{it}^{(s)} + \varepsilon_{it}.$$

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 ($\text{Exp}_t^{(j)}$), 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.[5] $D_{it}^{(j)}$ and $S_{it}^{(s)}$ are

---

[2] For reasons of confidentiality, I refrain from giving any information that could be used to identify these districts. However, it is important to note that these districts do not represent especially advantaged or disadvantaged populations. According to New Jersey state school district classifications, the average socioeconomic status of residents in these school districts is above the state median, but considerably below that of the most affluent districts. The proportion of students 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.

[3] In addition, according to district officials, students are not placed into classrooms based on ability or achievement. In support of this claim, I find that classroom dummy variables do not have significant predictive power for previous test scores within schools and grades, and actual classroom assignment produces a mixing of classmates from year to year that resembles random assignment. Results for these tests are available from the author upon request.

[4] Over this time period, districts administered the Comprehensive Test of Basic Skills (CTBS), the TerraNova CTBS (a revised version of CTBS), and the Metropolitan Achievement Test (MAT). The subject-area names are identical across all of these tests. Because I only examine variation in teacher quality within schools and years, potential effects of test acclimation at the district or school level will not be attributed to teachers.

[5] In equation (1) there is no explicit relationship between current test scores and past inputs except those that span across years, like α_i. Learning is a cumulative process, and current inputs, such as teacher quality, may affect both current and future student achievement. If the quality of current inputs and past inputs are correlated, conditional on the other control variables, such persistence 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,

indicator variables for whether the student had, respectively, teacher j and school s during year t.

Experience and year are collinear within teachers (except for the few who leave and return) so the identification of both teacher fixed effects and experience effects requires an assumption regarding the form of $f(Exp)$. I assume that additional teaching experience does not affect student test scores after a certain cutoff point $(\overline{Exp})$. Year effects are therefore identified from students whose teachers have experience above the cutoff. 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:

$$(2) \quad f(Exp_t^{(j)}) = \tilde{f}(Exp_t^{(j)})D_{Exp_t^{(j)} < \overline{Exp}}$$

$$+ \tilde{f}(\overline{Exp})D_{Exp_t^{(j)} \geq \overline{Exp}}.$$

This restriction is supported by previous research that suggests any gains from experience are made in the first few years of teaching (Rivkin et al., 2001). Moreover, the plausibility of this assumption can be examined by viewing the estimated marginal effect of additional experience at $\overline{Exp}$.

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 character of my results.

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 actual 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 with common variance, as shown by equation (3):

$$(3) \quad \theta^{(j)} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\theta^2) \Rightarrow \theta \sim \mathcal{N}(\mathbf{0}, \sigma_\theta^2 \mathbf{I}_J).$$

Equation (4) expresses that 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 (θ) and whose variance is estimated by the variance–covariance matrix of the teacher fixed-effects estimates $(\hat{\mathbf{V}})$:

$$(4) \quad \hat{\theta} \sim \mathcal{N}(\theta, \hat{\mathbf{V}}).$$

Given the distribution of the true fixed effects, the estimated teacher fixed effects are distributed as a mean-zero vector with variance equal to the sum of the true fixed effects variance and sampling error [equation (5)]. The underlying variance of teacher fixed effects (σ_θ^2) can be estimated via maximum likelihood:[6]

$$(5) \quad \hat{\theta} \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{V}} + \sigma_\theta^2 \mathbf{I}_J).$$

Table 1 shows F tests of the joint significance of teacher fixed effects and the joint significance of experience from estimates of equation (1).[7]

---

[6] 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.

[7] All standard-error calculations allow for clustering at the student level; $\tilde{f}(Exp_t^{(j)})$ is specified as a cubic, and $\overline{Exp}$ is set equal to ten years of experience. The time-varying student controls (X_{it}) I am able to include are indicator variables for being retained or repeating a grade. 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. Split-level classrooms refer to classes where students of adjacent grades are placed in the same classroom. The coefficient estimates for these variables are not shown for simplicity but are available from the author upon request. Notably, class size did not have a statistically

---

would lead to the same type of bias. In addition, test-score gains can be more volatile, since the idiosyncratic factors that affect test-score levels will affect gains to twice the extent (Thomas Kane and Douglas Staiger, 2001).

TABLE 1—STUDENT TEST SCORES AND TEACHER QUALITY

Statistic	Reading		Math	
	Vocabulary	Comprehension	Computation	Concepts
Number of observations	23,921	27,610	24,705	30,316
R^2	0.80	0.79	0.78	0.81
F test, H_0:	4.01	4.77	2.78	0.81
$f(\text{Exp}) = 0$	(0.01)	(<0.01)	(0.04)	(0.49)
F test, H_0:	4.43	2.75	3.72	5.30
$\{\theta\} = 0$	(<0.01)	(<0.01)	(<0.01)	(<0.01)

Notes: All regressions include teacher fixed effects, student fixed effects, and school-year effects. Numbers in parentheses are p values from F tests.

TABLE 2—WITHIN-SCHOOL VARIATION IN TEACHER QUALITY: THE STANDARD DEVIATION OF THE TEACHER FIXED-EFFECTS DISTRIBUTION

Measure	Raw SD	Adjusted SD	Number of 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

Notes: 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 are reported in parentheses.

Teacher fixed effects are significant predictors of test scores in all subject areas; the p values for the tests all fall below 0.01. Experience is a significant predictor of test scores for both reading subjects 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 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 0.1 standard deviations on the national scale.

It is difficult to know how the distribution of within-school teacher quality in these districts compares to the distribution of quality among broader groups of teachers, for example, statewide or nationwide. However, 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 among teachers at broader geographical levels may be considerably larger than the within-school estimates presented here.

To better interpret experience effects, I plot point estimates and 95-percent confidence intervals for the function $f(\text{Exp}_t^{(j)})$ in Figure 1. These results provide substantial evidence that teaching experience improves reading test scores. Ten years of teaching experience is expected to raise vocabulary and reading-comprehension test scores, respectively, by about 0.15 and 0.18 standard deviations (Fig. 1B). 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 declining marginal returns that approach zero as experience approaches the cutoff point. Marginal returns to experience appear to be linear for reading comprehension and suggest that my identification assumption may be violated in this case. 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 conser-

---

significant effect on student test scores in any subject area. The average past performance of students' classmates and other peer characteristics also were not significant predictors of student test scores. I do not include these measures of peer quality in the analysis presented here because the inclusion of past achievement as a control variable forces me to drop a substantial fraction of observations (i.e., an entire grade and year).
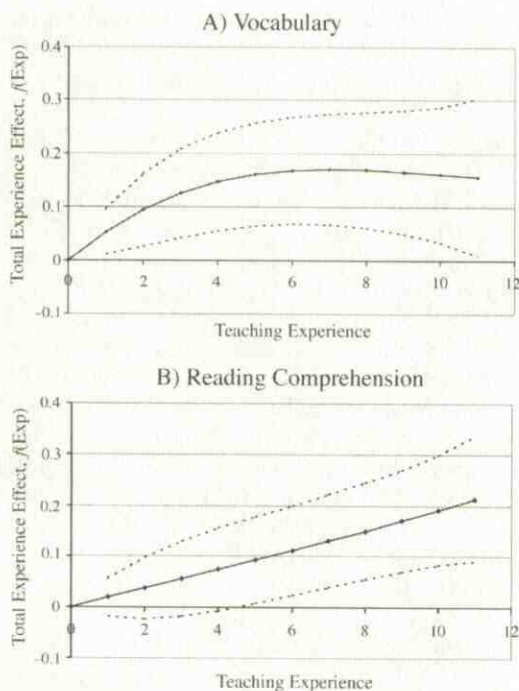
FIGURE 1. THE EFFECT OF TEACHER EXPERIENCE
ON READING ACHIEVEMENT, CONTROLLING
FOR FIXED TEACHER QUALITY

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

FIGURE 2. THE EFFECT OF TEACHER EXPERIENCE ON MATH
ACHIEVEMENT, CONTROLLING FOR FIXED TEACHER QUALITY

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

vative estimate of the impact of teaching experience on reading comprehension test scores.

Evidence of gains from experience for math subjects is weaker (Fig. 2). The first two years of teaching experience appear to raise scores significantly in math computation (about 0.1 standard deviations). However, subsequent years of experience appear to lower test scores, though standard errors are too large to conclude anything definitive about this latter trend. 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.

III. Conclusion

The empirical evidence above suggests that raising teacher quality may be a key instrument in improving student outcomes. However, in an environm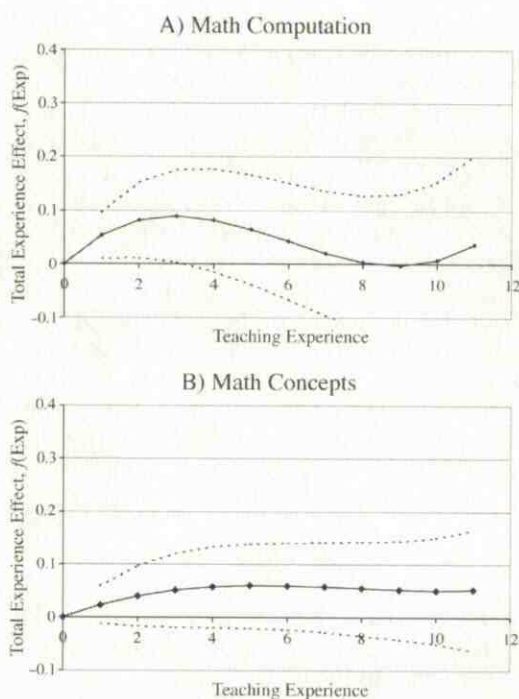ent where many observable teacher characteristics are not related to teacher quality, policies that reward teachers based on credentials may be less effective than policies that reward teachers based on performance. Test scores do not capture all facets of student learning. Nevertheless, test scores are widely available, objective, and are widely recognized as important indicators of achievement by educators, policymakers, and the public. Recent studies of pay-for-performance incentives for teachers in Israel (Victor Lavy, 2002a, b) indicate that both group- and individual-based incentives have positive effects on students' test scores.

Teacher evaluations may also present a simple and potentially important indicator of teacher quality. There is already substantial evidence that principals' opinions of teacher quality are highly correlated with student test scores (Murnane, 1975; Armor et al., 1976). Moreover, while evaluations introduce an element of subjectivity, they may also reflect valuable aspects of teaching not captured by student test scores.

Efforts to improve the quality of public-school teachers face some difficult hurdles, the most daunting of which is the growing shortage of teachers. William J. Hussar (1999) 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 United States during the 1999–2000 school year. 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 (Sean Corcoran et al., 2004; Caroline Hoxby and Andrew Leigh, 2004). 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.

REFERENCES

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

Armor, David; Conry-Oseguera, Patricia; Cox, Millicent; King, Nicelma; McDonnell, Lorraine; Pascal, Anthony; Pauly, Edward; Zellman, Gail; Sumner, Gerald and Thompson, Velma Montoya. "Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools." RAND (Santa Monica, CA) Report No. R-2007-LAUSD, August 1976.

Corcoran, Sean; Evans, William N. and Schwab, Robert M. "Changing Labor-Market Opportunities for Women and the Quality of Teachers 1957–2000." *American Economic Review*, May 2004 (*Papers and Proceedings*), *94*(2), pp. 230–35.

Hanushek, Eric A. "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *American Economic Review*, May 1971 (*Papers and Proceedings*), *61*(2), pp. 280–88.

————. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature*, September 1986, *24*(3), pp. 1141–77.

Hoxby, Caroline and Leigh, Andrew. "Pulled Away or Pushed Out? Explaining the Decline of Teacher Quality in the United States." *American Economic Review*, May 2004 (*Papers and Proceedings*), *94*(2), pp. 236–40.

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, Washington, DC, August 1999.

Kane, Thomas and Staiger, Douglas. "Improving School Accountability Measures," National Bureau of Economic Research (Cambridge, MA) Working Paper No. 8156, March 2001.

Lavy, Victor. "Paying for Performance: The Effect of Teachers' Financial Incentives on Students' Scholastic Outcomes." Working paper, Hebrew University, Jerusalam, Israel, July 2002a.

————. "Evaluating the Effect of Teacher Performance Incentives on Students' Achievements." *Journal of Political Economy*, December 2002b, *110*(6), pp. 1286–1317.

Murnane, Richard. *The impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger, 1975.

Park, Albert and Hannum, Emily. "Do Teachers Affect Learning in Developing Countries? Evidence from Matched Student–Teacher Data from China." Unpublished manuscript, University of Michigan, April 2002.

Rivkin, Steven G.; Hanushek, Eric A. and Kain, John F. "Teachers, Schools, and Academic Achievement." Working paper, The Cecil and Ida Green Center for the Study of Science and Society, Dallas, TX, September 2003.

Uribe, Claudia; Murnane, Richard J. and Willett, John B. "Why Do Students Learn More in Some Classrooms Than in Others? Evidence from Bogota." Working paper, Harvard Graduate School of Education, November 2003.